

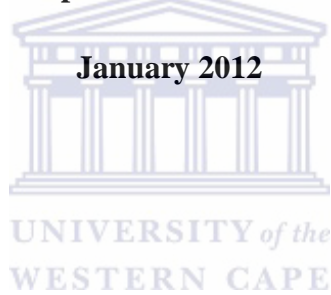
**EXPLORING THE SCALAR EQUIVALENCE OF THE PICTURE VOCABULARY
SCALE OF THE WOODCOCK MUNOZ LANGUAGE SURVEY ACROSS RURAL
AND URBAN ISIXHOSA-SPEAKING LEARNERS**

Qunita Brown

A mini-thesis submitted in partial fulfilment of the requirements for the degree of MA
(Research) Psychology in the Department of Psychology, University of the Western Cape

Supervisor: Professor SE Koch

Co-supervisor: M. Florence



Keywords: Differential item functioning, Scalar equivalence, Construct equivalence, Woodcock Munoz Language survey, Picture vocabulary scale, isiXhosa dialects, Secondary data analysis, Exploratory factor analysis, Bias, Equivalence

TABLE OF CONTENTS

	Page
ABSTRACT	i
DECLARATION	ii
DEDICATION	iii
ACKNOWLEDGEMENTS	iv
LIST OF ABBREVIATIONS	v
LIST OF TABLES	vi
LIST OF FIGURES	vii
LIST OF FORMULAE	viii
CHAPTER ONE: INTRODUCTION	
1.1 Contextual Background	1
1.2 Project Background	5
1.3 Rationale	7
1.4 Research Aim and Objectives	8
1.5 Overview of the study	8
CHAPTER TWO: LITERATURE REVIEW	
2.1 Introduction	10
2.2 Importance of Vocabulary	10
2.2.1 Vocabulary and intelligence	11
2.2.2 Vocabulary knowledge and reading comprehension	11
2.3 The Construct of Vocabulary	13
2.3.1 Defining word	13
2.3.2 Issues with larger lexical items	16
2.4 Vocabulary Knowledge	17
2.4.1 Precision of word knowledge	17
2.4.2 Dimensions of vocabulary knowledge	18



2.4.3 Stages of vocabulary knowledge	19
2.5 Vocabulary Ability	20
2.5.1 The context of vocabulary use	20
2.5.2 Four core dimensions of vocabulary ability	21
2.5.3 Metacognitive strategies for vocabulary use	22
2.6 Measurement of vocabulary	22
2.6.1 Introduction	22
2.6.2 Test format and issues in the measurement of vocabulary	23
2.6.2.1 Objective language tests	24
2.6.2.2 Multiple-choice format	24
2.6.2.3 Matching format	25
2.6.2.4 Checklist formats	25
2.6.3 Alternatives in the designs of test format	26
2.6.3.1 Discrete-embedded	26
2.5.3.2 Selective-comprehensive	26
2.5.3.3 Context-independent-Context-dependent	27
2.7 Testing facets of vocabulary	27
2.7.1 Vocabulary size	27
2.7.1.1 Deciding on what counts as a word	28
2.7.1.1 Deciding which words to test	29
2.7.2 Quality of vocabulary	30
2.7.2.1 The role of context in assessing quality of vocabulary	30
2.7.2.2 Receptive and productive vocabulary	31
2.7.2.3 Recognition and recall	32
2.7.2.4 Comprehension and use	33
2.8 Validity in vocabulary tests	34
2.9 Summary	37



CHAPTER THREE: THEORETICAL FRAMEWORK OF EQUIVALENCE AND BIAS

3.1 Introduction	39
3.2 Equivalence	39
3.2.1 Levels of Equivalence	39
3.3 The Taxonomy of Bias	40
3.4 Bias in vocabulary tests	42
3.4.1 The Peabody Picture Vocabulary test	42
3.5 Research on Monolingual tests with reference to Bias and Equivalence	46
3.6 Conclusion	47

CHAPTER FOUR: METHOD

4.1 Introduction	48
4.2 Research Design	48
4.3 Sampling Procedure	48
4.4 Participants	48
4.5 Data Collection Tool	51
4.5.1 Adaptation and translation process of the adapted isiXhosa version of the WMLS	53
4.5.2 Psychometric properties of the WMLS	54
4.5.3 Psychometric properties of the PV Scale	54
4.6 Data Collection Procedure	55
4.7 Data Analysis	55
4.7.1 Introduction	55
4.7.2 Research objective 1	55
4.7.2.1 Factor analysis	55
4.7.2.2 Exploratory factor analysis	56
4.7.2.3 Tucker's phi coefficient	56
4.7.2.4 Executing the factor analysis	57



4.7.2.5 The reporting of the factor analysis	58
4.7.3 Research objective 2	59
4.7.4 Research objective 3	59
4.8 Ethical considerations	60
CHAPTER FIVE: RESULTS	
5.1 Introduction	61
5.2 Construct equivalence of the PV scale across the two groups	61
5.2.1 Steps in conducting the factor analysis	61
5.2.2 Construct Equivalence results with the DIF items included	62
5.2.2.2 The Tucker's Phi	62
5.2.2.3 Scatter plots of the Factor pattern coefficients with DIF items included	65
5.2.3 Construct Equivalence results of the factors with the DIF items removed	67
5.2.3.1 Factor Analysis results	67
5.2.3.2 Tuckers's phi with DIF items removed	70
5.2.3.3 Scatter plots of the factor pattern coefficients with the DIF items removed	70
5.3 Cronbach's Alpha of the factors after the deletion of the DIF items	72
5.4 Naming of the factors	73
5.5 Summary	74
CHAPTER SIX: DISCUSSION AND CONCLUSION	
6.1 Introduction	75
6.2 Discussion of the Results	75
6.2.1 Results of the EFA	75
6.2.2 Results of the Cronbach's Alpha per factor after the deletion of the DIF items	78
6.2.3 Implications of the findings	79
6.3 Limitations of the study	79
6.4 Conclusion	79
6.5 Recommendations for future research	80

REFERENCES

81

APPENDICES

94



ABSTRACT

The fall of apartheid and the rise of democracy have brought assessment issues in multicultural societies to the forefront in South Africa. The rise of multicultural assessment demands the development of tests that are culturally relevant to enhance fair testing practices, and issues of bias and equivalence of tests become increasingly important. This study forms part of a larger project titled the Additive Bilingual Education Project (ABLE). The Woodcock Munoz Language Survey (WMLS) was specifically selected to evaluate the language aims in the project, and was adapted from English to isiXhosa. Previous research has indicated that one of the scales in the adapted isiXhosa version of the WMLS, namely the Picture Vocabulary Scale (PV), displays some item bias, or differential item functioning (DIF), across rural and urban isiXhosa learners. Research has also indicated that differences in dialects can have an impact on test takers' scores. It is therefore essential to explore the structural equivalence of the adapted isiXhosa version of the WMLS on the PV scale across rural and urban isiXhosa learners, and to ascertain whether DIF is affecting the extent to which the same construct is measured across both groups. The results contribute to establishing the scalar equivalence of the adapted isiXhosa version of the WMLS across rural and urban isiXhosa-speaking learners. Secondary Data Analysis (SDA) was employed because this allowed the researcher to re-analyse the existing data in order to further evaluate construct equivalence. The sample of the larger study consisted of 260 learners, both male and female, selected from a population of Grade 6 and 7 learners attending schools in the Eastern Cape. The data was analysed by using the statistical programme Comprehensive Exploratory Factor Analysis (CEFA) and the Statistical Package for Social Sciences (SPSS). Exploratory factor analysis and the Tucker's phi coefficient were used. The results indicated distinct factor loadings for both groups, but slight differences were observed which raised concerns about construct equivalence. Scatter plots were employed to investigate further, which also gave cause for concern. It was therefore concluded that construct equivalence was only partially attained. In addition, the Cronbach's Alpha per factor was calculated, showing that internal consistency was displayed only for Factor 1 and not for Factor 2 for the rural group, or both factors for the urban group. Scalar equivalence across the two groups must therefore be explored further.

DECLARATION

I declare that “Exploring the Scalar Equivalence of the Picture Vocabulary Scale of the Woodcock Munoz language across rural and urban isiXhosa speaking learners” is my own work, that it has not been submitted before for any degree or examination at any other university, and that all sources I have used or cited have been indicated and acknowledged as complete references.

Qunita Brown



January 2012

Signed:

DEDICATION

I dedicate this thesis to the memory of my grandmother, Magdalena, the woman who raised and nurtured me and taught me the value of education. May you rest in peace, you will never be forgotten.



ACKNOWLEDGEMENTS

I would like to thank the following individuals for making this thesis possible:

My family, for all their support and encouragement during this challenging year. To my boyfriend, Clint, thank you so much for your patience, understanding and words of encouragement. You will never know how much it meant to me. Special thanks should go to my sister Carmelita, for standing by me and urging me on to complete this thesis. I would also like to thank all my friends who supported me especially Teza and Dale.

A special thank you to my supervisor Prof Elize Koch, for her guidance and valued contributions and for urging me on when it was needed. I loved working with such a passionate and dedicated researcher. Thank you for giving me the opportunity to learn from you.

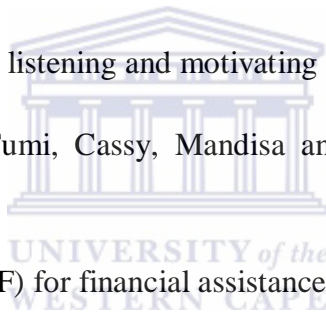
Shazly Savahl for understanding, listening and motivating me when I really needed it.

My M.A colleagues: Sabirah, Tumi, Cassy, Mandisa and Maya for your friendship and support.

The National Research Fund (NRF) for financial assistance.

Helen Allen for editing.

Maria Florence for providing feedback.



LIST OF ABBREVIATIONS

ABLE	Additive Bilingual Education Project
AAE	African American English
CEFA	Comprehensive Exploratory Factor Analysis
CFA	Confirmatory Factor Analysis
DIF	Differential Item Functioning
EFA	Exploratory Factor Analysis
ITC	International Test Commission
L1	First language speakers
L2	Second language speakers
LWI	Letter-Word Identification
PAF	Principal Axis Factoring
PCF	Principal Components Factoring
PPVT	Peabody Picture Vocabulary Test
PV	Picture Vocabulary
SDA	Secondary Data Analysis
SE	Standard English
SPSS	Statistical Package for the Social Sciences
VA	Verbal Analogies
WMLS	Woodcock Munoz Language Survey

LIST OF TABLES

		Page
Table 1	Distribution of participants per language group	49
Table 2	Distribution of participants per grade	49
Table 3	Distribution of participants per gender	50
Table 4	Description of Scales (WMLS)	52
Table 5	The pattern matrix loading for the rural isiXhosa dialect group	63
Table 6	The pattern matrix loading for the urban isiXhosa dialect group	64
Table 7	The Tucker's Phi coefficient per factor	65
Table 8	Pattern matrix loadings for the rural isiXhosa dialect group with the DIF included	68
Table 9	Pattern matrix loadings for the urban isiXhosa dialect group with the DIF items excluded	69
Table 10	The Tucker's Phi Coefficient per factor	70
Table 11	The Cronbach's Alpha for the two factors across the two dialect groups	72

LIST OF FIGURES

	Page
Figure 1: Gender differences between rural and urban isiXhosa learners	49
Figure 2: Gender differences between rural and urban isiXhosa learners	51
Figure 3: Scatter plot of the factor pattern coefficients for factor 1 across both rural and urban isiXhosa dialect groups	66
Figure 4: Scatter plot of the factor pattern coefficients for factor 2 across both rural and urban isiXhosa dialect groups	67
Figure 5: Scatter plot of factor pattern coefficients for factor 1 for both rural and urban isiXhosa dialect groups	71
Figure 6: Scatter plot of factor pattern coefficients for factor 2 for both rural and urban isiXhosa dialect group.	72



LIST OF FORMULAE

		Page
Formula 1	The Tucker's Phi	57



CHAPTER 1

INTRODUCTION

1.1 Contextual Background

According to Oakland (2004), tests are administered in many countries with participants ranging from newborns to the elderly. A psychological test is an objective and standardised measure of a sample of behaviour (Anastasi & Urbaniak, 1997). Tests are used to describe current behaviours and other qualities, to attempt to predict future behaviours, evaluate progress, aid in counselling and guidance (Oakland, 2004), identify education and training needs, and assist in making decisions regarding the placement of persons in jobs or a field of study, and are helpful when diagnosing disorders (Foxcroft & Roodt, 2009). Such tests are most often developed in Western contexts and standardised for Western populations, and then applied and used in non-Western contexts as well, notwithstanding several issues regarding their use across cultures and languages, such as fairness and bias.

In South Africa, policy and legislation have brought testing and assessment issues to the forefront (Claasen, 1997; Foxcroft, 1997; Meiring, Van de Vijver, Rothmann & Barrick, 2005). Assessment is essentially a process-orientated activity intended to gather a wide array of information by using assessment measures (tests) and other information from various sources (a person's history, interviews, etc.) (Foxcroft & Roodt, 2009). All the information derived from this process is then evaluated and integrated in order to reach a conclusion or make a decision. "Testing", which refers to the use of tests and measures, entails the measurement of behaviour, and is thus an essential element of the much broader evaluative process known as "assessment". This thesis deals mainly with the issue of testing. The rise of multicultural assessment as a result of political changes in South Africa, demands the development of tests that are culturally relevant, to ensure fair testing practices (Foxcroft, 1997). "Multicultural assessment" includes the testing of individuals from different cultural and linguistic backgrounds (Foxcroft, Roodt & Abrahams, 2009). Inherent in multicultural assessment are issues of bias and equivalence in tests. Test developers, test users and psychologists are legally obligated to provide evidence demonstrating that all assessment measures are fair (Kanjee & Foxcroft, 2009). These issues will be discussed in more depth in the literature section of this thesis. Language plays a huge role in these issues, particularly that of dialect differences in languages.

In terms of language and dialect, Hudson (1996) notes that there are two major differences between the two. Firstly, there is a difference in size because a language is larger than a dialect. In essence, a language contains more items than a dialect. For instance, people may refer to English as a language containing the sum total of all the terms in all its dialects, with Standard English (SE) as one dialect among many others (Indian English, Yorkshire English, etc.). Secondly, these two varieties differ as regards the prestige bestowed on each, in that a language has prestige while a dialect lacks it. If we were to apply this logic, it would seem that SE is not a dialect at all, but rather a language, whereas the varieties which are not utilised in formal writing are dialects. In other words, whether or not a variety is used in formal writing has a direct influence on its prestige. Unwritten languages in Britain are often known as dialects, irrespective of whether or not there is a “proper” language to which they are linked or related.

International research has indicated that dialect differences in many languages impact on language processing development and literacy development in children (Apel & Thomas-Tate, 2009; Green, 2002; Yiakoumetti, 2006). In Greece, there are differences between Cypriot dialect (CD) and Standard Modern Greek (SMG), and in the United States differences exist between African American English (AAE) and Standard English (SE) (Apel & Thomas-Tate, 2009; Yiakoumetti, 2006). In South Africa, dialect differences have also been found between Standard Afrikaans and Kaapse Afrikaans (Cape Afrikaans) (Deumert, 2002). These differences can be at the level of phonology and morphology, with significant differences at a lexical level. Vocabulary knowledge forms a large subsection of the lexicon (Apel & Thomas-Tate, 2009; Deumert, 2004; Green, 2002; Yiakoumetti, 2006). “Phonemes” are the units of sound, and each language has its own phonological rules which direct the combination of sounds (Doctor & Knight, 1993; Hoff, 2009). “Morphemes” are units which carry meaning, and “morphology” is the system of rules about combining the smallest units of language into words.

With regard to AAE and SE, researchers have shown that AAE has an impact on spelling, vocabulary, reading accuracy, word recognition, phonemic awareness, and morphological awareness in the reading development of African American children (Apel & Thomas-Tate, 2009; Johnson, 2005; Stockman, 2000). Possibly as a result of the differences in dialects, students who speak AAE tend to score lower than their Caucasian peers on national literacy assessments (Apel & Thomas-Tate, 2009; Green, 2002). A substantial amount of research pertaining to AAE and education has focused primarily on language and reading arts, but

recently interest has emerged in exploring AAE in the educational realm in relation to communication disorders (Green, 2002). More specifically, a large proportion of research on child AAE in this area is due to assessment issues or challenges that African American children were faced with in the absence of normative data for children acquiring AAE. Green (2002) notes that because no language assessment tests were designed for children who were acquiring AAE, their language would have to be assessed according to tests that were constructed for children using mainstream English. Obviously these tests were not used in accordance with the intended purpose, and thus ran the risk of showing that the speakers were not using language appropriate for their age group, when in all likelihood many of the speakers would have been using language approximately in accordance with the rules used in the AAE-speaking community.

According to Caltrax (1996), one of the language varieties that occurs in communal repertoires of literate societies, is the so-called “standard language”. “Standard language” refers to the formal written form of the language that is taught in educational institutions and used in publications and the media. To be more specific, a standard language can be defined as “a codified form of a language, accepted by, and serving as a model to, a larger speech community” (Garvin & Mathiot, 1968, p.365). In addition, whereas normal language development takes place in a rather disorganised way and is mainly below the threshold of consciousness of the speakers, standard languages are viewed as the result of a deliberate and direct intervention by society (Caltrax, 1996). This intervention then gives rise to a standard language, where before there were only dialects (non-standard varieties).

Hudson (1996) lists several stages that a language has to pass through to become a “standard”:

- a) *selection*: the first stage where a particular variety is selected and developed into a standard language.
- b) *codification*: where the linguistic features of such a variety must be written down in handbooks, dictionaries, grammatical forms, terminology and orthography. Orthographic rules are rules for combining letters in a meaningful way, and just like phonological rules, different languages have different orthographic rules (Doctor & Knight, 1993).

c) *elaboration of function*: when the scope of the use of the language is broadened, so that it is now used in schools, the media, religious activities and literature (Hudson, 1996).

d) *acceptance*: when the community has to accept the variety and incorporate it as its national language. Once this has occurred, the standard language serves as a unifying force of the state.

Non-standard languages differ markedly from the standard form in their manner of acquisition and their specialised functional roles (Hudson, 1996). As mentioned above, standard languages can only be taught or acquired formally at school or in adult literacy classes (Mansour, 1993). Pride and Holmes (1979) state that non-standard language cannot perform the same social functions that a standard one can; it will primarily be used in that specific local village or tribe. According to Caltreux (1996), the main non-standard language varieties are dialects.

There are 12 identified isiXhosa dialects (Webb, 2002). However, only two of these dialects are included in Standard isiXhosa that is used in the domains of education, religion and formal meetings (Caltreux, 1996). The standard dialect of isiXhosa is mostly spoken in rural areas in the central regions of the Eastern Cape. Thipa (1989) lists a few factors which can be regarded as main differences between urban and rural (standard) Xhosa, which are as follows:

- It appears that speakers of urban Xhosa display a greater tendency to borrow from Afrikaans and English than speakers of rural Xhosa.
- Urban Xhosa tends to be regarded as more “innovative” than rural Xhosa, which tends to be viewed as more conservative.
- Owing to the above-mentioned, urban Xhosa is more likely to undergo rapid changes than rural Xhosa.
- Lastly, rural Xhosa is characteristic of speakers who have been the least exposed to Western experiences and influences. The differences between the standard and urban varieties of isiXhosa are the main focus of this study.

Despite the acknowledgement that different dialects within a language have an impact on spelling, vocabulary and so on (Apel & Thomas-Tate, 2009; Green, 2002; Yiakoumetti, 2006), there are very few studies pertaining to how differences in the standard variety (used

in the rural setting) and non- standard variety (spoken in urban setting) of isiXhosa impact on students' reading and writing skills. In addition, issues of bias become important as these differences may not be acknowledged in tests measuring vocabulary. The discussion of dialect differences between AAE and SE has illustrated the dangers of not taking cognisance of this factor. If researchers were to ignore these significant differences and their impact on test results, a distortion of the results may occur (test scores may be erroneously attributed to low levels of vocabulary knowledge) or there may be a misdiagnosis of a language disorder which can severely affect the test-taker's life.

1.2 Project Background

This thesis forms part of a larger project entitled: The Additive Bilingual Education Project (ABLE) which was implemented in 2003 specifically for isiXhosa-speaking children in the rural areas of the Eastern Cape, South Africa (Koch, Landon, Jackson & Foli, 2009). The Woodcock Munoz Language Survey (WMLS) is a test assessing cognitive academic language proficiency, and was specifically chosen to research the language aims of the ABLE project. The test was adapted from English into isiXhosa after consulting the International Test Commission (ITC) guidelines for the adaptation of tests (ITC, 2002; Koch et al., 2009). Currently, the original version of this test is used extensively in the USA in evaluating bilingual programmes, and English and Spanish versions of the test have been constructed (Woodcock & Munoz-Sandoval, 2001).

ABLE has three main objectives (Koch et al., 2009). Firstly, researchers aim to assess the long-term effect of additive (or "late-exit transitional") bilingual curriculum delivery with regard to language proficiency in English and isiXhosa in particular. It examines the cognitive development and academic achievement of a group of isiXhosa-speaking learners from a rural area by comparing them to a group of learners from a similar contextual background, who have been exposed to subtractive bilingual education. Secondly, ABLE aims to describe currently what form additive bilingual curriculum delivery takes in practice in South Africa. Lastly, it intends to provide a description of the effect of this model on learners and teachers, on the school itself, and on the wider community associated with the school. Demonstrating the success of the ABLE project is contingent on the assessment tools that are used, and therefore it is pivotal to establish equivalence of the two versions of the test.

Researchers from the broader study obtained permission to adapt the original version into a South African English version as well as an isiXhosa version (Koch, 2009). To adapt the test into isiXhosa, a multilingual and multidisciplinary team was assembled, which consisted of language educators, accredited translators, bilingual English- and isiXhosa-speaking linguists, mainly monolingual English-speaking language educators, and a bilingual psychometric expert with a research psychology background. With regard to the processes involved in the test adaptation, the instrument was adapted during two workshops, one of which commenced at the end of 2004 and the other at the beginning of 2006. The main adaptation work occurred during the first workshop and subsequently the first round of data collection took place, where exploratory analyses pertaining to the equivalence of the two versions were conducted. Only after these analyses did the second workshop commence, largely aiming at applying changes to the test, based on the first round of results.

The team focused on identifying the underlying linguistic and psychological processes measured by the test, and taking the above-mentioned into account in the adaptation process. Because the main focus was on these processes, various items were re-written instead of only being translated (Koch, 2009). The strategies that were employed in the adaptation process included extensive relexification, which is “the translation of roots and use of totally different phrases, because of the lack of overlap in metaphors in the two languages or the lack of available words in the target language” (Koch, 2009, p.305). The team also paid attention to the grading of difficulty in the items, because all the scales of the original WMLS were graded in terms of item difficulty (*easy to difficult*). Because of this characteristic of the scales, a need arose to select other words instead of translating the items from English into isiXhosa (Koch, 2009).

In terms of challenges faced by the team with regard to the adaptation of the PV scale, the original version had several historical and culturally loaded items, including words that lacked equivalent isiXhosa words (Koch, 2009). Some of the solutions to this problem included making use of loan words as well as exchanging the culturally loaded English words with words that were loaded in favour of the Xhosa culture. Previous research has demonstrated that the adapted isiXhosa version of the PV scale displays relatively good internal consistency across rural and urban isiXhosa learners, with a Cronbach’s Alpha of .77 for both groups (Silo, 2010). In terms of internal consistency with regard to the adapted English and isiXhosa versions of the test, the Cronbach’s Alpha values range from .64 to .91 for the different scales. The Cronbach’s Alpha for the English version of the PV scale for

first-language speakers of English is .73, while it is .81 for the isiXhosa group (Koch, 2009). The psychometric properties of the adapted English and isiXhosa versions of the test are currently being established, and this study forms part of the psychometric evaluation of the WMLS.

1.3 Rationale

Currently not one of the versions of the WMLS has yet been normed for the South African population, as the research on the equivalence (this concept will be explained in detail in the third chapter) of the two versions of the test is in the process of being conducted (Koch et al., 2009). According to Poortinga (1989), equivalence essentially refers to whether test scores obtained can be compared in different cultural groups. In other words, the focus here is whether any scores obtained in different cultures can be compared. This research has produced promising results on two scales, Letter Word Identification (LWI) and Verbal Analogies (VA) (Arendse, 2009; Haupt, 2010).

Silo (2010) has investigated whether any item bias (this concept will be explained in detail in a later section), using logistic regression, was observed across rural and urban isiXhosa learners on the various scales of the adapted isiXhosa version of the WMLS. The results indicate that item bias has occurred, and that this threatens the scalar equivalence of the WMLS. Of particular interest with regard to this study is the PV scale, where six items displayed bias or DIF across rural and urban isiXhosa learners. Some of these results were attributable to dialect differences, such as the finding that on a few of the items, the rural isiXhosa learners performed slightly better than the urban isiXhosa learners. A possible explanation for this finding is that rural isiXhosa learners may have been more exposed to some of the pictures in this scale. In terms of items favouring urban isiXhosa learners, this may have been due to being more exposed to media (television) and the borrowing of English words into isiXhosa.

According to Hambleton, Marendia and Spielberger (2005) dialects within a language may impact on the validity of an adapted test. To control for this, it is essential to choose which dialect is of interest, or whether the aim of testing is to generate an adaptation that will apply across dialects within a language. Bekker (2005) state that the following factors may contribute to differences in performance across dialects: educational background; familiarity with tests and test-taking skills; familiarity with syntax and words in the native language; and

lastly, different degrees of proficiency and acculturation in the language of the test in members of the same cultural group.

When conducting DIF analysis with small samples, problems often arise (Robin, Sireci & Hambleton, 2003). Most notably, group ability differences and small groups within ability groups can lead to over-identification of DIF items and unstable results. It is thus essential to further explore the impact of DIF on the structural equivalence of the PV scale, and whether the removal of DIF items would improve construct equivalence, should it be a problem.

This study is important for two reasons, first, to contribute to a further understanding of the impact of dialect differences on tests of vocabulary, and secondly to make a contribution to the development of unbiased tests in the indigenous African languages.

1.4 Research Aim and Objectives

The study's overall aim is to assess the scalar equivalence of the adapted isiXhosa version of the picture vocabulary scale of the WMLS across rural and urban isiXhosa-speaking learners.

The specific research objectives are to:

1. Evaluate the construct equivalence with the DIF items included.
2. Evaluate the construct equivalence with the DIF items excluded.
3. Evaluate the Cronbach's Alpha of the factors after the deletion of the DIF items.

1.5 Overview of the Study

This chapter has aimed to introduce the reader to the contextual background of the study and how it will contribute to the broader project within which it is situated. Inherent in this discussion was the outlining of central concepts pertinent to the study, which specifically comprised a background to the ABLE project and the rationale for the investigation.

Chapter 2 will consist of the literature review. Owing to the emphasis on the PV scale of the adapted isiXhosa version of the WMLS, literature pertaining to vocabulary will be discussed.

Chapter 3 will outline the theoretical framework used in the study, namely the theory of equivalence and bias that guided the researcher during the course of this study.

Chapter 4 will comprise the methodology that was used to achieve the aims of the study. In particular, an exploration of the research design, sampling methods, instrument and data

collection procedures, statistical techniques employed in the study, and ethical concerns, will be provided.

Chapter 5 will contain the reporting of the findings according to each objective of the study.

Chapter 6 will comprise a thorough discussion of the findings of the study. Implications of the study as well as its limitations will be provided. Lastly, recommendations for future research will be made.



CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Owing to the fact that this study primarily deals with the PV scale of the WMLS, it is important to provide a clear understanding of what is being measured and how dialect differences can impact on vocabulary development. This chapter will contain three main sections. The first section is concerned with the importance of vocabulary and the links between vocabulary and intelligence, as well as vocabulary and reading comprehension. This will be followed by a discussion of the construct “vocabulary”. The construct is complex, and a thorough theoretical understanding of the construct has implications for its measurement. Lastly, issues pertaining to measurement of vocabulary are extensively discussed to unpack the complexity around the measurement of this construct.

2.2 Importance of Vocabulary

According to O’Rourke (1974), people communicate in a variety of ways, but most importantly, we communicate verbally through the use of language. Unlike a photograph, which is entirely representational, language is symbolic. Words are independent units of writing and speech, and vocabulary growth is not restricted to the periphery of the learner’s life, but is central to his/her development. When a new word is learned, it often produces a chain reaction, reminding the person of another word or situation and leading to a search for a new application of the word. Thus, a word may be viewed as not only the means of describing an experience, but also a motive for seeking an idea.

Vocabulary development is a vital part of every student’s life. Not only does it affect their thoughts, aspirations and actions, but more crucially, their success. Generally speaking, success with words equals success in many areas, particularly in academic achievement. Vocabulary development is of the utmost importance because it allows people to exchange ideas, and aids in acquiring new experiences. Developing one’s vocabulary allows one to see conceptual relationships, by putting handles on objects and ideas to enable effective manipulation. Thus, people’s ability to name things greatly influences the extent of their cognitive skills (O’Rourke, 1974).

2.2.1. Vocabulary and intelligence

The strong association between vocabulary and general intelligence is one of the most robust findings in the history of intelligence testing (Anderson & Freebody, 1981). In a seminal study, one of the first of its kind, Terman (1918) reported a correlation of .91 between mental age and intelligence quotient (IQ) (as assessed by the Stanford Revision of the Binet-Simon scale) and the vocabulary subscale. He suggested that the vocabulary measure alone constitutes a good estimate of the performance on the entire scale, and can be used as a short measure. Following Terman's suggestion, various studies have been conducted with different age groups to explore the link between vocabulary and intelligence (Anderson & Freebody, 1981). These studies reported correlations between vocabulary subtest scores and total test scores on various different IQ and achievement tests, ranging from .71 to .98 (Elwood, 1939; Mahan & Witmer, 1936; Lewinski, 1948; Raven, 1948; Spache, 1943). These findings led to scholars realising that vocabulary is important, and many tests have been devised to measure this construct (Read, 1997). These tests are often used to assess verbal ability and provide an estimation of general intelligence (Chan, Cheung, Sze, Leung & Cheung, 2008). Consequently, vocabulary scales are often included in popular tests of intellectual functioning, such as the Wechsler and Stanford Binet intelligence scales (Chan et al., 2008). More specifically, the findings of older studies examining this association have indicated that vocabulary test results correlate highly with those of mental development and reading (O'Rourke, 1974).

2.2.2 Vocabulary knowledge and reading comprehension

Studies have shown that vocabulary is a key variable in reading comprehension (Anderson & Freebody, 1985; Beck, Perfetti & McKeown, 1982; Coyne, Simmons & Kame'enui, 2004a; Cunningham & Stanovich, 1997; Klare, 1975; Quian, 2002; Raptis, 1997; Stahl & Nagy, 2006; Tannenbaum, Torgesen & Wagner, 2006). According to Ricketts, Nation and Bishop (2007), "reading comprehension" refers to the ability to understand connected texts. Stahl and Fairbanks (1986) examined the findings of 41 studies that dealt with the impact of vocabulary instruction on comprehension. These results revealed an average effect size of .91 for vocabulary, which theoretically would raise the comprehension of an average child from the 50th percentile to the 83rd percentile. In terms of predictive power, subsequent studies have demonstrated that vocabulary knowledge predicts not only listening but also reading comprehension performance, with positive correlations ranging from .6 to .8 (Pearson,

Hiebert & Kamil, 2007). In addition, Jalongo and Sobolak (2011) acknowledge that this association has long-lasting implications for students at both high and low vocabulary levels.

Many researchers have postulated that the relation between vocabulary and reading comprehension is reciprocal across development, because reading provides an opportunity to learn new word meanings (Beck, Perfetti & McKeown, 1982; Verhoeven & Van Leeuwe, 2008). Reading comprehension can only be beneficial or successful when word forms are readily identified and word meanings are easily assessed, which places considerable demands on the underlying linguistic capacities of the child (Verhoeven & Van Leeuwe, 2008). The existing literature indicates that children with poor reading comprehension tend to show relatively low levels of vocabulary knowledge (Nation, Clarke, Marshall & Durand, 2004), and they are not skilled at using the text to infer the meanings of new words (Cain, Oakhill & Lemmon, 2004). Ricketts, Nation and Bishop (2007) note that data pertaining to studies of children with reading comprehension impairment provides a rich source of evidence regarding the important role that vocabulary plays in reading development. More specifically, children who are characterised as poor comprehenders (children who are defined as having age-appropriate reading accuracy skills, but have specific difficulty with reading comprehension) demonstrate weaknesses in listening comprehension and vocabulary (Catts, Adlof & Weismer, 2006).

A recent longitudinal study was conducted with a representative sample of 2143 Dutch children throughout the elementary school period, to ascertain the specific effects of word decoding, vocabulary and listening comprehension abilities on the development of reading comprehension (Verhoeven & Van Leeuwe, 2008). One of the specific aims was to test two theoretical frameworks for the prediction of reading comprehension, namely; the lexical quality hypothesis and the simple reading view. The lexical quality hypothesis states that the development of a child's reading comprehension will be supported by their word knowledge, which includes the precision of their phonological, orthographic and lexical-semantic representations in addition to the sheer number of words that they know. "Semantics" refers to how language conveys meaning (Doctor & Knight, 1993). In terms of the lexical quality hypothesis, vocabulary and word decoding are assumed to be critical determinants of reading comprehension (Verhoeven & Van Leeuwe, 2008). The simple reading view states that listening comprehension, in addition to word decoding skills, leads to the development of reading comprehension. Thus, in this study it was postulated that reading comprehension was

a product of listening comprehension, word decoding and listening comprehension in addition to word-decoding skills would predict the development of reading comprehension.

The results provided empirical support for both hypotheses (Verhoeven & Van Leeuwe, 2008). In particular, for each of the linguistic variables involved in the study, significant progress from one to the next was consistently observed. In terms of the stability of the measures, these were found to be high across time, which suggests that individual differences between students remain across grades. Despite the support for both hypotheses, the findings reflect that in subsequent grades, vocabulary is an important variable that predicts reading comprehension directly, whereas listening comprehension demonstrates a reciprocal relationship with vocabulary. Thus, it would seem that familiarity with the words in a text can substantially facilitate reading, and conversely, that skilled comprehension of a text can result in vocabulary growth (Perfetti & Hart, 2001). The findings of this study and previous studies have emphasised that vocabulary is pivotal in reading comprehension.

2.3 The Construct of Vocabulary

Before considering how to test vocabulary, it is imperative to explore the nature of what one wants to assess. In this section the nature of this construct will be explicated to demonstrate the many issues that arise in the measurement of vocabulary. In addition, this section will illustrate the complexity of vocabulary.

It is commonly assumed that vocabulary is an inventory of individual words with associated meanings. From this viewpoint, vocabulary knowledge would entail knowing the meanings of words, and therefore the aim of a vocabulary test would be to ascertain whether learners can match a word with a synonym, a dictionary-type definition or an equivalent word in their own language. Current developments in applied linguistics and language teaching, however, have created the need to address many questions relating to vocabulary, such as what a word is, as well as the issue of lexical items (Read, 2000).

2.3.1 Defining words

The definition of “word” is the first question that needs to be addressed (Read, 2000). Not only does this interest many linguists on a theoretical level, but in terms of testing, many practical reasons exist for asking this question. The major issues with regard to what constitutes a word are primarily due to the intricacies of base words, word families and the

concept of homographs. These issues all have major implications for testing, which will become apparent later in the chapter.

One important distinction in order to understand the definition of “word” is the distinction between “*tokens*” and “*types*” which applies to the counting of words in a text (Read, 2000). The number of “*tokens*” is the same as the number of word forms, which means that individual words occurring more than once get counted each time that they are used. The numbers of “*types*” are the total number of the different word forms; in other words, a word that is repeated many times (in many forms) is counted only once. In addition, the relative proportions of types and tokens, which is termed the *type-token ratio*, is a widely used measure of the language development of not only native speakers of a language but also language learners. The relevance of demarcating these differences will become clearer in the following paragraphs where examples will be given in order to illustrate how this distinction can impact on testing and ultimately on test scores.

Words like *a*, *to*, *and*, *the*, *in* and *that* lead to questioning whether they can be regarded as vocabulary items (Read, 2000). These words, namely pronouns, conjunctions, articles, prepositions, auxiliaries and so on, are known as “function words” and are viewed as belonging more to the grammar of a language than to its vocabulary. However, unlike content words, adjectives, nouns, “full” verbs and adverbs have little if any meaning in isolation, and mainly serve to provide links within sentences, to modify the meanings of words and so on. This has implications for testing, because what is generally assessed in vocabulary tests is the knowledge of specific content words. Read (2000) notes that even if researchers or test developers were to restrict their attention to content words, another problem would be the fact that these words come in a variety of forms. For example, we have the word *society*, but there are also *societies*, *society’s* and *societies’*. In this instance, we would normally regard these as different forms of the same word. In addition, grammatically speaking, what is involved is adding inflectional endings to a base form, without changing the meaning or the word class of the base form. The base and inflected forms are referred to as the “lemma” (Cooper & Van Dyk, 2003). When a study is undertaken that involves counting the number of words (in the sense of types) either in a spoken or written text, researchers normally lemmatise the tokens, in order for the inflected forms to be counted as instances of the same lemma as the base form.

According to Cooper and Van Dyk (2003), not only do base words take on the form of inflectional endings, but also a variety of derived forms, which results in a change to the word class, therefore adding a new element of meaning. For instance, consider the word *leak* with inflected forms such as *leaks*, *leaking* and *leaked* as well as derivatives: *leaky*, *leakiness*, *leakage* and *leaker* (Read, 2000). Even though a distinction exists between the literal “loss of a fluid” and the more metaphorical “loss of secret information”, all these words can be considered to be closely related in form and meaning. Such a set of word forms that share a common meaning is referred to as a “word family” (Cooper & Van Dyk, 2003). Read (2000) maintains that the situation becomes extremely complex when dealing with a word such as *society* and the many other words that resemble it to varying degrees in meaning and form: *social*, *socially*, *sociable*, *unsociable*, *sociability*, *socialise*, *socialisation*, *socialism*, *socialist*, *socialistic*, *socialism*, *sociology* and so on. All these words share the same *soci-* form, and it would appear that they possess a common underlying meaning. However, collectively, these words express quite a range of meanings, and we cannot assume that they are members of the same word family. Thus, the problem lies in how to separate them into word families, which has obvious implications for testing.

According to Bauer and Nation (1993), the issue of distinguishing word forms and word families becomes crucial in relation to measures of vocabulary size. This is apparent in the widely varying estimates of how many words a native speaker knows, because some researchers are counting word forms while others are focusing on word families. Even if the test is not designed to measure vocabulary size, another question that arises is what exactly can be inferred from learner performance on particular test items. For example, consider this: an item is constructed to assess knowledge of the word *critical* and many learners answer it correctly. Do researchers then credit the learners with knowing just the word *critical*, or is it reasonable to assume that they also know the words *crisis*, *critically* and *criticism*? In other words, what is actually being assessed? Is it the individual word form or perhaps the whole word family to which that word belongs?

Read (2000) states that the process of what exactly constitutes a word gets more complex with the introduction of homographs. These are single word forms that have at least two meanings that are so different that they obviously belong to different word families. One example often cited is the word *bank*, which has two major meanings (an institution that provides financial services, and the sloping ground beside a river). What is apparent is that no underlying meaning exists that can usefully link these two definitions, so we can assume that

we are dealing with distinct word families. The complexity of homographs has highlighted that in the testing context, researchers cannot just assume that simply because learners have demonstrated knowledge of one meaning, they have acquired any of the other meanings.

This discussion of the definition of a word in the English language has emphasised that the term “word” can refer to a variety of lexical units (Read, 2000). The complexity of what constitutes a word has obvious implications for testing, because scholars and researchers tend to use different criteria with regard to the issues just discussed. It may appear that certain assessment procedures call for a clearer definition of what the relevant unit is, and explicit criteria for distinguishing one unit from another, which would dramatically aid in the testing of vocabulary.

2.3.2 Issues with larger lexical items

Another crucial point with regard to vocabulary is that it consists of more than just single words. According to Read (1997), there are phrasal verbs (such as *move out*), and compound nouns (*fire fighter*) which are viewed as lexical units, that consist of more than one word form. Then there are idioms (e.g. *let the cat out of the bag*), and in studies of second-language (L2) acquisition, these sentences and phrases cause great difficulty for learners mainly because the whole unit has a meaning that cannot be worked out by just knowing what the individual words mean.

These multi-word items have been recognised as playing an important part in vocabulary learning (Read, 1997). Many scholars have highlighted that fluent speakers and writers have a large amount of other kinds of “prefabricated language” at their disposal (Read, 2000) Pawley and Syder (1983) argue that the ability to speak fluently is contingent upon acquiring thousands of memorised sentence stems and whole sentences that are lexicalised to varying degrees. In addition, Pawley and Syder (1983) maintain that memorised sentences and phrases are the normal building blocks of fluent discourse, and yet simultaneously they provide models for constructing many new sequences that become memorable and as a result enter the stock of familiar usages.

Sinclair (1991) has attempted to amalgamate the aforementioned perspectives by postulating that two principles are needed to provide an adequate explanation of how texts are constructed. The first principle is known as the “open-choice principle”. This view is essentially based on Chomsky’s work, and regards sentences as being creatively produced on

the basis of an underlying system of rules. Sentences then contain slots that can be filled by a wide range of possible words, depending largely on the language user's choice. Despite this, Read (2000) acknowledges that a large body of corpus research has demonstrated that in practice, lexical choices are more limited than expected if only open-choice principle was operating. What normally happens is that words commonly come together in combinations of two, three, four or more, that appear to form relatively fixed expressions known as "collocations". When Sinclair (1991) became aware of this, he maintained that the open-choice principle should be complemented by the "idiom principle", which he explained as follows: "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments" (Sinclair, 1991, p.110). According to Read (2000), this view helps to explain why very frequent content words such as *take*, *make* and *get* appear to contribute very little specific meaning of their own, but have to be understood in relation to the entire phrases in which they occur. Sinclair (1991) thus believes that while linguists have traditionally relied on the Chomskian view as the basis for their work, the idiom principle at least deserves some attention in the construction and interpretation of texts. Existing literature has shown that linguists have not paid enough attention to lexical phrases, and as a result they are not well researched or documented (Read, 2000). It is only through the surge of interest in discourse analysis and corpus linguistics that the significance of lexical phrases has been acknowledged.

2.4 Vocabulary Knowledge

Owing to the complexity of vocabulary knowledge, it is not surprising that in studies pertaining to vocabulary knowledge, different scholars emphasise different aspects, which leads to confusion in the literature (Read, 1997). In this section, the varying aspects of vocabulary knowledge will be discussed, namely: precision of word knowledge; the dimensions of vocabulary; and the stages in the development of vocabulary knowledge.

2.4.1 Precision of word knowledge

Dolch and Leeds (1953 as cited in Read, 1997) use items which they claim assess precision of word knowledge, which relates to what the test-takers know about the specific meaning of each target word, rather than merely having a vague idea about it. This might represent one way of defining vocabulary knowledge, but the drawback is that this approach assumes that each word has only one meaning which has to be known precisely. Read (2000) notes that

words often have several different meanings; for instance, consider the word *fresh* as in *fresh bread*, *fresh ideas*, *a fresh breeze*, *fresh supplies* and so on. If we were to take this aspect into consideration, we would need to add a dimension of range of meaning in addition to precision.

To take it further in terms of conceptualisation, vocabulary knowledge therefore involves more than simply word meaning (Read, 2000). There are multiple components of word knowledge which include pronunciation, spelling, grammatical form, frequency, collocations and restrictions on the use of the word, as well as the differences between receptive and productive knowledge, which will be discussed later in the chapter.

Cronbach's (1942) framework added to the complexity because he devised a way to analyse the scope of vocabulary knowledge. In particular, he referred to five types of behaviour involved in understanding a word, which are as follows: 1) *generalisation* (being able to define the word); 2) *application* (selecting an appropriate use of the word); 3) *breadth of meaning* (recalling the various meanings); 4) *precision of meaning* (applying the use of the word correctly to all possible situations; and lastly, 5) *availability* (being able to use the word productively).

2.4.2 Dimensions of vocabulary knowledge

Henriksen (1999) has tried to provide some clarification about the issue of what constitutes knowledge of words, by proposing that researchers or scholars should recognise three distinct dimensions of vocabulary knowledge. The first dimension is known as “partial-precise” knowledge. Many researchers have emphasised that the learner must be allowed to be vague about the meaning of a word at first. Precision will come later, and lexical development can then be viewed as moving or progressing from rough categorisation or vagueness to more precision and mastery of the meaning. Brown (1994) notes that in the process of acquiring word meaning, the learner's knowledge of a specific lexical item moves from simply recognising the word (word recognition) through different degrees of partial knowledge towards precise comprehension. Researchers have to recognise and emphasise that no first-language (L1) speaker will ever develop an exhaustive knowledge of a word's meaning potential (Henriksen, 1999).

The second dimension is known as “depth” of knowledge. This semantisation process involves a progression along both dimension one (partial-precise knowledge) and dimension

two (depth of knowledge). This development along dimension one is mainly linked to the mapping process (creating extensional links through the use of labelling and packaging), whereas dimension two is primarily associated with network building (creating intentional links).

The last dimension is known as “receptive-productive” (Henriksen, 1999). The major distinction to note here is between having some knowledge of a word and being able to use it in writing and speech. This dimension is often viewed as a continuum, although many difficulties arise in defining how and exactly at what point words become available for productive use (Read, 2000). The issues pertaining to the receptive-productive continuum will be discussed later in the chapter. Read (2000) notes that Henriksen’s proposed conceptual framework is useful in providing a better basis for conceptualising vocabulary knowledge and for dissecting what aspects of the construct are actually being measured in specific research studies.

2.4.3 Stages of vocabulary knowledge

A more developmental approach to describing vocabulary knowledge was postulated by Dale (1965). Several researchers of L1 vocabulary have constructed scales which represent the varying degrees of partial knowledge that people can have with regard to the meaning of words that they know (Read, 2000). Dale’s (1965) basic four stages of word knowledge have been described as being extremely useful in conceptualising the various stages of vocabulary knowledge, and are discussed as follows: *Stage 1*: Never having seen the term before; *Stage 2*: Knowing that there is such a word, but not knowing what it means; *Stage 3*: Having context bound and a vague knowledge of the word’s meaning; *Stage 4*: knowing the word well and remembering it. Dougherty-Stahl and Bravo (2010) note that the final stage of Dale’s conceptualisation can be broken down further into additional stages, comprising the ability to name other words that are related to the word under consideration, and having precise versus general knowledge.

2.5 Vocabulary Ability

According to Chapelle (1994 as cited in Read, 2000), “vocabulary ability” refers to knowledge of language plus the ability to put language to use in context. This theorist developed a framework of vocabulary ability which has implications for testing, and identified three components of vocabulary ability, namely: 1) The context of vocabulary use;

2) Four core dimensions of vocabulary ability; and 3) Metacognitive strategies for vocabulary use.

2.5.1 The context of vocabulary use

As regards context and vocabulary, the whole text in a reading forms the context that people draw on, in order to interpret the individual lexical item within it (Chapelle, 1994 as cited in Read, 2000). However, if we were to look at context from a different perspective, more specifically from a communicative point of view, context would consist of more than just a linguistic phenomenon. Attention should be given to Bachman and Palmer's (1996) various types of pragmatic knowledge (refer to the section dealing with metacognitive strategies), because understanding exactly what is meant by "vocabulary ability" will only be strengthened if we draw on this knowledge. In particular, focus should be given to the social and cultural situation in which lexical items are used, because this can significantly influence their meaning. In terms of the above-mentioned, there are three ways in which context can dramatically affect lexical meaning, namely: 1) when there are differences across generations and between more formal and colloquial uses of words; 2) differences in interpretation across language varieties; and 3) when there are differences between everyday usage and more specialised terminology in specific fields of study. This section has illustrated that the influence of context is very important, and a definition of vocabulary ability will be enhanced if context is added into the equation.

Chapelle (1994 as cited in Read, 2000) also uses Halliday and Hasan's (1989) systemic linguistic theory, which looks at context from a more social perspective rather than a purely linguistic one. This theory maintains that context includes three complex elements, namely field, tenor and mode, which can be utilised to analyse the way in which features of spoken or written language relate to aspects of the social situation in which the language is being used. "*Field*" refers to the type of activity in which the language users are engaged, as well as the subject matter involved. "*Tenor*" refers to the relative social status of the language users and their role relationship. "*Mode*" encompasses the channel of communication, and in particular the features that distinguish writing from speech. According to this theory, the kind of vocabulary ability that learners need for reading a newspaper at home is very different from that required for listening to a chemistry lecture in a classroom. This has obvious implications for testing, for the above-mentioned have illustrated that scholars, especially test developers,

should take cognisance of context at a social level because this influences the kind of vocabulary ability that learners draw from when their vocabulary ability is being assessed.

2.5.2 Four core dimensions of vocabulary ability

The second component of Chapelle's (1994 as cited in Read, 2000) framework of vocabulary ability has received the most attention from L2 teachers and applied linguists alike. Four dimensions have been outlined by Chapelle (1994 as cited in Read, 2000):

- **Vocabulary size**

This basically refers to the number of words that a person knows (Chapelle, 1994 as cited in Read, 2000; Cooper & Van Dyk, 2003). In studies of L1 acquisition, scholars' attempts to measure the total size of native speakers normally involved taking a sample from a large unabridged dictionary (Read, 2000). When working from a communicative approach to vocabulary ability, it is essential to strive not to just seek to measure vocabulary size as an absolute size, but rather to measure it in relation to particular contexts of use (Chapelle, 1994 as cited in Read, 2000).

- **Knowledge of word characteristics**

It would be logical to assume that, just as L1 speakers do, L2 learners know more about some words than others (Read, 2000). More specifically, their understanding of particular words may range from vague to more precise (Cronbach, 1942). In addition, learners are likely to have some confusion about certain words that they have learned because the words might share common features (Laufer, 1990 as cited in Read, 2000). Once again, just as with vocabulary size, the extent to which a learner knows a word varies according to the context in which the word is used (Read, 2000).

- **Lexicon organisation**

This dimension deals largely with the way in which words and other lexical items are stored in the brain (Chapelle as cited in Read, 2000). In terms of this dimension, Meara (1984) states that researchers need to pay more attention to developing tests that investigate the developing lexicon of L2 learners and the ways in which their lexical storage differs from that of L1 speakers.

- Fundamental vocabulary processes

The last dimension refers to the specific processes that people use in order to gain access to their knowledge of vocabulary, not only for understanding but also for writing and speaking, such as lexical inferencing (Chapelle, 1994 as cited in Read, 2000). Psycholinguists have played a significant role in identifying a substantial number of processes of this kind. The data reflects that these processes operate more quickly and more automatically for native speakers than for less proficient learners, who not only have gaps in their knowledge of L2 words, but their mental lexicon might also not be as efficiently organised (Read, 2000).

2.5.3 Metacognitive strategies for vocabulary use

This is the last component of Chapelle's (1994 as cited in Read, 2000) framework and is what Bachman (1990) defines as "strategic competence". These are strategies that are used by all language users to enable them to manage the ways that they use their vocabulary ability in communication. It is important to note that it is only in situations when people have to undertake unfamiliar or cognitively demanding communication tasks that these strategies become conscious or predominant. Learners have a strong need for metacognitive strategies in certain communication situations because they have to overcome their lack of vocabulary ability in order to function effectively (Bachman, 1990). Chapelle's (1994 as cited in Read, 2000) framework of vocabulary ability has enabled scholars to move towards a definition that encompasses a wider range of testing purposes and at the same time is in line with Bachman and Palmer's (1996) general construct of language ability.

2.6 Measurement of vocabulary

2.6.1 Introduction

According to Read (2000), some of the most groundbreaking research in the measurement of vocabulary was done by vocabulary acquisition researchers as opposed to language testers.

Language testers seem to have neglected vocabulary tests and focused more on integrative and communicative measures of language proficiency. Other notable contributors to the understanding of vocabulary measurement are reading researchers with an emphasis on reading English as a first language. In this section, some of the issues with regard to the measurement of vocabulary resulting from this body of work will be explored.

2.6.2 Test format and issues in the measurement of vocabulary

In order to grasp the complexity of issues regarding the measurement of vocabulary, it is first necessary to discuss conventional test formats, because many of the present issues around vocabulary testing are directed at the perceived inadequacies of conventional vocabulary tests, particularly the formats of these tests.

2.6.2.1 Objective language tests. The rise of testing vocabulary in the school setting is inextricably linked to the development of objective testing, especially in the United States (Read, 1997). When learning material is divided into small units, each of which can be assessed through supplying a test item with a single correct answer that can be specified in advance, this is known as an “objective test” (Anastasi & Urbina, 1997). The most common form that these items take is the multiple-choice format. These tests are also regarded as objective because they can be scored without requiring any judgement on the part of the scorer as to whether the response is correct or not. Spolsky (1995 as cited in Read, 2000, Schmitt, 1999) outlines how psychometrics gave rise to objective testing, and how this became a dominant influence on all aspects of the American school curriculum during the period just after the First World War. These tests became so popular that they progressively replaced traditional essay examinations from the 1930’s onward (Read, 1997). Vocabulary was assessed by presenting students with a list of foreign words requiring them to match them with their English translations. Other earlier tests used multiple-choice items in a similar way, by providing a L2 word in the stem and four or five words as the options.

The following are the advantages of objective language tests for vocabulary:

- Words can be treated as separate independent linguistic units with a meaning expressed by a synonym, a translation equivalent or a short defining phrase (Read, 2000). Because of this, it was relatively easy to construct a set of multiple-choice items consisting of a word followed by four or five possible meanings, or a matching test consisting of short definitions, or jumbled lists of words.
- Psychometric theory demands tests that are proven to maintain excellent technical characteristics (Anastasi & Urbina, 1997). Multiple-choice tests proved to do just that (Read, 2000). In particular, well-written items could effectively discriminate among learners based on their level of ability, and therefore these tests were extremely reliable.

- According to Anderson and Freebody (1981), objective vocabulary tests did not simply measure vocabulary, but were also useful and valid indicators of language ability in a broad sense.

Owing to these obvious advantages, it was not surprising that objective vocabulary tests became a significant part of the discrete-point approach to L2 testing, where testing of students was focused mainly on their knowledge of individual structural elements of the language (Read, 2000). In addition, many authors supported this approach and recommended the use of objective test items such as blank filling, matching, multiple-choice, picture labelling and word translation (Read, 1997).

The following are the different item formats that are available in objective tests:

2.6.2.2 Multiple-choice format. The multiple-choice item format has been the most widely used method, not just for L1speakers but for L2 learners as well (Read, 2000). However, in Wesche and Paribakht's' (1996) article, several criticisms of this format were outlined. These are as follows:

- They can sometimes be difficult to construct, and require intensive field-testing, analysis and refinement.
- The student may know another meaning for the word, but not the one required by the test.
- The student may pick the right word by a process of elimination, and has in any case a 25% chance of guessing the right response in a four-alternative format.
- These items may assess test learners' knowledge of distracters rather than their ability to correctly identify the exact meaning of the target word.
- It is likely that the learner could miss an item because there is a lack of knowledge or understanding of syntax in the distracters.
- This format allows only a very limited or restricted sampling of a learner's total vocabulary.

Despite these criticisms, Wesche and Paribacht (1996) acknowledge that multiple-choice items will continue to be a first choice for many test developers, not only for vocabulary, but for other aspects of language proficiency as well, primarily because they are very convenient to administer, and many well-established procedures exist for developing them.

Notwithstanding their widespread use, ongoing research with regard to these tests is being conducted, particularly in the field of L2 learning.

2.6.2.3 Matching formats. According to Cooper and Van Dyk (2003), “matching” entails selecting the test item that closely correspond to an appropriate synonym or suitable definition from a number of options. Read (2000) states that the matching format and those formats that require a translation or equivalent word, are simpler to construct than multiple-choice ones, but these formats represent a low level of word knowledge. In addition, these formats do not provide any indication of whether the students will understand a word when they come across it in use, particularly if it has a different meaning from the one they have learnt.

2.6.2.4 Checklist formats. The yes/no checklist requires learners merely to indicate whether a word is known (by means of simply marking it with a tick) (Cooper & Van Dyk, 2003). This format is either validated by the inclusion of non-words in the corpus, or by asking the learners to define a sample of known words. Read (2000) notes that the checklist format is really the simplest possible format for testing vocabulary, and it has been utilised with L1 speakers at least since 1890. One of the major criticisms directed at the checklist in its classic form is that there are no means of ascertaining how validly the test-takers are reporting their knowledge of the words. For example, a test-taker can have a different idea from the researcher as to what “knowing a word” means and may be genuinely mistaken about certain words, by confusing one with another and so on. Despite these criticisms, the checklist format is obviously quite appealing to researchers who focus on vocabulary size, because this format enables the researcher to present the learner with several hundred words in order to maintain a reliable basis for making his or her estimates. In addition, for some purposes, the checklist format provides satisfactory results. However, in other situations, especially when students are assessed individually rather than being participants in a study, it is pivotal to find some direct evidence of whether or not the words are indeed known in some sense. This can be attained by using formats such as the matching or multiple-choice formats.

Overall, it appears that obtaining a good estimate of vocabulary size is a rather complex task. In particular, at all three levels, namely defining the units to be counted, selecting a sample and deciding upon a test format, challenging issues or questions need to be addressed before a really reliable measure can be obtained (Read, 2000).

2.6.3 Alternatives in the designs of test format

In this section the various possibilities in test format designs for vocabulary tests will be discussed, specifically in view of the issues surrounding objective tests as discussed above. The work on vocabulary measurement by Read (2000) is quite useful and important because he has outlined the different alternatives of measurement designs. He states that these designs should be considered dichotomies. “Dichotomy” refers to the division into two usually contradictory parts or opinions (American Heritage Dictionary of the English Language, 2000). Read’s (2000) alternatives to consider consist of 1) discrete-embedded, 2) selective-comprehensive and 3) context-independent-dependent .

2.6.3.1 Discrete-embedded. According to Read (2000), discrete tests are tests which measure vocabulary as a separate construct from other aspects of language ability. In other words, a discrete test is designed to view vocabulary as a distinct construct, separated from other components of language competence (Read, 1997). In contrast, an embedded vocabulary design refers to a design that contributes to the assessment of a larger construct (Read, 2000). An example of an embedded design is found in reading tasks consisting of a written text followed by a set of comprehension questions. In this case, the vocabulary item scores are not counted separately; in other words, they simply form part of the measure of the learners’ reading comprehension ability. It is important to note that many tests with a discrete design do require the learners to respond to words which are presented in a short sentence or in isolation, but this is not what makes this design discrete. Rather, it is the fact that this design is primarily emphasising the construct of vocabulary knowledge.

2.6.3.2 Selective-comprehensive. A selective measure is based on a set of target words specifically selected by the test-writer, and the test-takers are assessed according to how well they have shown their knowledge of the meaning or use of those words (Read, 1997). Test items of this kind are often constructed because they are easy to distinguish between and count, whether manually or by computer, and good resources are often available for test-writers to draw on, usually in the form of word-frequency lists and dictionaries. In addition, the target words may be selected as individual words and then incorporated into separate test items, or the test-writer can first choose a suitable text and then use certain words from it as the basis of the vocabulary assessment (Read, 2000).

On the other hand, a comprehensive design refers to a design that takes cognisance of all the vocabulary content of a written or spoken text (Read, 1997). For example, consider this: a

speaking test is administered to students, in which they are rated on various criteria, including their range of expression. In this case, the raters are not listening for particular words or expressions, but in principle they are focusing on forming a judgement of the quality of the test-taker's overall vocabulary use.

2.6.3.3 Context-independent: context-dependent. A design that is context-dependent is one in which the test-taker requires some understanding of the context in order to be confident in choosing the right option (Read, 1997). This researcher notes that if lexical phrases have to be evaluated in terms of their suitability for the social context, it hardly makes any sense to assess them in the isolated, context-independent way in which individual words have traditionally been assessed.

From the above discussion, it would seem that when deciding on what direction to take in terms of measuring vocabulary, the researcher or linguist needs to consider the purposes of the measure. This will aid them in design, because many of the previous issues discussed (the construct of vocabulary) such as what is a word, as well as larger lexical items, can affect which design is preferred and ultimately the interpretation of the results.

2.7 Testing facets of vocabulary

In this section, the testing of facets of vocabulary will be elaborated upon. These include vocabulary size and the quality of vocabulary.

2.7.1 Vocabulary size

Vocabulary size can be defined as the number of words that a person knows (Cooper & Van Dyk, 2003; Read, 1997, 2000). According to Read (2000), the number of words a learner knows is important in educational settings, and because of this, many studies have been conducted that focused specifically on vocabulary size. Four major reasons for this are outlined below.

- Reading researchers are very interested in estimating the vocabulary size of native speakers of English as they progress from childhood to adulthood (Anderson & Freebody, 1981). This research mainly has to do with the association between vocabulary and reading comprehension. The results of such studies are pivotal for it has significant implications with regard to the way that reading programmes are designed and implemented.

- By estimating native speaker vocabulary size at different ages enables teachers to be provided with a target for the acquisition of vocabulary for children entering school who have little or no knowledge of the language used as the medium of instruction (Cummins, 1981).
- International students who prepare to undertake upper secondary or tertiary education through a new medium of instruction quite simply do not have the time to achieve a vocabulary size that even comes close to that of a native speaker (Read, 2000). Researchers thus attempt to ascertain the minimum number of words needed to cope with the language demands of their studies (Read, 2000). Sutarsyah, Nation and Kennedy (1994) study found that knowledge of 4000 to 5000 words would be a prerequisite for understanding an undergraduate economics textbook written in English.
- In many countries where English is a foreign language, university students are taught through the medium of the national language but they also need to read English texts related to their field of study. As with the international students discussed above it is often useful to calculate a realistic minimum vocabulary size for these students. Many scholars assume that in order to read independently, students should at least know 95% of the running words in a text, this implies that on average only one word in 20 will be unfamiliar to them (Read, 2000).

Researchers who conduct research on vocabulary size are not claiming that students can meet their language needs merely by increasing the number of words that they know (Read, 2000). It would be foolish not to acknowledge that reading comprehension involves grammatical competence, an understanding of how texts are organised, some sort of background knowledge of the subject matter and other abilities, in addition to vocabulary knowledge. The important point to note is that adequate knowledge of words is a prerequisite for effective language use. Those learners who find their vocabulary to be below a certain threshold level tend to struggle to decode the basic aspects of a text, to the extent that they find it difficult to develop any higher-level understanding of the content. If we accept that vocabulary size has important uses as a concept, the big question remains how to measure it? Some of the “how” issues are discussed below.

2.7.1.1 Deciding on what counts as a word. This has already been discussed in a previous section and will not be covered extensively in this section. Just to reiterate, the larger

estimates of vocabulary size for native speakers tend to be calculated on the basis of individual word forms, whereas more conservative estimates regard word families as the units to be measured (Read, 1997).

A difficulty regarding constructing tests is how to separate these words into families (Read, 2000). Nagy and Anderson (1984) had this problem when they attempted to estimate how many words American children were exposed to in the books prescribed in certain schools. They noticed that many words in their sample were semantically related in varying degrees, and they tried to develop a scale of relatedness to try to help in sorting out whether two word forms belonged to the same family. Their scale proved to be not entirely satisfactory because a large amount of subjective judgement was involved, and because it was based on untested assumptions with regard to what children found easy or difficult in interpreting the meanings of words. Bauer and Nation (1993) devised an alternative approach to defining membership of word families, by making use of such criteria as productivity, regularity and frequency of the prefixes and suffixes that are added to base words.

It would seem from the discussion above that the identification of units to be counted is a crucial step in research pertaining to vocabulary size. Aside from the difficulties of distinguishing between base and derived words, researchers have to decide how to deal with abbreviations, homographs, proper nouns, compound words, idioms and multi-word units (Read, 1997).

2.7.1.2 Deciding which words to test. According to Read (2000), it would be practically impossible to test all the words that a native speaker of a language might know. In conventional vocabulary size tests, researchers usually start with a large dictionary and then draw a sample of words, perhaps 1% of the dictionary entries. Next, they test how many of the selected words are known by a group of subjects. Lastly, the test scores are then multiplied by 100 to provide an estimate of the total vocabulary size.

This may seem to be an easy and relatively simple process, but as stated by Nation (1993b) there might be many problems with it. For example, dictionary headwords are not the most appropriate sampling units, because of the problem with base and derived word forms. Secondly, if you were to choose the first word on every sixth page, you might run the risk of producing a sample where common words are overrepresented because these words take up much more space in the dictionary than low-frequency words. Lastly, there are technical questions with regard to the size of the sample required to produce a reliable estimate of

vocabulary size. According to Meara (1996a as cited in Read, 2000), it is challenging to estimate an indefinite large quantity, perhaps even tens of thousands of items, from a small sample of only a few hundred words.

2.7.2 Quality of vocabulary

Despite the advantages of employing vocabulary size tests, one major limitation is that these tests can only provide a superficial indication of how well any specific word is known (Read, 1997). This criticism has been applied to various objective vocabulary tests, not only those that were designed for estimating total vocabulary size. In terms of studies focusing on depth or quality of vocabulary, the existing literature is quite limited, even in studies involving native speakers of English (Boyle, 2009; Douherty-Stahl et al., 2010). Despite the paucity of studies relating to the quality of vocabulary, the existing results suggest that ascertaining the quality of learners' vocabulary appears to have value for a variety of purposes, such as assessing the ability of children of immigrant communities to be educated through their second language (Read, 2000). In the Netherlands, for example, bilingual children from Moroccan and Turkish backgrounds obtain relatively low levels of achievement in school, and linguistic research has indicated that their vocabulary size in Dutch is significantly smaller than that of their monolingual Dutch-speaking peers. This is consistent with Cummin's (1981) findings on vocabulary knowledge of children in Canadian schools.

2.7.2.1 The role of context in assessing quality of vocabulary. There has been very little research that explicitly addressed the role of context in assessing the quality of knowledge of words (Read, 2000). One study that is often cited in the literature is a study conducted by Stalnaker and Kurath (1935 cited in Read, 2000), in which they compared two methods of testing knowledge of German vocabulary. One of the tests is known as the "Best-answer test" which consisted of multiple-choice items, with each target word being presented in isolation. The second test used was referred to as the "context test" and involved constructing a reading passage containing all 100 of the target words. The test-takers were required to supply the English equivalent of each underlined word. When these two tests were administered to German students at the University of Chicago, remarkably similar results were obtained. More specifically, the tests were highly correlated with each other, and the findings reflected very similar correlations with the two other measures of the students' ability employed in the study, namely intelligence test scores and teacher ratings in terms of their achievement in German. From the findings, the authors concluded that the two tests were equally valid

measures of essentially the same ability. In addition, while no recommendations for which type of test should be preferred, the results suggested that there was no real advantage in testing words in context.

However, Read (2000) believes that this is not the case. He maintains that his distinction between context-independent and context-dependent tests is extremely relevant in this discussion. He notes that Stalnaker and Kurath's (1935) context test was primarily based on a text that was specially written to contain all of the 100 preselected words, in the style of a graded reader. Although Stalnaker and Kurath (1935 cited in Read, 2000) maintained that each response had to fit into the context in which it appeared, it is probable that at that elementary level of language learning, the students could have treated each underlined word as an isolated item in most cases, without even needing to refer to the context to get to the correct response. To the extent that this was true, the context test could have been regarded as a context-independent measure of test-takers' knowledge of the target words. In conclusion, despite the lack of research evidence regarding the role of context in testing the quality of vocabulary, language teachers and testers staunchly maintain that vocabulary should always be presented in context.

2.7.2.2 Receptive and productive vocabulary. As users of first and second languages, many people acknowledge that the number of words recognised and understood is larger than the number used in everyday speech and writing (Read, 2000). This distinction between receptive and productive vocabulary is accepted by researchers and scholars working in the fields of first and second vocabulary development, and is frequently referred to by the alternative terms "passive" and "active" (Jalongo & Sobolak, 2011; Millet et al., 2008).

Melka (1997) notes, however, that basic problems still exist in conceptualising and measuring the two types of vocabulary, despite a body of literature being assembled on the subject. Conceptually, the difficulty stems from finding appropriate criteria for distinguishing words that contain receptive status from those that form part of a person's productive vocabulary. As mentioned earlier, many researchers and scholars assume that words are first known receptively and only later become available for productive use. It has been suggested that it could be useful to think in terms of a receptive-to-productive continuum, which represents increasing degrees of knowledge or familiarity with a word. Therefore, when a new word is encountered, learners have limited knowledge of the word, and it is possible that they may not even remember it until they encounter it again. Also, only after the learners gain more

knowledge of the word's spelling, pronunciation, grammar, meaning, range of use and so on, will they be able to use it themselves. The major concern, then, stems from trying to locate the threshold at which the word passes from receptive to productive status. The question remains: is there a certain amount of word knowledge that is needed before productive use is even possible? If a continuum exists, this is simply not a smooth one; a fluid boundary exists with a huge amount of interaction between receptive and productive vocabulary.

In terms of measurement, the lack of an adequate conceptual definition severely hampers the test-measurement process (Read, 2000). The existing literature primarily relates to estimations of size with regard to receptive and productive vocabulary of native speakers and learners. The same problems encountered with vocabulary size tests are relevant here. Melka (1997) has often stressed that there has been no consistency in the way that receptive and productive vocabulary have been measured. Commonly used formats such as multiple-choice, translation and illustration, as well as checklist format, have all been employed by various researchers in the assessment of receptive and productive vocabulary. At the very least, a consensus has to be reached about what counts as a receptive measure and exactly what constitutes a productive one. Read (2000) has tried to solve this problem by describing each type of vocabulary with the use of other terms such as "recognition and recall" and "comprehension and use". These will be discussed below.

2.7.2.3 Recognition and recall. According to Read (2000), "recognition" in the context of vocabulary testing means that the test-takers are presented with a target word and are then required to demonstrate that they understand its meaning. In the case of "recall", the test-takers are provided with some form of stimulus which is designed to evoke the target word from their memory. A simple example of this distinction is found in the literature pertaining to experimental research on vocabulary learning. In these studies "recognition" means providing the participants with the L1 translation of an L2 word, and "recall" refers to the opposite process; they provide the L2 word in response to the L1 translation. This process was utilised by Takala (1984) in his two-way translation in order to estimate the receptive and productive vocabulary of Finnish learners of English. In conclusion, Read (2000) postulates that the major difference between these two types of vocabulary is being able to recognise the word when it is presented and being able to recall it when prompted to do so. Recognition and recall can then represent aspects of vocabulary which can be assessed by the use of selective and relatively context-independent test design.

2.7.2.4 Comprehension and use. “Comprehension” and “use” are ways of distinguishing between reception and production, and are quite different from the way it was delineated above (Read, 1997). “Comprehension” as it is used here means that learners can understand a word when they come across it in context while reading or listening, whereas “word use” means that the word occurs in their own writing or speech. To adequately assess these aspects of vocabulary requires test tasks that are comprehensive and context-dependent. Thus, the researcher might test comprehension by getting the learners to listen to a story or talk comprising numerous target words and then ascertaining how well they understood the words in context. In a similar manner, use can be assessed by setting controlled tasks such as retelling a story, picture description or translation, which is designed to elicit a range of target vocabulary.

However, from the perspective of vocabulary researchers investigating reception and production, these tasks may be viewed as unsatisfactory. Presenting the target words in a whole or spoken text is a relatively inefficient use of testing time, and also the range of vocabulary that can be covered is restricted to words that are related to the topic. Moreover, it is very likely that some of the word meanings can be inferred from the context rather than being already known. In terms of the use tasks, the learner may not apply some of the target words that the researcher wants to test, by unconsciously or deliberately avoiding them (Read, 2000). Because of this, vocabulary researchers have a tendency to prefer very selective and controlled tasks, by presenting the target words in isolation or in a limited sentence context.

Read (2000) maintains that he is not arguing that only comprehension and use tasks are valid measures of both types of vocabulary. Instead he suggests that there is a place for recognition and recall tasks in research and in helping to make decisions about learners. However, problems occur when the terms “reception” and “production” are used unsystematically to refer to both distinctions. This can lead to the assumption that if a recall task has been constructed, then researchers can infer that learners who provide the correct answer are able to, and in fact do, employ the target word appropriately and correctly in their own writing and speech.

The above discussion has illustrated that the terms “reception” and “production” are too broad. In undertaking a vocabulary measurement research project, which involves making this distinction, researchers urgently need to define which specific learner ability each one

refers to. Addressing this issue will lead to providing a better basis for designing suitable testing tasks. In conclusion, the twofold distinction does not resolve these issues, but what it does is highlight the point that even if researchers are dealing with degrees of vocabulary knowledge or ability, no simple continuum running from minimal receptive knowledge to advanced productive ability exists (Read, 2000).

2.8 Validity in vocabulary tests

The concept of validity has undergone many changes (Huysamen, 2002). In the past, four distinct types of validity have been identified, namely construct validity, content validity, predictive validity and concurrent validity, where each type was linked to a different aim in testing. However, many researchers now argue for a unitary concept of validity as all inferences based on test scores are assumed to refer to some underlying construct (Linn, 1994 as cited in Huysamen, 2002; Messick, 1989). In addition Moss (1995 as cited in Huysamen, 2002) states that all validity research must be steered by the principles of scientific inquiry reflected in construct validity.

According to Read (2000), researchers and scholars assume that tests labelled “vocabulary” are measures of lexical knowledge and nothing else. However, Read (1997) notes that the distinction between a vocabulary test and other tests of language ability are not easy to establish by means of statistical analysis. Farr and Carey (1986) reviewed issues relating to measurement in first language reading research and found that a high degree of overlap occurs between tests of vocabulary and the other sub-skills associated with reading. What has often been used in validating vocabulary tests are correlational procedures, but generally what this actually entails is simply correlating one vocabulary measure with another (Read, 1997). In order to demonstrate that vocabulary is a separate component of language ability, it is necessary to make use of a systematic procedure to investigate the relationships between vocabulary tests and other language tests, and this entails the most fundamental kind of research that language testers undertake: construct validation of tests (Bachman, 1990).

In the past, researchers have neglected issues of validity with regard to vocabulary (Schmitt, 1999). However, it seems that recently, researchers have directed their attention to issues concerning validity and vocabulary testing. Studies on vocabulary-item formats focused primarily on aspects such as appropriateness of difficulty, reliability and test rapidity (Henning, 1991; Schedl, Thomas & Way, 1995). When concerns with validity were addressed, it was typically not a comprehensive assessment of construct validity but rather

only one of the facets of validity, such as criterion validity that was assessed. The way to ascertain criterion validity was basically just examining subparts of a test and then comparing them to an estimate of total vocabulary size obtained from the complete test (Henning, 1991). Schmitt (1999) concedes that construct validity was largely ignored because conventional vocabulary tests measured breadth of knowledge where estimates were provided of an individual's vocabulary size. The implication of the aforementioned is the fact that no other methods were used to find a more comprehensive view of validity. Owing to Messick's (1989) notion of a unitary concept of construct validity, Bachman (1990) has proposed that in terms of vocabulary and validity testing, all traditional types of validity (predictive, content and concurrent) should be thoroughly investigated as facets of construct validity.

Schmitt (1999) states that predictive validity of some integrated tests can be operationalised quite easily as the purpose for which the test language is going to be used. Because of the complexity of vocabulary, it is not clear what scores on a vocabulary test could logically predict. While there is some truth in the view that vocabulary tests can predict success in language-related activities like writing (Laufer and Nation, 1995), reading (Laufer, 1992) and producing correct morphology (Schmitt & Meara, 1997), researchers must be careful about linking predictive validation for any single language component, such as vocabulary, to global language performance, particularly when we consider the various other factors which also apply, such as motivation, proficiency and the testee's first language (Schmitt, 1999). Ideally, scores from vocabulary-size items should be used in predicting some lexically based language aspect. Unfortunately, a large enough body of knowledge does not exist with regard to L2 acquisition, to state anything conclusive about the consistency or rate of vocabulary learning. Because of this, predicting future vocabulary knowledge such as quality of vocabulary from vocabulary size scores is unrealistic. The discussion above has indicated that currently, predictive validity may not be the best way to demonstrate construct validity.

“Concurrent validity” refers to the accuracy with which a measure can identify or diagnose the current behaviour or status of the characteristics or specific skills of an individual (Roodt, 2009). This definition implies that a comparison of test scores with another measure is undertaken at approximately the same time (Schmitt, 1999). To be more specific, normally this would involve comparing the sample test scores with those from an established standard test measuring the same construct. In terms of L2 lexis, a major problem is that no such established standardised test can be utilised that has demonstrated adequate validity for this purpose.

To reiterate, the discussion above suggests that there are no appropriate means to establish predictive validity, while the lack of concurrent measures seems to severely limit these approaches as principal means of establishing the construct validity of vocabulary items. In addition, content validity can be established from corpus evidence, but this is perhaps not adequate in and of itself (Schmitt, 1999).

According to Read (2000), most conventional vocabulary tests have aimed to assess knowledge of the meaning of a specific set of words. From this perspective, the relationship between test content and the construct may appear to be quite straightforward. After all, what else could the test be measuring if it presents the test-takers with a set of words with little or no context and then requires them to pick the correct definition or synonym? It also seems quite obvious that a vocabulary test of the above-mentioned kind is measuring something different from a grammar test where the learners are required to recognise a context where the present perfect form of a verb must be used or the preposition identified. Read (2000) notes that test developers and users need to be aware of two major sources of influence on test scores, namely the ability or knowledge represented by the construct, and the testing task. In the field of construct validation these are generally known as “trait” and “method” (Campbell & Fiske, 1959). More specifically, Campbell and Fiske (1959) constructed a methodology referred to as multitrait multimethod (MTMM) construct validation, which permitted these researchers to separately evaluate the contributions of traits and methods to test scores. However, these studies are complex and time-consuming, as a considerable number of carefully planned tests must be given to a large number of test-takers (Read, 2000). Owing to this limitation, only a few studies have been undertaken in the field of language testing.

Despite the paucity of such studies, two important studies have been conducted to ascertain whether it was possible to statistically distinguish between knowledge of grammar and vocabulary (Read, 2000). Corrigan and Upsur (1982) constructed three vocabulary and three grammar tests, each of which assessed knowledge of the same language items using different test methods. For example, in the first method, the language item (word or structure) was presented aurally on tape; in the second method it was presented in a printed sentence; and thirdly it was cued by a picture. The responses required by the test-takers were similarly varied. These tests were then administered to adult English-Spanish language learners from diverse language backgrounds studying at a university in the US. The MTMM procedure entailed a systematic comparison of the correlations between the six tests. The general

principle in this procedure was to establish that vocabulary was an independent trait, therefore it was imperative to demonstrate that the vocabulary tests correlated with each other more highly on average than did the pairs of tests where the methods and traits were crossed or mixed (for instance the picture-based vocabulary test and the aural grammar test). The results revealed that the average correlation between the pairs of vocabulary tests (.216) was actually lower than that for the tests in which the methods and traits were mixed (.257). The results were also lower than the mean correlation of the tests which employed the same method, for example picture-based versus picture based grammar (.358). These results suggested that the authors failed to produce evidence for the construct validity of their tests of vocabulary knowledge. In particular, the results indicated that vocabulary did not emerge as a distinct trait, because the item used had a greater influence on the learner's performance than the issue of whether it was a vocabulary test or not.

Because of the low correlations obtained in the above-mentioned study, Arnaud (1989) attempted to attain more conclusive results by conducting a similar study, utilising tests that were tailored for a more homogeneous group of English learners, namely first-year students attending a French university. Once again the traits used were vocabulary and grammar, which were assessed by three test methods: error recognition, picture-cued multiple choice and French-to-English translation. Despite the study being an improvement in terms of reliability, Arnaud (1989) was also unable to demonstrate that grammar or vocabulary existed as a separate construct. Taking quite a pessimistic stance, Arnaud (1989) concluded that it would never be possible to show that vocabulary is a distinct trait if researchers continue to use the MTMM methodology. In conclusion, these studies have emphasised the difficulty of isolating precise elements of language for measurement purposes (Read, 2000). The studies have highlighted the point that researchers need to be cautious in making assumptions about what aspect of the language is being assessed simply on the basis of the label that the test has been given.

2.9. Summary

This chapter has dealt with vocabulary, especially the importance thereof, the construct of vocabulary and the measurement of vocabulary. As vocabulary is an extremely complex construct, it was necessary to extensively explore all aspects relevant to it because these issues have major implications for testing and ultimately the interpretation of test scores which impact on students' lives. In terms of construct validity and vocabulary measures, it

appears that researchers and test developers have neglected validity issues. Evidence for certain types of validity has been obtained (predictive, concurrent) but researchers have not looked at validity from a comprehensive and holistic viewpoint. The construct “vocabulary” is extremely complex and therefore a thorough investigation of construct validity is required to adequately conclude that vocabulary measures are actually measuring what they are intended to measure.

In Chapter 1, the impact of dialect differences was explored and it was demonstrated that these differences can impact on the valid testing of vocabulary. It is therefore also essential to explore the possible influence that dialect could have exerted, when assessing isiXhosa learners’ scores on the PV scale.

The next chapter outlines the theoretical framework that was used to guide the study.



CHAPTER 3

THEORETICAL FRAMEWORK OF EQUIVALENCE AND BIAS

3.1 Introduction

In this chapter, the theoretical framework that was used in this study will be discussed. More specifically, the theoretical framework of equivalence and bias will be discussed after which research on bias in both monolingual and adapted language versions of tests in general, and the Peabody Picture Vocabulary test, specifically, will be provided. “Test adaptation” refers to changing a test in order to make it more applicable to a specific context (Kanjee & Foxcroft, 2009). The relevance of these issues will become apparent in this chapter.

3.2 Equivalence

“Equivalence” is a technical psychometric term referring largely to the comparability of test scores obtained in different cultural groups, where the most significant question raised is whether obtained scores in different cultures can be meaningfully compared (Van de Vijver & Leung, 1997). In other words, for measures to be considered equivalent, individuals with the same or similar construct or ability but belonging to a different group (such as English or isiXhosa-speaking learners) should receive the same or similar scores on the different language versions of the items or measure (Kanjee & Foxcroft, 2009). In addition, equivalence can be regarded as a function of the characteristics of an instrument and of the cultural groups involved (Van de Vijver & Leung, 1997). Bias and equivalence are closely related but slightly different concepts. If scores are found to be unbiased, they can be regarded as free from nuisance factors and therefore equivalent and can be used for comparison across cultures and language groups (Van de Vijver & Leung, 1997; Van de Vijver & Rothmann, 2004).

3.2.1 Levels of equivalence

Van de Vijver and Leung (1997) have divided equivalence into different levels which takes on a hierarchal form namely, construct inequivalence, construct equivalence, measurement unit equivalence and scalar equivalence. Higher levels of equivalence are more difficult to attain (Van de Vijver & Rothmann, 2004). Construct inequivalence occurs when there is incomparability of constructs across language groups and is analogous to “comparing apples and oranges” (Van de Vijver & Rothmann, 2004, p. 3). When the same construct is measured across all cultural groups studied this is known as construct equivalence or (“structural

equivalence”) (Meiring et al, 2005; Van de Vijver & Leung, 1997). In terms of statistical techniques, the most common analysis employed in examining construct equivalence is factor analysis (Van de Vijver & Rothmann, 2004). In addition, Van de Vijver and Rothmann (2004), states that if the instrument utilised yields the same factors in different cultural groups, then strong evidence is obtained indicating that the instrument measures the same underlying construct. It is important to note that construct equivalence does not presuppose the use of identical instruments across cultures (Przeworski & Teune, 1970 as cited in Van de Vijver & Rothmann, 2004). For instance, an instrument assessing depression may be based on partially or entirely different indicators in each cultural group and still display construct equivalence (Van de Vijver & Rothmann, 2004).

According to Van de Vijver and Rothmann (2004), measurement unit equivalence is obtained when the scales of instruments have the same units of measurement but differ in origin, such as Kelvin and Celsius scales in temperature measurement. More specifically, the units of measurement are equal in both groups but the origins are not (Van de Vijver & Leung, 1997). This form of equivalence assumes interval- or ratio- level scores (Van de Vijver & Rothmann, 2004).

“Scalar equivalence” refers to the highest level of equivalence and is obtained when two metric measures have the same measurement unit and origin and measures the same construct (Meiring et al., 2005). Only when this form of equivalence is attained can direct comparisons be made by employing statistical tests such as analysis of variance and the t-test leading researchers to conclude that average scores obtained in two cultures are equal or different (Van de Vijver & Rothmann, 2004). Furthermore, this form of equivalence assumes that identical ratio and interval scales are applicable across cultural groups (Van de Vijver & Leung, 1997). Claims to the highest form of equivalence can be quite controversial. In particular, researchers and test developers sometimes claim scalar equivalence when only construct equivalence has been established. This has often occurred with regard to personality questionnaires, where an exploratory factor analysis essentially demonstrated similar loadings in various cultural groups, these scholars often argue that scores on these instruments show scalar equivalence.

3.3 The Taxonomy of Bias

Van de Vijver and Leung (1997) have proposed a taxonomy of bias comprising of three types, namely construct, method and item bias. When the construct that is being measured is

not found to be identical in all groups, this is known as construct bias. This form of bias can arise from several sources, for example, in terms of the definition of a construct, there may show an incomplete overlap across cultures. In some instances, the presence of construct bias can be studied by using factor analysis or another technique intended at detecting the structure underlying an instrument. In addition, cross-cultural differences in factor analytic solutions point to construct bias. However, it is important to note that the researcher cannot detect the presence of construct bias with regard to the instrument if it has only been administered once to the target group.

“Method bias” is an umbrella term comprising of all sources of bias resulting from method and procedure of a study and includes sample, administration and instrument bias (Van de Vijver & Rothmann, 2004). Three types of method bias have been identified. The first is known as sample bias and refers to confounding sample differences. In the literature pertaining to cognitive differences between literate and illiterate individuals, this field has experienced many challenges because of sample bias. More specifically, a comparison between the two groups will always be a comparison between schooled and unschooled persons. In the aforementioned example, when studying the impact of literacy then it is almost inevitable that the study will become a study on the influence of schooling. It is therefore essential to consider that sample bias can increase with the cultural distance between the samples.

The second source of method bias is known as administration bias (Van de Vijver & Rothmann, 2004). “Administration bias” can be the result of differences in the mode or procedures used in administering an instrument. For instance, when interviews are conducted in participant’s homes, physical conditions (noise and the presence of others) are extremely difficult to control. Another cause of administration bias can be ambiguity in the instructions and guidelines with regard to questionnaires or a differential application of these instructions.

Method bias can also be due to communication problems between the examiner and examinee (Van de Vijver & Leung, 1997). These could occur from differential interviewing skills and language problems. Communication problems could also arise from the use of locally inappropriate modes of address or other violations of local norms. This type of bias usually affects scores at the level of the whole instrument. Statistically, method bias will be detected in the data as a significant effect for cultural group in a t-test or a significant main

effect for cultural group in an analysis of variance (if we assume that the method bias is sufficiently large enough to reach statistical significance).

The last type of bias is known as “item bias” or “differential item functioning” (DIF) and is a generic term for all anomalies at the item level such as inapplicability of an item to a specific culture or poor translation of items (Van de Vijver & Rothmann, 2004). DIF is reflected in an item when individuals who have the same ability, but differ in terms of culture and language, do not have the same probability of getting the item correct (Dorans & Holland, 1993; Kanjee & Foxcroft, 2009). There are various statistical techniques available to assist researchers and test developers in detecting DIF items (Van de Vijver, 1998). When DIF items are identified and removed from any test, this greatly increases the validity and reliability of test scores (Hambleton & Kanjee, 1995).

In terms of DIF, a distinction is made between uniform and non-uniform bias. If the main effect of language or culture is significant this refers to uniform bias. In other words, the likelihood of answering the item correctly is greater for one group than the other. If the interaction of ability level and language or culture is significant this refers to non-uniform bias. In other words, the difference with regard to the likelihood of the two groups obtaining a correct answer is not identical at all ability levels (Zumbo, 1999).

3.4 Bias in Vocabulary Tests.

When standardized vocabulary tests are used with multicultural populations but were developed for a homogenous cultural or language population, these tests tend to provide a negatively biased view of vocabulary knowledge (Restrepo, Schwanenflugel, Blake, Neuharth-Prichett & Ruston, 2006; Stockman, 2000). This is often observed in performance differences between speakers of AAE and speakers of SE on language assessments despite research indicating that language acquisition does not differ for the two groups (Restrepo et al., 2006). To illustrate the issues relating to bias in vocabulary tests, an in-depth look into a popular vocabulary measure will be provided.

3.4.1 The Peabody Picture Vocabulary test

The Peabody Picture Vocabulary test (PPVT) was first published in 1959 by Dunn and is one of the oldest and most widely used standardised vocabulary tests (Stockman, 2000). This test was designed to measure receptive vocabulary in the English language, and is often used to assess a person’s verbal ability, or form part of a wider battery of tests measuring cognitive

functioning (Haitana, Pitana & Rucklidge, 2010). The first edition consisted of 150 plates, each comprising of four pictures (Jensen, 1975). In terms of administration, the examiner names one of the pictures and the test-taker is tasked with pointing to the correct picture. The vocabulary ranges from very common, easy and concrete words to very rare words with abstract concepts. According to Stockman (2000), many speech-language pathologists utilize the PPVT-III as a screening instrument for verbal ability and in evaluating receptive vocabulary, despite many researchers cautioning them about the inappropriate use of the measure for these purposes (Gray, Plante, Vance & Henrichsen, 1999).

According to Jensen (1975), “cultural bias” in tests refers to the extent that the test contains cultural content that is normally uncharacteristic to the members of one group but not the members of another group. The test is then liable to be biased with regard to comparisons of the test groups between the groups or predictions based on their scores. To the above-mentioned author all the criticism levelled at psychological tests especially mental tests frustrated him, especially the assumption that these tests are culturally biased against certain minorities, specifically blacks, and that these tests were culturally biased in favour of middle-class whites.

In order to dispel these assumptions, Jensen (1975) undertook a major study, aimed at exploring whether cultural bias occurred with a sample of 600 white and 400 black children ranging from 6 to 12 years of age from schools in California, US. The measures used were the PPVT and the Raven’s Progressive Matrices. The results observed were unequivocal; none of the various subjective indices of cultural bias displayed any significant indication of bias in these tests. In particular, the correlation of raw scores with age, rank order of item difficulty, internal consistency reliability, relative difficulty of adjacent items were found to be substantially the same in white and black groups. Thus, Jensen’s objective was obtained: he empirically demonstrated that two popular standardised tests did not display cultural bias with this sample.

Despite these earlier promising results, various subsequent findings reported numerous issues with regard to cultural bias. Rock and Stenner (2005) found that the PPVT-R (a revised edition) produced different results for White American and African American kindergarten - aged children. More specifically, the results suggested that African American first- grade children had approximately half the vocabulary of White American first graders. In addition, the discrepancy between the scores of the two groups remained close to one standard

deviation apart, even after systematically controlling for such factors like parental education, low- birth weight and socio-economic status (SES). Rock and Stenner (2005) then concluded that the test might be biased and recommended that further studies should be conducted in order to explore the factors pertaining to bias in the revised edition of the PPVT-R.

The PPVT-III was standardised on a representative U.S sample, and was alleged to be a culturally valid test mainly because of the inclusion of a substantial representative group of ethnic minority children within a wider norm group (Stockman, 2000). While ascertaining, the appropriateness of the PPVT-III for use with an at-risk sample of African American schoolchildren, Washington and Craig (1999) reported that the scores of the two groups did not differ significantly from the norm group. Because of this the authors concluded that the PPVT-III is a valid and culturally fair test that is suitable for use with African American children.

However, findings of a study conducted by Restrepo et al., (2006) provided evidence that were not consistent with previous research conducted on the PPVT-III. The sample consisted of 210 high-risk, preschool children from a south-eastern state in the US. In particular, African American children and European American children were assessed. When comparing the two groups on two popular standardised vocabulary tests, children who speak AAE and children whose mothers had low education levels tended to score lower on both measures than speakers of SE whose mothers had obtained a high -school diploma or some form of higher education. The above-mentioned authors acknowledge that these differences could be because of the failure of standardised tests to take into account the interplay of language social practices and SES on testee's performances. Furthermore, the bias displayed towards speakers of AAE could be the result of dialect differences of the tester and testee.

Haitana et al., (2010) reviewed all the available literature pertaining to the PPVT-III and they note that it was clear that there is a paucity of research aimed at systematically investigating the use of this measure with people from diverse cultural backgrounds. In addition, the above-mentioned authors acknowledge the well-intentioned stance of many researchers to reduce any previously existing bias, but many researchers maintain that the mere inclusion of an ethnic minority norm group does not make a test unbiased. Rather for a test to be considered unbiased, it is essential that the test also measures culturally appropriate knowledge, utilising methods of testing that are appropriate for people with differing cultural backgrounds (Palmer, 2004).

Haitana et al., (2010) aimed to ascertain the cultural appropriateness of the PPVT-III with a sample of 46 Maori children from three different age groups. One of the main reasons for using this measure in Haitana et al's (2010) study was because many researchers in New Zealand have used this instrument to measure verbal knowledge and receptive vocabulary development in children (Phillips, McNaughton & MacDonald, 2004; Reese & Cox, 1999; Reese & Read, 2000). This measure was chosen as a research tool because it is a well known test of emergent language (Phillips et al., 2004). In terms of reliability, Reese and Read (2004) administered the test to a mixed ethnicity sample of New Zealand children and reported that the test-retest reliability matched those observed by the American referenced norm group. Owing to this finding the researchers concluded that the PPVT-III is a useful instrument with regard to measuring language ability in children from multiple economic and ethnic backgrounds (Reese & Read, 2000).

In terms of the findings of Haitana et al's (2010) study, the results revealed that the PPVT-III is an appropriate measure for use with Maori children as a receptive measure, even though a number of suggestions were discussed as to ways in which the administration and interpretation of the test scores could be modified when working with Maori in order to dramatically minimise the impact of cultural bias. As with all studies, this one had a few limitations. The biggest limitation was small sample sizes which severely limited the range of analyses that were possible. In addition, an inadequate amount of participants were sampled that attended Maori-medium schools, which hampered the possibility of conducting more comprehensive error analyses. Due to these limitations the authors have called for additional research to be undertaken to establish if changes to potential culturally biased items may enhance the validity of the PPVT-III for use with Maori children.

Despite many researchers undertaking studies on the PPVT, a framework which can guide bias at every level was not used, such as the framework proposed by Van de Vijver and Leung (1997). For example, regarding the study conducted by Reese and Read (2004) to assess reliability is not nearly sufficient enough; the researcher or test developer has to investigate all aspects of bias in order to conclude that any test is free from bias. Thus, it is essential to undertake a thorough investigation of bias because only when this has been done, can the researcher or test developer conclude with full confidence that the adapted version of a test is equivalent.

3.5 Research on Monolingual tests with reference to Bias and Equivalence

“Monolingual testing” refers to tests that are only available in one language but are administered across many diverse language groups (Koch, 2007). The reason for discussing monolingual testing in this thesis is because this study is about a monolingual test that is used with different dialect groups. Monolingual tests are popular internationally, and South Africa has become one country of many to use monolingual tests in measuring individuals on a particular trait or aspect (Foxcroft, Paterson, le Roux & Herbst, 2004). Research to provide evidence of the equivalence of test scores across language or cultural groups (and in the case of this study, dialect groups) are therefore as important for monolingual tests as is the case for different language versions.

Allahouf and Abramzon (2008) share similar views because they state that the utilisation of monolingual tests across two language groups is extremely problematic, because a single test form cannot assess proficiency if there is a large gap or variation with regard to the nature of language ability between the two groups. In addition, these authors maintain that if different test forms are used this will lead to major implications with regard to issues of fairness and standardisation. Allahouf and Abramzon (2008) conducted a study that made use of the Hebrew Proficiency Test (HPT). The sample included participants who were Arabic and Russian L1 speakers. The researchers were interested in examining differences relating to performance on L2 test items between the two groups from different L1 backgrounds. In terms of proficiency differences, these were quite small, which by implication increased the accuracy of DIF detection. The results indicated that Arabic speakers performed better than the Russian speakers. Interestingly, the results indicated that grammar and vocabulary items favoured the Arabic speakers mainly because of the similarities between Hebrew and Arabic and because of the presence of cognates in this test. As a result of these findings, the researchers concluded that the HPT functioned differently across these two groups.

Rossier (2004) was interested in determining the cross-cultural equivalence of various personality inventories that are frequently used. More specifically, he focused on personality traits in Switzerland and Burkina Faso. His results demonstrated that the structural equivalence of tests is severely influenced by the theoretical differences on which the tests are based or constructed. In addition, he postulated that when tests are based on theories that are sensitive with regard to cultural context and environmental influences, structural equivalence is less likely to be observed.

Koch and Dornbrack (2008) evaluated bias in monolingual assessment in the South Africa context. More specifically, this study aimed to evaluate the utilisation of language criteria for admission to higher education in South Africa. Despite the fact that policies are in place highlighting multilingualism and that higher education institutions have adopted these multilingual language policies, the predominant languages of learning and teaching are still English and Afrikaans. Due to the aforementioned, students who are isiXhosa first- language speakers from disadvantaged educational backgrounds would suffer major repercussions. Koch and Dornbrack (2008) maintain that these criteria with regard to admission put these students at a severe disadvantage because their educational backgrounds are not being considered when applying to a higher education institution. The results of this study indicated a differential effect of the English- only biased criteria on access. In addition, the findings of this study demonstrated that by evaluating students' performance on a single language and regarding it as representative of their academic literacy with regard to the language of teaching and learning is not only biased but extremely problematic and has concerns in terms of fairness.

3.6 Conclusion

In this chapter, the theoretical framework of equivalence and bias was extensively discussed. All the issues pertaining to this framework were discussed. Specific attention was devoted to bias, where a case study was provided in order to illustrate the various issues surrounding tests. In addition, an account of monolingual assessment and how this relates to bias and equivalence was provided. Furthermore, problems pertaining to monolingual tests were briefly mentioned.

The next chapter outlines the methodology that was used in order to achieve the aims of the study.

CHAPTER 4

METHOD

4.1 Introduction

This study forms part of a broader study ABLE and made use of secondary data analysis (SDA) on data collected for the larger project. SDA can be viewed as the analysis of data that has been previously collected and analysed. Accordingly, the subsequent sections pertaining to the method, namely; the sampling procedures and sample characteristics reflect those of the broader study. In addition, the current study primarily dealt with the equivalence of the adapted isiXhosa version of the WMLS, particularly focusing on the PV scale, across rural and urban isiXhosa dialect groups.

4.2 Research Design

Due to the fact that a comparison of pre-existing groups is made, a differential research design was employed (Gravetter & Forzano, 2009). This design is used when participant characteristics such as gender or language automatically assign participants to groups. In this study the context in which participants are located, namely; the rural or urban South African contexts, defined each isiXhosa dialect group.

4.3 Sampling Procedure

Purposive convenience sampling was used in the main study to select participants. This sampling technique is useful as it allows the researcher to select the sample based on the researcher's knowledge of the population and the objectives of the study (Babbie & Mouton, 2001). This sampling technique allowed the researcher to control for confounding variables such as gender, grade and context (rural and urban area) by assigning equal numbers of participants to the above-mentioned variables.

4.4 Participants

The sample of the larger study consisted of 260 male and female learners who were selected from a population of grade 6 and 7 learners attending schools in the Eastern Cape from both rural and urban areas. Learners from ex-model C schools were excluded from the study as the researcher wanted to control for the effect of not being exposed to academic isiXhosa at school level. IsiXhosa speaking learners at ex- model C schools generally do not take

isiXhosa as a home language at school level. A description of the participants in terms of context (rural & urban) and grade is presented below.

Table 1:

Distribution of participants per language group

AmaXhosa learners	N	%
Rural	127	49
Urban	133	51
Total	260	100

The above table indicates that the size of the groups were fairly similar. More specifically, there are only six more participants in the urban isiXhosa speaking group than in the rural group.

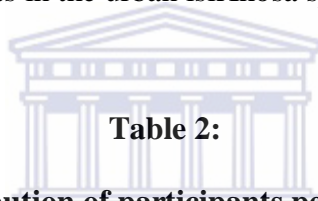


Table 2:

Distribution of participants per grade

AmaXhosa learners	Grade			
	6	%	7	%
Rural	60	65.93	67	39.64
Urban	31	34.06	102	60.35
Total	91	100	169	100

Table 2 indicates that there were more grade 7 isiXhosa learners than grade 6 isiXhosa learners in the urban regions. More specifically, in terms of grade 6 students, the sample comprised of 65.93 % of students from rural areas. However, the opposite was observed in terms of grade 7 students where the majority 60.35% of the sample comprised of students from urban areas. Below a graphic representation of the difference per grade is presented.

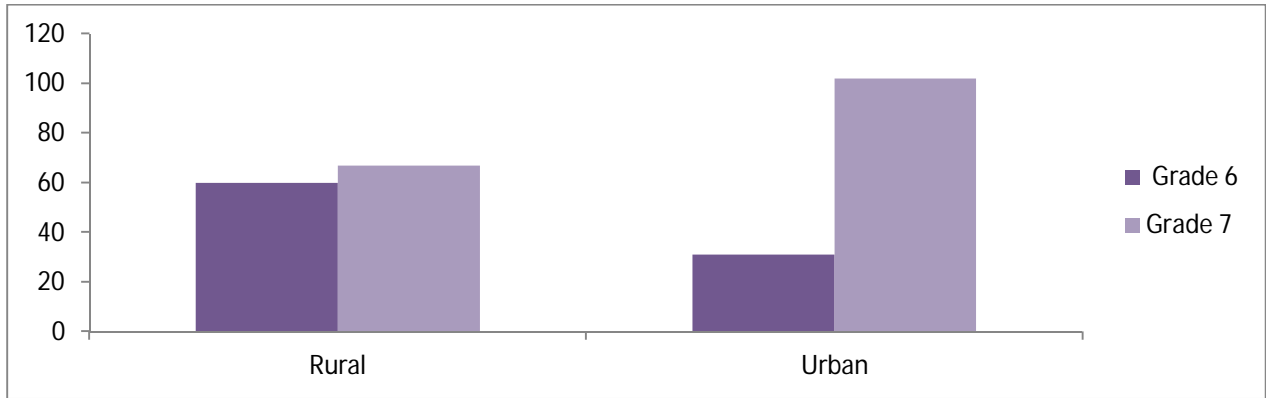


Figure 1: Grade differences between rural and urban isiXhosa learners

Figure 1 demonstrates that there were more grade 7 isiXhosa learners in both dialect groups. However, unlike the grade 6 learners, the proportion of grade 7 learners is higher for the urban isiXhosa learners than for the rural learners.

Table 3:
Distribution of participants per gender

Group	Gender				Total
	Female	%	Male	%	
Rural	63	44.68	64	53.78	127
Urban	78	55.30	55	46.21	133
Total	141	100	119	100	260

The above table indicates that the sample comprised of more males in the rural group. In particular, 53.78% of the sample consisted of males in this group. However, this table also indicates that there were more females than males in the urban group. More specifically, 55.30% of the sample comprised of females in the urban group.

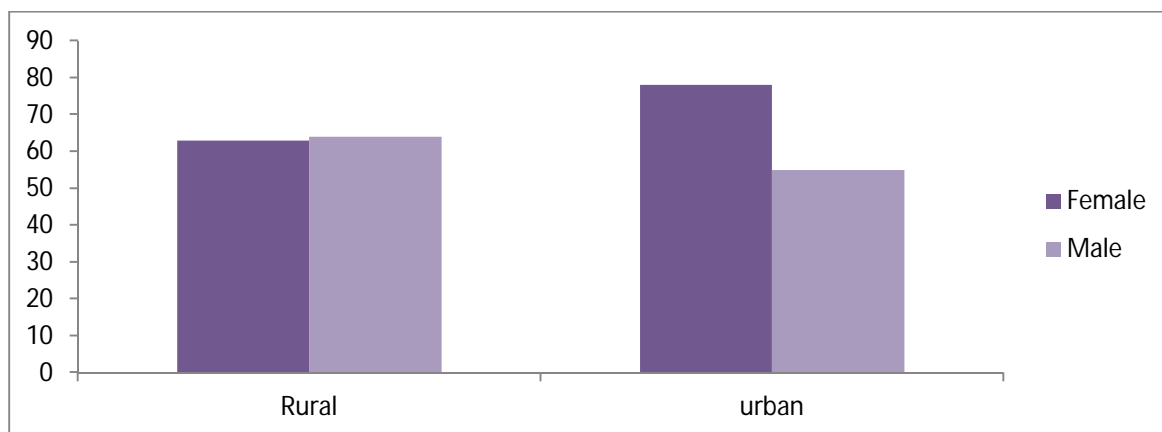


Figure 2: Gender differences between rural and urban isiXhosa learners

The above figure demonstrates that there were slightly more males than females in the rural group, whilst the opposite occurred in the urban group, where more females than males were observed.

4.5 Data Collection Tool

The PV scale forms part of the WMLS. In this section I will firstly describe the test and then focus on the PV scale. The WMLS assesses proficiency in language (reading and writing) and provides an overall measure of language competence and cognitive academic language proficiency levels. The WMLS comprises of four scales which are Scale 1: Picture Vocabulary (PV), Scale 2: Verbal analogies (VA), Scale 3: Letter Word Identification (LWI) and Scale 4: Dictation (Woodcock & Munoz-Sandoval, 2001). The description of the scales with regard to test requirements and measurements is provided in Table 4. In addition, specific groupings or combinations of these four scales form clusters, which are broad measures for interpretative purposes. Cluster interpretation is extremely useful mainly because it minimizes the danger of generalizing the score for a single, narrow behaviour to a broad, multifaceted ability. Employing cluster interpretation, results in an increase in validity because the score that serves as the basis for interpretation comprises of multiple components of a broad ability.

The PV scale of the WMLS measures the ability to name familiar and unfamiliar pictured objects. The scale also taps into breadth and depth of school- related knowledge and experience of the testee and measures oral expression. Despite the scale containing a few receptive items at the beginning, users should note that the scale is primarily an expressive semantic task at the single-word level. There are 57 items in the scale with the difficulty level

increasing as the objects pictured appear less and less frequently in the environment (Woodcock & Munoz-Sandoval, 2001).

Table 4:

Description of scales (WMLS)

SCALE	TEST REQUIREMENTS	MEASURES	RESPONSE STYLE	NUMBER OF ITEMS
Picture Vocabulary (PV)	Testee names the familiar and unfamiliar pictured objects that involve breadth and depth of school-related knowledge and experience.	Oral language, including, language development and lexical knowledge.	Oral (word)	Total= 57
Verbal Analogies (VA)	Testee completes oral analogies requiring verbal comprehension and reasoning.	Reasoning using lexical knowledge.	Oral (word)	Total= 35
Letter-Word Identification (LWI)	Testee reads familiar and unfamiliar letters and words.	Letter-Word Identification skills.	Oral (letter, word, name)	Total= 57
Dictation (Dict)	Testee responds in writing to questions which require verbal comprehension, knowledge of letter forms, spelling, punctuation, capitalisation, and word usage.	Prewriting Skills (for early items), Ability to respond in writing to a variety of questions.	Motor (Writing)	Total= 56

4.5.1 Adaptation and translation process of the adapted isiXhosa versions of the WMLS

As discussed in chapter one, the main aim of the adaptation process was not just to focus on the literal translation of the test but to understand and identify the underlying psychological and linguistic processes as measured by the test and subsequently be cognisant of these factors in the adaptation process (Koch, 2009). Permission from the test developers to adapt the test was obtained and is presented in (Appendix A).

In terms of the PV, what the team found particularly challenging is the fact that the English version of this scale comprised of many historical and culturally loaded items; also many words in this scale lacked equivalent isiXhosa words. Some of the responses to these challenges were to make use of loan words and also replace the culturally loaded English words with words that were more loaded in favour of the Xhosa culture. The latter approach may not have been appropriate as this introduced possible bias in favour of rural as compared to urban speakers of isiXhosa (Koch, 2009).

With regard to the VA scale, the most challenging aspect was the way in which the prompts were framed. For example, the English prompts were in the form of '*mother is to father as sister is to...*'. It is important to note that the analogies increase in difficulty in terms of the underlying logic and pattern that needs to be identified for the analogy to be completed. Nevertheless, the form remains the same throughout the scale. While this phrasing appears to be quite an unnatural way of speaking in English, it still makes grammatical sense in the language. On the other hand, in isiXhosa this way of phrasing a prompt would make no sense at all grammatically. The team then responded to this issue by choosing a completely different form with regard to the phrasing of the prompt (Koch, 2009).

The main challenge with regard to the LWI was the fact that isiXhosa consists of a regular phoneme-based orthography when compared to the irregular orthography of English. In addition, in terms of mastering phoneme-based orthography, this is much easier in isiXhosa than in English when children are starting to develop reading skills. Owing to the aforementioned, the team had to focus on the length of words (because of the nature of sentence construction in isiXhosa, e.g. words tend to be long and possibly may contain various levels of information) and the clicks in isiXhosa as well as the identification of relevant phoneme clusters (Koch, 2009).

In terms of the Dictation scale, the team occasionally had to use completely new words, since articles as employed in English did not work nearly as well for isiXhosa. In addition, comparable issues with regard to writing convention in the two languages had to be identified, taking cognisance of the grading of items in terms of difficulty (Koch, 2009).

4.5.2 Psychometric properties of the WMLS

Norms for the English and Spanish versions of the test were provided from subjects in the United States, Spain, Central America and South America (Woodcock & Munoz-Sandoval, 2001). In terms of the reliability of the WMLS, standard errors of measurement (SEMs) and internal consistency reliability coefficients were established for all English forms and clusters across their scope of intended use. Reliability for the original version were calculated for the USA population, using the split-half procedure, more specifically, the use of odd and even raw scores were used and were corrected for length by the Spearman-Brown formula. With regard to the cluster reliabilities, this was calculated by employing the use of Mosier's (1943) procedure, median reliabilities ranged from .80 to .93 for the scales and .88 to .96 for the clusters. In terms of validity, the WMLS was evaluated on construct, content and concurrent validity.

4.5.3 Psychometric properties of the Picture Vocabulary Scale

This particular study forms part of a broader study aimed at investigating the psychometric properties of the adapted South African versions of the WMLS, and hence will serve to add to the psychometric information currently being assembled for the South African population. Thus far, the WMLS has not yet been normed for the South African population. Consequently, a complete psychometric report of the test for the South African context is currently not available, even though research is currently in progress (Koch, 2009). Previous research has demonstrated that both adapted versions (English and Xhosa) of the WMLS displayed promising results on two of the scales, namely; the LWI (Arendse, 2009; Koch 2009; and VA (Haupt, 2009; Koch, 2009; Roomaney, 2010; Silo, 2009). However, in terms of the adapted English version problems in equivalence have been identified at a structural level with regard to the VA scale (Ismail, 2010).

With regard to the PV scale, previous research explored whether any item bias was observed across rural and urban isiXhosa learners on the adapted version of the WMLS (Silo, 2010). The results identified six items that could be considered as DIF items, namely; items (19, 29,

33, 21, 23, and 35). The first three items had a large effect size whereas the last three items displayed a moderate effect size. It is therefore imperative to explore whether the presence of the DIF items had an effect on structural equivalence (these terms were explained in chapter 3) which will improve with the removal of the identified DIF items from the scale. The Cronbach's Alpha for both rural and urban isiXhosa dialect groups was .77 (Silo, 2010).

4.6 Data Collection Procedure

The main researcher of the larger study obtained ethical clearance from the Nelson Mandela Metropolitan University (NMMU) previously known as the University of Port Elizabeth (UPE) and permission to collect data from the Department of Education in the Eastern Cape (Appendix B and C). Permission was also sought from the principals of the schools and the learner's parents (Appendix D and E). Prior to data collection, it was essential to train test administrators with regard to administering the WMLS in order to facilitate the process of standard administration of the instrument. The data collection took place in 2007 and 2008, where the instrument was administered to learners in a school environment. Thereafter the data was stored in a safe place and later captured and cleaned by the main researcher.

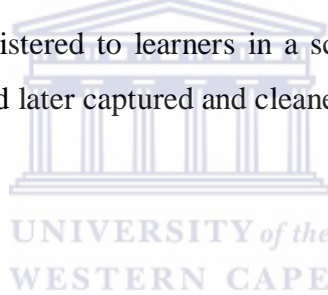
4.7 Data Analysis

4.7.1 Introduction

Due to the use of secondary data, the current researcher made use of existing data of the main study to conduct the relevant statistical test utilising the statistical programme of CEFA (Comprehensive Exploratory Factor Analysis) Version 3.04 (Brown, Cudeck, Tateneni & Mels, 2004) and the Statistical Package for the Social Sciences (SPSS) . The data analysis techniques used in this study will be discussed according to the specific research objectives of the study.

4.7.2 Research objective 1

The research objective was to evaluate the construct equivalence with the DIF items included. In terms of assessing the construct equivalence of the adapted isiXhosa PV scale, exploratory factor analysis was used because this allowed the researcher to identify a latent subset of factors or characteristics that underlie a specific domain (Schaap & Vermeulen, 2008).



4.7.2.1 Factor analysis. According to Campbell, Walker and Farrell (2003), factor analysis is a procedure for reducing the complexity of data by attempting to identify an underlying set of relationships between variables. In addition, factor analysis is a statistical method that had not been extensively used until the advent of computer-based computation because of the complexity and size of calculations that needed to be undertaken. Factor analysis has three main uses: 1) to understand the structure of a set of variables. 2) to construct a questionnaire with the aim of measuring an underlying variable and 3) to reduce a data set to a more manageable size while at the same time, retaining as much of the original information as possible (Field, 2009).

Two broad approaches are linked to data reduction pertaining to factor analytic techniques namely; exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) (Hair, Anderson, Babin & Black, 2010). Of the two approaches, the exploratory approach is more popular, and is utilised when the data under investigation is to be analysed from a theoretical perspective, and the various factors to be extrapolated are identified and referred to as *post facto* (Campbell et al., 2003). Therefore, with regard to EFA, the researcher has very little or even no knowledge about the factor structure, whilst, CFA assumes that the factors are known or hypothesised *a priori*.

4.7.2.2 Exploratory factor analysis. Due to the exploratory nature of EFA, this analysis does not make use of inferential statistics (Costello & Osborne, 2005). Therefore, this design is regarded as the most appropriate for utilisation when exploring a data set, because it was not designed to test hypotheses or theories. In terms of estimation techniques with regard to EFA, the principal components factoring (PCF) and principal axis factoring (PAF) are the most sought after approaches (Hair et al., 2010). In EFA, a variety of alternatives are available in order to characterise the relationships between the variables. With regard to EFA, the Pearson correlation matrix is usually employed. In order to be useful, the Pearson correlation matrix needs data that is interval-scaled as opposed to the Spearman correlation matrix which essentially calculates correlations for ordinal scaled data (Thompson, 2004). However, when the variables in the analysis are of a dichotomous nature, this can often lead to artificial factors; therefore the use of tetrachoric correlations as opposed to the conventional Pearson correlation is recommended in order to increase the validity of the results in the case of dichotomous variables (Kubinger, 2003). This was the approach of this study.

4.7.2.3 Tucker's phi coefficient. To assess the congruence of the construct across rural and urban isiXhosa learners the Tucker's phi coefficient was employed. The Tucker's phi coefficient is often used to evaluate the similarity of factors across different groups (Zumbo, Sireci & Hambleton, 2003). To put it differently, the Tucker's phi enables the researcher to know how similar the pattern of high and low factor loadings are, across different groups. The Tucker's phi was calculated by employing the use of a freeware software program by MarleyWatkins titled Rc. The Tucker's phi formula can be presented as follows:

$$r_{xy} = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 y_i^2}}$$

(Van der Vijver & Leung, 1997)

In terms of analysing the Tucker's phi, observed values that are higher than .95 are viewed as evidence of factorial similarity, whereas observed values less than .85 may indicate non-negligible incongruities (Van de Vijver & Leung, 1997). This is regarded as a rule of thumb and therefore no hypothesis is required. However, some scholars have used a more relaxed Tucker's phi value of .90 or .80 as an indication of factorial similarity (Van der Oord et al., 2005). In this study, a Tucker's phi value of .80 will be regarded as evidence for factorial similarity.

4.7.2.4 Executing the factor analysis. Hair et al., (2010) state that factor analysis follows a linear process structure with specified steps. The steps and the specific decisions for this study will now be described.

The first step is to decide on the method of extraction or estimation. This study used a Common factor analysis in order to determine whether the variables shared underlying latent factors. This form of analysis is primarily concerned with common or shared variance, and is useful for data reduction (Hair et al., 2010).

After determining the method of extraction the next step involves selecting the number of factors to retain. The researcher first ran the analysis with one factor to be retained and then

decided that two factors would be adequate due to the variance left over and the number of items that were not loading on the factor.

The next step in this process entailed making a choice with regard to the rotation method (Costello & Osbourne, 2005). The aim of rotation is to simplify and clarify the data structure. It is important for the researcher to note that rotation cannot improve the basic aspects of the analysis, such as the amount of variance extracted from the items. For this particular study, an oblique rotation was used, because it produces correlated factors facilitating easy interpretation (Hair et al., 2010).

Field (2009) states in the case of oblique rotation, the researcher has to examine the pattern matrix table to ascertain the contribution of the factor loading. Consequently, this is the next step in the factor analysis process. "Factor loading" refers to the co-ordinate of a variable along a classification axis. In addition, the factor loading can be viewed as the Pearson correlation between a factor and a variable. Factor loadings are used to assess the substantive importance of a given variable to a given factor. Typically, when evaluating the relative contribution of each item to a factor, a critical value of .30 is considered as important. However, it is crucial to note that the significance of a factor loading is contingent upon the sample size. As a result of a relatively small sample size ($n = 260$), a strict critical value of .40 was used in order to assess the factor loadings of the two factors (Hair et al., 2010). Items that loaded on more than one factor (cross-loadings) were viewed as poor items, in addition, at least three items should load on a factor if it is to be considered a stable factor.

The factor analysis was run separately for the two isiXhosa dialect groups. More specifically, the analysis was first run with the rural group (a two-factor solution was selected as producing more stable results) and then with the urban group. The data for the urban group was specified to include the same items as well as using a two-factor solution. The other steps pertaining to this process will be discussed in the results section.

4.7.2.5 The reporting of the factor analysis. The reporting will consist of the following steps:

1. The Pattern Matrices of each language group with the DIF items will be presented and discussed.
2. The Tucker's phi of the factors with the DIF items included will be presented and discussed.

3. A scatter plot for each language group will be produced and compared in order to cross validate the findings of the Tucker's phi.
4. The Pattern Matrices of each language group with the DIF items removed will be presented and discussed.
5. Steps 2, 3 and 4 will be repeated with the DIF items removed.

To facilitate easy interpretation the results will be presented for the two phases of the factor analysis separately.

4.7.3 Research objective 2

The objective was to evaluate the construct equivalence with the DIF items excluded. Silo's (2010) study identified six items as DIF items (19, 29, 33, 21, 23, and 35). For this particular study, only the DIF items that displayed a large effect size were excluded from the analysis. Thus, items 19, 29, 33 were removed from the final analysis. As a result of the above, 20 items were included in the analysis. All the steps of research objective 2 were repeated.

4.7.4 Research objective 3

The objective was to evaluate the Cronbach's Alpha of the factors per group after the deletion of the DIF items. "Reliability" essentially refers to the consistency of scores obtained by the same persons when they are re-examined either with the same test on different occasions, or with different sets of equivalent items or under other variable examining conditions (Anastasi & Urbina, 1997). Also, in the broadest sense, test reliability indicates the degree to which individual differences in test scores are attributable to 'true' differences with regard to the characteristics under consideration and the extent to which they are attributable to chance errors. In other words, measures of test reliability allow the researcher or test developer to estimate what proportion of the total variance of test scores is error variance.

Cronbach (1951) also shares similar views pertaining to the above definition of test reliability. More specifically, reliability, including internal consistency measures was viewed by Cronbach (1951) as the proportion of test variance that was attributable to group and general factors where specific item variance or uniqueness was considered error.

Cronbach's alpha is usually used as a measure of the reliability of a set of questions in a survey instrument (Grau, 2007). It basically, measures the interrelatedness of a set of items,

even though a high value for alpha does not imply unidimensionality (where the items measure a single latent construct).

In terms of interpreting Cronbach's Alpha, an acceptable level of reliability has traditionally been set at .70 or higher (Grau, 2007). For this particular study a Cronbach's Alpha value of .70 will be regarded as acceptable and thus illustrate that internal consistency has been met.

The Cronbach's Alpha per factor for each group was calculated. SPSS was utilised in order to achieve this objective. More specifically, a reliability analysis was chosen and was run separately for both groups. Once again, only items that had a loading of .40 were included in the analysis. In terms of the rural group and the first factor, 14 items were included in the analysis. These were items (14, 24, 25, 26, 28, 30, 32, 35, 36, 38, 39, 40, 45, and 47). With regard to the second factor and the rural group, 7 items were included in the reliability analysis. In particular, these were items (3, 8, 9, 13, 15, 45, and 47). For the urban group and the first factor, 9 items were included in the analysis, namely; items (3, 25, 26, 28, 30, 32, 35, 36, and 38). For the second factor and the urban group, 11 items were included in the analysis, which were items (3, 8, 9, 13, 15, 34, 38, 39, 40, 45, and 47).

4.8 Ethical Considerations

Due to using SD, all the relevant ethical considerations were undertaken by the main researcher. Ethical clearance was obtained. In the main study, data was stored in a safe place and in the present study this also occurred. It was reported that participation in the study caused no harm to participants. Permission was obtained to re-analyse the data (Appendix F).

CHAPTER 5

RESULTS

5.1 Introduction

This chapter focuses on the overall aim which is to assess the scalar equivalence of the adapted isiXhosa version of the PV scale of the WMLS across rural and urban isiXhosa-speaking learners by examining construct equivalence. The statistical procedures employed in this study were EFA and the Tucker's phi coefficient, as well as the Cronbach's Alpha. The results of these statistical techniques are summarised in tables and graphs in order to facilitate with analysing and interpreting the data. This chapter serves as a basis for the subsequent chapter in which the implications of the results will be elaborated upon.

Owing to the use of SDA, the current researcher will not be examining the group differences, such as the mean scores and mean item characteristics as this was previously explored in Silo's study (2010).

5.2 Construct Equivalence of the PV scale across the two groups

5.2.1 Steps in conducting the factor analysis

The first phase of the analysis required the selection of a two-factor solution using the data of the rural isiXhosa dialect group first. The following steps were as followed:

- 1) A two-factor solution was specified following from the finding that a one factor solution left too much variance unexplained.
- 2) Items that displayed no variance in either one of dialect groups were removed for the initial solution. They were PV 1, 2, 4, 5, 6, 7, 10, 11, 12, 16, 37, 41, 42, 43, 46, 48, 49, 50, 51, 52, 53, 54, 55, 56, and 57. All participants answered these items either correctly or incorrectly.
- 3) Given the sample size of 127 and 132 respectively for the rural isiXhosa dialect and the urban isiXhosa dialect groups, a strict cut off score of .40 was imposed with regard to the size of the factor loadings (Hair et al., 2010). After the first running of the EFA, items, 17, 18, 20, 21, 22, 23, 27, and 31, were removed due to not meeting the criterion.
- 4) This resulted in 23 items ranging from PV3 to PV47 being used for the final solution. This solution provided a stable structure for the final analysis.

Subsequently, the analysis on the data for the urban isiXhosa dialect group was specified to include the same items as well as utilising a two-factor solution. The Tucker's phi coefficient and scatter plots per factor were used to assess factor congruence.

With regard to the final phase of the factor analysis, for both groups the DIF items identified by Silo (2010) were removed. Once again, the Tucker's phi coefficient and scatter plots per factor were employed to assess factor congruence.

5.2.2 Construct Equivalence results with the DIF items included

5.2.2.1 Factor analysis results. The results of the pattern matrix for the adapted isiXhosa version of the PV scale are illustrated in tables 5 and 6 across the two dialect groups.

Table 5 indicate the loadings on factor 1 and factor 2. The two factors are characterised by high factor loadings and the sufficient numbers of items loading on a particular factor. Fourteen items loaded on the first factor. While fewer items loaded on factor 2 (seven), most had high loadings. No items loaded on none of the factors.



Table 5:

The pattern matrix loadings for the rural isiXhosa dialect group

Item	Factor 1	Factor 2
3	-0.08	0.79
8	0.16	0.85
9	0.10	1.00
13	0.21	0.75
14	0.44	0.04
15	0.05	0.56
19	0.50	-0.37
24	0.51	-0.23
25	0.53	-0.31
26	0.96	0.04
28	0.63	0.34
29	0.36	0.14
30	0.92	0.11
32	0.87	-0.07
33	0.40	0.04
34	0.36	0.09
35	0.94	0.11
36	0.61	-0.07
38	0.91	-0.14
39	0.64	-0.37
40	0.73	-0.14
45	0.39	-0.61
47	0.39	-0.61

Table 6 also indicates loadings on both Factor 1 and Factor 2.

Table 6:

The pattern matrix loadings for the urban isiXhosa dialect group

Item	Factor 1	Factor 2
3	0.62	-0.61
8	0.35	-0.73
9	0.36	-0.73
13	0.17	-0.58
14	0.39	0.07
15	-0.06	-0.55
19	0.28	-0.12
24	0.31	0.10
25	0.39	-0.05
26	0.27	0.36
28	0.47	0.32
29	0.51	0.14
30	0.46	-0.04
32	1.00	-0.03
33	0.26	0.41
34	0.15	0.51
35	0.74	0.10
36	0.54	0.14
38	0.30	0.44
39	0.14	0.68
40	0.07	0.94
45	0.33	0.72
47	0.33	0.72

In terms of the urban isiXhosa dialect group, the number of items loading on the factors was more similar for the two factors than was the case for the rural group. However, a few discrepancies were observed. Item 3 loaded on both factors. In addition, in the rural group

items 19 and 24 loaded on the first factor, but for the urban group these same items did not load on either factor. Also, for the rural group items 15, 26, 33, 38, 39 and 40 loaded on the first factor, but for the urban group these items loaded on the second factor.

It appears therefore that the factor analysis solution that was derived for the rural group does not hold for the urban group. This is further investigated in the next two sections.

5.2.2.2 The Tucker’s phi coefficient per factor with DIF items included. The following table includes the Tucker’s phi coefficients on the factor analysis results with the DIF items included.

Table 7:

The Tucker’s Phi coefficient per factor

Factor 1	Factor 2
.83	.85

For this study, a relaxed Tucker’s phi value of .80 will be used as an indication of factorial similarity (Van der Oord et al., 2005). Thus, we can conclude that the results at this stage are promising for both factors; even though there were discrepancies with regard to the loadings, they did not affect factor congruence. However, if the criterion of .95 (Van de Vijver & Leung, 1997) is used, factor incongruence exists. Factor congruence is thus tentatively accepted, but further investigation using scatter plots will contribute to a better understanding of construct equivalence across the two groups.

5.2.2.3 Scatter plots of the factor pattern coefficients with DIF items included. This section consists of two scatter plots. The first scatter plot represents the factor pattern coefficients for Factor 1 across the two dialect groups. The second comprises of the factor pattern coefficients for the second factor across the two dialect groups.

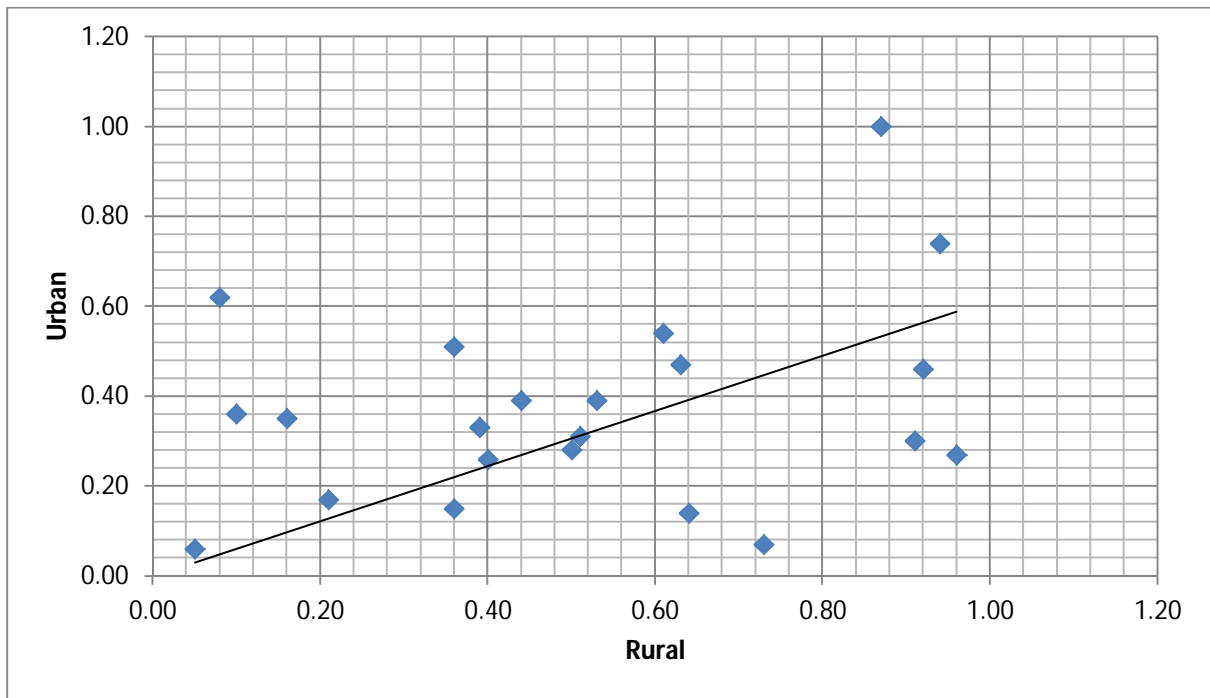


Figure 3: Scatter plot of the factor pattern coefficients for factor 1 across both rural and urban isiXhosa dialect groups

In figure 3 above, one observes the relation of the items towards the identity line. A number of items are reasonably aligned to the identity line. However, a few items fell further from the identity line. Some item loadings were higher for the rural isiXhosa dialect group and some were higher for the urban group. These results serve to confirm the results of the Tucker's Phi and we can then conclude that even though Factor 1 is approaching structural equivalence between the two dialect groups, it is not perfectly congruent.

The scatter plot for Factor 2 was then produced and can be found in figure 4 below.

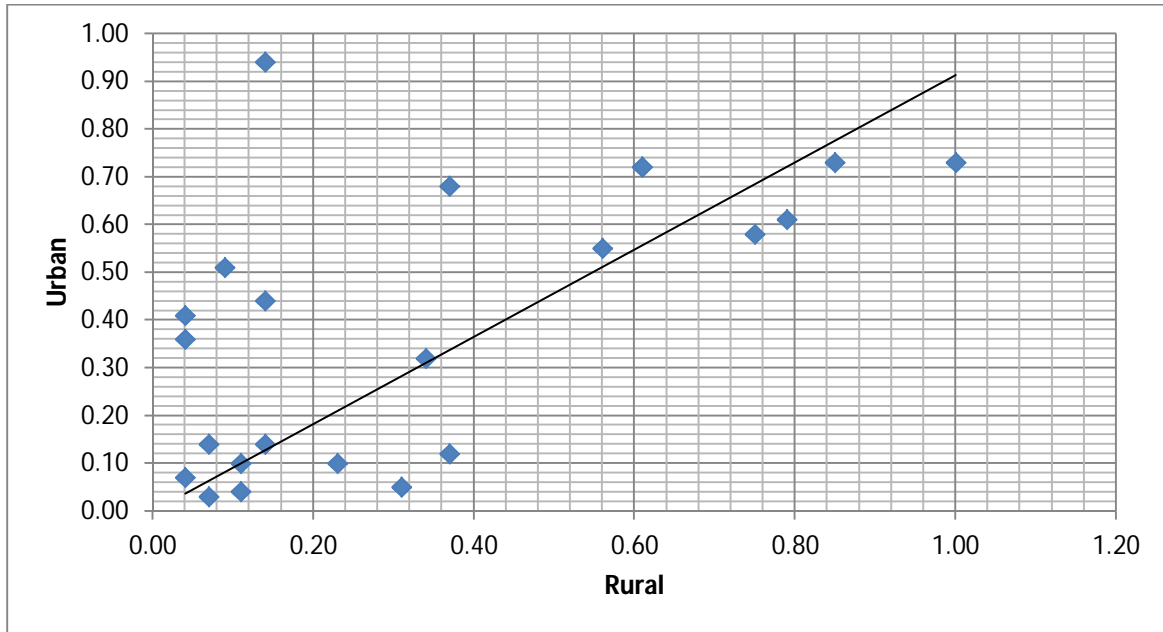


Figure 4: Scatter plot of factor pattern coefficients for Factor 2 across both rural and urban isiXhosa dialect groups

In figure 4 above (all negative values were converted to positive values); one observes the relation of the items towards the identity line. Once again, a number of items were closely aligned, indicating structural equivalence. However, it appears that the factor loadings for the urban isiXhosa dialect are higher than the rural isiXhosa dialect group thus indicating that this factor is better defined for the urban isiXhosa dialect group. The above results once again confirm the results of the Tucker’s phi and one can then conclude that while Factor 2 is also approaching structural equivalence, it is not perfectly congruent.

In the next section the impact of the DIF items on equivalence (congruence) are explored.

5.2.3 Construct Equivalence results of the factors with the DIF items removed

5.2.3.1 Factor analysis results. As stated in the previous section, with regard to the PV scale Silo’s (2010) study identified six items as DIF items (19, 29, 33, 21, 23, and 35). For this particular study, only the DIF items that displayed a large effect size were excluded from the analysis. Thus, items 19, 29, 33 were removed from the final analysis. As a result, 20 items were included in the analysis.

The following tables (8-9) represent the two-factor solution for the rural and urban isiXhosa dialect group of the adapted isiXhosa version of the PV scale with the DIF items excluded.

Table 8:

The pattern matrix loadings for the rural isiXhosa dialect group with the DIF items excluded

Item	Factor 1	Factor 2
3	0.05	0.77
8	0.20	0.80
9	0.13	0.96
13	0.26	0.73
14	0.43	0.02
15	0.04	0.60
24	0.52	-0.23
25	0.54	-0.04
26	0.97	0.01
28	0.61	0.33
30	0.93	-0.09
32	0.87	-0.08
34	0.34	-0.08
35	0.92	0.12
36	0.60	-0.07
38	0.92	-0.16
39	0.58	-0.38
40	0.72	-0.14
45	0.40	-0.70
47	0.40	-0.70

The results reveal distinct loadings on factor 1 and factor 2 for the rural isiXhosa dialect groups. Most of the items loaded on factor 1 (13). Items 45 and 47 now load on both factors, while no items loaded on none of the factors.

Table 9:

The pattern matrix loadings for the urban isiXhosa dialect group with the DIF items excluded

Item	Factor 1	Factor 2
3	0.57	0.66
8	0.32	-0.79
9	0.33	-0.79
13	0.12	-0.58
14	0.32	0.07
15	-0.09	-0.53
24	0.35	0.15
25	0.43	-0.04
26	0.42	0.39
28	0.41	0.29
30	0.51	-0.05
32	0.99	-0.08
34	0.08	0.47
35	0.75	0.06
36	0.52	0.15
38	0.41	0.43
39	0.26	0.66
40	0.15	0.93
45	0.32	0.67
47	0.32	0.67

The above results once again indicate a sufficient number of loadings on both factors for the urban isiXhosa dialect group. Item 3 loaded on both factors, and item 14 loaded on none; in other words, fewer items than before loaded on none of the factors. Of the 20 items included in the analysis, 16 of the items loaded on the same factors for both groups. Only items 34, 39

and 40 loaded on separate factors. In particular, for the rural group these items loaded on the first factor but for the urban group, these same items loaded on the second factor.

It thus seems as if factor congruence might have improved with the removal of the large DIF items.

5.2.3.2 The Tucker's Phi with DIF items removed

Table 10:

The Tucker's Phi coefficient per factor

Factor 1	Factor 2
.88	.88

After excluding the three items that displayed a large effect size with regard to DIF, the congruence indicator for both factors improved. If we were to use the relaxed Tucker phi value of .80 as an indication of factorial similarity (Van der Oord et al., 2005) then we can conclude that these factors are congruent. However, Van de Vijver and Poortinga (2002) states that values higher than .95 are regarded as evidence for factorial similarity, whereas values lower than .90 or .85 (Ten Berge, 1986) point to non-negligible incongruities. Owing to the above, we can only tentatively conclude that the two factors are approaching congruence. We will investigate this further in the next sections.

5.2.3.3 Scatter plots of the factor pattern coefficients with DIF items removed. This section once again comprises of two scatter plots. The first scatter plot represents the factor pattern coefficients for factor 1 across rural and urban isiXhosa dialect groups with the DIF items removed. The second scatter plot represents the factor pattern coefficients for Factor 2 across the two isiXhosa dialect groups with the DIF items removed.

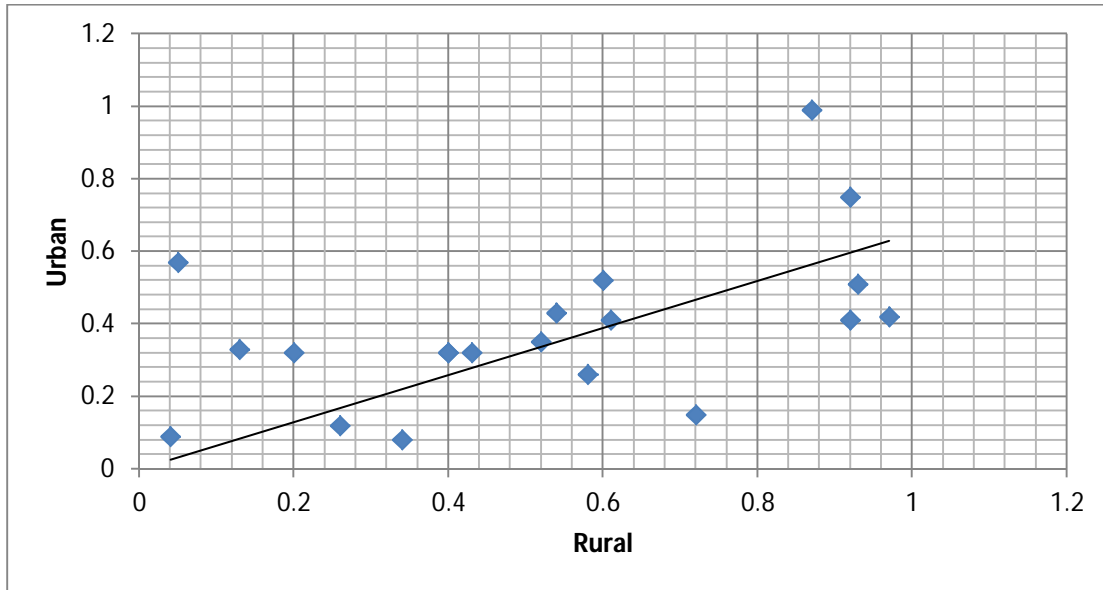


Figure 5: Scatter plot of factor pattern coefficients for Factor 1 for both rural isiXhosa and urban isiXhosa dialect groups

In figure 5 above, one observes the relation of the items towards the identity line. More items than before are closely aligned to the identity line across both dialect groups for Factor 1. However, a few of the items fall further from the identity line, again some indicating higher loadings in the urban group, and some with higher loadings in the rural group.

Thus, the above results confirm the results of the Tucker's Phi and one can then conclude that despite Factor 1 approaching structural equivalence and the improvement in the congruence, it is not perfectly congruent.

The scatter plot for Factor 2 was also produced and can be observed in Figure 6 below.

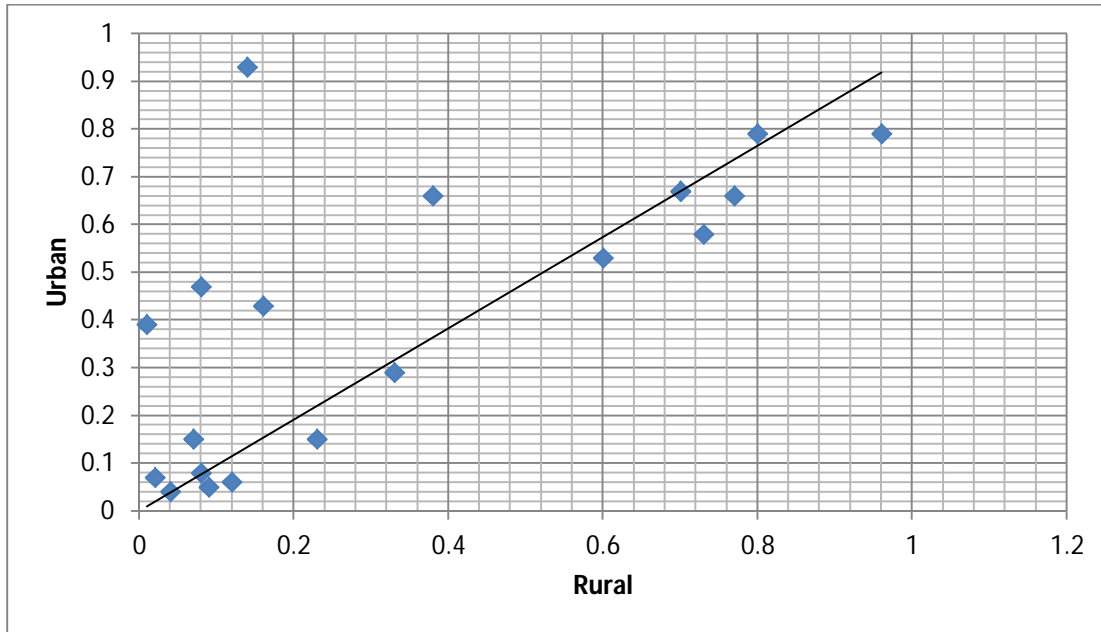


Figure 6: Scatter plot of factor pattern coefficients for Factor 2 for both rural and urban isiXhosa dialect groups

In the above figure, one observes the relation of the items towards the identity line. Again, more items than before are aligned to the identity line. However, it is clear that in this factor the urban group had higher loadings for a number of items than the rural group, indicating a more stable factor in the urban group.

Again, these results confirm the results of the Tucker’s phi and one can therefore conclude that despite Factor 2 approaching structural equivalence, this factor is not perfectly congruent.

The Cronbach’s Alpha for the two factors will now be investigated for both groups.

5.3 Cronbach’s Alpha of the factors after the deletion of the DIF items

Table 11:

The Cronbach’s Alpha for the two factors across the two dialect groups

Group	Cronbach's Alpha	
	Factor 1	Factor 2
Rural isiXhosa	.79	.55
Urban isiXhosa	.69	.45

Schmitt (1996) states that for research purposes an alpha coefficient of .70 or higher can be regarded as acceptable or adequate. However, amongst measurement experts the acceptable value of the alpha coefficient has been fiercely debated not just for research but for clinical purposes as well (Schmitt, 1996; Streiner, 2003). Nunnally (1967) in the first version of his book recommended that in the early stages of research an alpha coefficient of .50 to .60 is adequate, .80 is sufficient for basic research tools and .90 as the minimal tolerable estimate for clinical purposes, with an ideal of .95. With regard to acceptable alpha coefficients for clinical purposes this is extremely crucial for these values are used to make decisions about individuals (Streiner, 2003). Streiner (2003) argued that Nunnally's (1967) recommendations are extremely useful for research purposes but maintains that the alpha coefficients for clinical scales are too high and suggests that the ideal value should be .90.

In terms of factor 1 for the rural group, the results indicate that this factor (.79) has a relatively high internal consistency. However, with regard to factor 2 of the rural group and both factors for the urban group, the results have indicated that the alpha coefficient is below the acceptable value or criteria and thus can be regarded as not displaying internal consistency. In addition, it appears that the two groups differ on their Cronbach's Alpha on both factors. Compared to the reliability of this scale in the two groups in Silo's study (2010) (Rural: .77, and Urban: .77) it therefore appears as if the reliability improved slightly for the rural group on factor 1 (after the DIF items were removed), but not for the urban group, while the reliability for both groups on factor 2 is much lower than for the overall scale.

5.4 Naming of the factors

According to Hair et al., (2010) when a stable factor solution has been derived, the next step would be to attempt to assign some meaning to the factors. This process entails substantive interpretation of the pattern of factor loadings for the variables, which includes their signs, in an attempt to name each of the factors. In addition, all significant factor loadings usually are used in the interpretation process. Thus, variables with higher loadings influence to a large degree the name or label assigned in representing a factor.

For this particular study, items that loaded at .9 or above were used in order to name the factors, while the loadings in the rural group (as the reference group) were used for the naming. Four items had the highest loadings. More specifically, items 26 (magnet, .96), 30 (theatre stage, .92), 35 (printing press, .94) and 38 (thermostat, .91) all contributed to the naming of the first factor. Due to the aforementioned, Factor 1 was labelled culturally-

influenced vocabulary, because the items with the highest loadings comprised of pictures that required words with a clear cultural (mostly western culture) influence. In other words, factor 1 content has a tendency to have a strong urban or western influence.

With regard to factor 2, items 32 (mine, .1.00) and 40 (fish gill, .93) had the highest loadings and thus lead to factor 2 being labelled as nature-orientated and familiar, because these items comprised mostly of images that are more familiar to students in the rural areas.

5.5 Summary

This chapter primarily focused on the statistical analyses that were used in order to evaluate three specific objectives outlined in Chapter 1 of the study.

The first objective evaluated the construct equivalence with the DIF items included. EFA was used in order to assess this objective. More specifically, CEFA was used where a two factor solution structure was specified because a one factor solution left too much of the variance unexplained. In terms of the results, distinct loadings were observed for both groups. However, slight differences were observed on the loadings as the results of the Tucker's Phi indicated slight congruence. The use of scatter plots was then employed to further investigate and confirmed slight congruence only.

When the DIF items that were identified in Silo (2010) were removed, distinct loadings on both factors once again occurred for both groups, but slight differences were once again observed despite the improvement seen in the results of the Tucker's phi. Further examination was warranted, where scatter plots were once again used which served to confirm slight congruence, thus raising certain concerns.

With regard to the third objective which was to evaluate the Cronbach's Alpha of the factors after the deletion of the DIF items. SPSS was employed in order to assess this objective. For the rural group, only Factor 1 obtained a relatively high internal consistency. However, Factor 2 raised certain concerns with regard to reliability because for both groups internal consistency was not displayed.

In the next chapter, Discussion and Conclusion, the results of this chapter will be discussed in light of the overall aim of this study, which is to assess the scalar equivalence of the adapted isiXhosa version of the PV scale of the WMLS across rural and urban isiXhosa speaking learners.

CHAPTER SIX

DISCUSSION AND CONCLUSION

6.1 Introduction

The overall aim was to assess the scalar equivalence of the adapted isiXhosa version of the PV scale of the WMLS across rural and urban isiXhosa speaking learners. To put it differently, the purpose was to assess whether the scores on the PV scale can be utilised across the two dialect groups. The overall aim was evaluated by means of three objectives, namely to:

- 1) Evaluate the construct equivalence with the DIF items included
- 2) Evaluate the construct equivalence with the DIF items excluded and
- 3) Evaluate the Cronbach's Alpha of the factors after the deletion of the DIF items

This chapter will thus primarily focus on the major findings of the results and a comprehensive discussion will be provided in order to identify the implications of these results as well as the limitations pertaining to the study. Recommendations for future research will be briefly discussed based on these results and the chapter will conclude with a few remarks on the present study.

6.2. Discussion of the Results

The following results will be discussed in terms of the three objectives of the study in order to ultimately evaluate the main aim of the study of scalar equivalence.

6.2.1 Results of the exploratory factor analysis

The first objective was assessed by employing the use of CEFA and was conducted across the two dialect groups. In addition, in this analysis the DIF items were included. The results revealed that two factors were distinguishable in both groups as indicated by their high factor loadings. Despite these initial promising results, slight differences across the two groups were observed that warranted further investigation. The results of the Tucker's phi and scatter plots thus confirmed that only while the solutions approached congruence, it was not perfect.

The second objective was also assessed by employing the use of the aforementioned statistical tools, this time with the DIF items excluded. The results once again indicated that two distinct factors were present for both groups. Sufficient Tucker's phi values once again indicated that the two factors were approaching factorial similarity, but that the results continued to raise concern about construct equivalence. The scatter plots supported the concern.

Previous research on this scale (Silo, 2010) identified DIF items which were attributable to dialect differences. After assessing the impact of the DIF items in the current study we can conclude that even after deleting these items, structural equivalence was only partially achieved. From the scatter plots it appeared as if especially factor 1 tended towards lower congruence and therefore structural inequivalence. It seems therefore that while we may be able to use the scale across the two dialect groups, we still need to interpret differences in the scores with caution. Van de Vijver and Leung (1997) stated that DIF can be regarded as being an important source with regards to information about cross-cultural differences. As a result the unbiased items within the PV could assist to define cultural commonalities of the construct.

The naming of the factors indicated that factor 1 comprises of items that can be regarded as culturally influenced. A number of these items have a strong Western influence, while others, as a result of the adaptation process, tended towards a more traditional isiXhosa cultural influence (Koch, 2009). The finding that this cultural influence leads to lower factorial congruence is in line with Thipa's (1989) views where he stated that rural amaXhosa comprises of speakers of the language that have been least exposed to Western experiences and influences, while urban isiXhosa have less exposure to traditional cultural artifacts and practices. In Silo (2010)'s study it was also found that the two groups did equally well on this scale. However, given the fact that the urban group consisted of more Grade 7 learners where we would have expected a higher vocabulary score, this may indicate that the structural inequivalence has an impact on the score comparability across the two groups. Jensen (1975) stated that cultural bias refers to the extent to which a test contains cultural content that is normally uncharacteristic to members of one group but not members of the other group. Cultural bias may explain low structural equivalence of this factor across the two groups, a finding that remains even after the removal of the large DIF items.

The structural differences and item bias (Silo, 2010) observed between the two groups are also in line with findings conducted on the PPVT-III where differences were observed amongst children who spoke AAE and those who spoke SE (Restrepo et al., 2006). In particular, bias was observed towards speakers of AAE and this was attributed to possible dialect differences of the tester and testee. This could possibly explain why the factors approached congruence but did not display perfect congruence.

The items in the PV scale of the adapted isiXhosa version of the WMLS are presented in a discrete way, where test takers are required to name objects where no context is provided. This format may not be appropriate. If the vocabulary items would have been presented in a more comprehensive design, then cognisance of all the vocabulary content of a written or spoken text could have occurred (Read, 1997). This design proposed by Read (2000) could be used where learners are rated on various criteria, which broadens their range of expression thus definitely increasing the quality of the learners overall vocabulary use and providing more knowledge with regard to vocabulary.

Read (2000) states that with regard to vocabulary, many questions need to be addressed because these could ultimately affect the measurement of this construct. The problems with congruence could be the result of issues pertaining to 'defining words' such as the intricacies of base words, word families and homographs. In addition, slight congruence can also be explained by the difficulties with larger lexical items. In other words, vocabulary comprises of more than just single words, and includes phrasal verbs and compound nouns, which cannot be inferred by just knowing the individual word (Read, 1997). These issues with congruence can be overcome if vocabulary tests were more contextually based. More specifically, test developers should consider alternative types of test design with regard to vocabulary. Read (1997) suggested that tests measuring vocabulary should be more comprehensive and context- dependent where the test takes cognisance of all the vocabulary content of a written or spoken text because this will lead to forming a judgement of the quality of a students' overall vocabulary use

In a similar vein, Bachman and Palmer (1996) stated that vocabulary knowledge will greatly be enhanced if we were to draw on various types of pragmatic knowledge. More specifically, test developers should direct their attention to the social and cultural situation in which lexical items are used for this can dramatically impact on their meaning. One of the ways context can affect lexical meaning are the differences in interpretation across language

varieties (standard and non-standard). The fact that the factors approached structural congruence but certain concerns were raised points to the possibility that the social context of vocabulary use could have impacted on the results of the test scores. This was observed in the many items that favoured urban isiXhosa speaking learners. Taking cognisance of the cultural and social context and consequently constructing items that are culture-free could improve the congruence of the two factors.

6.2.2 Results of the Cronbach's Alpha per factor after the deletion of the DIF items

In order to evaluate this objective, SPSS was employed. The general trend observed here was the fact that Factor 1 displayed internal consistency only for the rural group. In terms of the urban group, both factors were problematic with regards to internal consistency. These findings are similar to Ismail (2010) where the results of the adapted English version of the WMLS of the VA scale across English and isiXhosa learners of factor 2 were also below the acceptable level and hence did not display internal consistency. These results could be because Factor 2 for the rural group and both factors for the urban group had lower items in the final analysis. In order to overcome this challenge, researchers can further develop items of a similar nature than the unbiased items in both factors and then attempt to evaluate the alphas again.

With regard to alpha and the acceptable level of reliability, consensus has not yet been reached amongst measurement experts (Coolican, 2004; Cortina, 1993; Grau, 2003; Nunnally, 1967; Schmitt, 1996; Streiner, 2003). More specifically, Klassen (2003) states that an alpha value that is equal to or greater than .6 is considered a minimum acceptable level, despite some authors' arguing for a stronger standard of at least .7. Thus traditionally, .7 has been considered as the acceptable level although interpretation of alpha in specific contexts is generally more complicated than that (Schmitt, 1996). In particular, Schmitt (1996) noted that a high alpha is possible even when the item responses are multidimensional. In addition, the level of alpha is also linked to the number of items being tested. Cortina (1993) demonstrated how the value of alpha varied according to the number of items tested and how alpha usually declined as the number of dimensions increased. What Cortina (1993) did not indicate however, was although a high level of alpha does not guarantee unidimensionality nor does it automatically indicate high average item intercorrelations, a low level of alpha is often associated with multidimensional data. Thus, the low alpha values of the factors are worrying and warrant further research. This is especially important for this study because although .7 is

normally regarded as satisfactory, (Schmitt, 1996) but in terms of diagnosis, .9 and up is regarded for selection of diagnosis (Nunally, 1967; Streiner, 2003). Thus error in selection increases with low alpha values which could lead to a misdiagnosis of a language disorder, with severe consequences for the test taker.

6.2.3 Implications of the findings

In the present study it was speculated that the removal of previously identified DIF items would contribute to structural equivalence and hence ultimately bring the researcher closer to establishing scalar equivalence across the two dialect groups on the adapted isiXhosa version of the PV scale. It was thus assumed that any structural differences observed in previous research was the result of the presence of DIF items and that if these items were subsequently removed, the structural differences would disappear. However, despite stable factors being identified for both groups and with both factors approaching structural equivalence, perfect congruence was not attained.

6.3 Limitations of the study

Due to the fact that this study employed the use of secondary data, the sampling procedure was not executed with the current study in mind but rather for the broader project. In this study generalisability was not a core factor; it is important, though, to note that this impacts on the external validity of the adapted versions of the WMLS. However, this limitation is not regarded as a main limitation pertaining to the study because the main concern here was to assess whether the WMLS is an appropriate instrument to utilise in the ABLE project. Due to the above-mentioned concerns with issues of internal validity were thus regarded as more pertinent than external validity at this stage.

Another major limitation of the current study in the modest sample size utilised. Despite the sample adhering to the minimum sampling criteria needed for the different statistical methods that were employed, a larger sample size might have yielded more significant results. In addition, more Grade 7 students were sampled in the urban group. Research has indicated that ability does appear to impact on differences in factor structure (Sireci & Khaliq, 2002).

Keeping the limitations in mind, the conclusions of the study will now be presented.

6.4 Conclusion

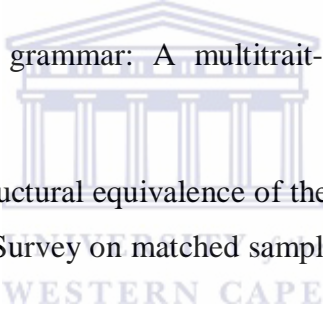
The central aim or goal of this study was to explore the scalar equivalence of the adapted isiXhosa version of the PV scale of the WMLS across rural and urban isiXhosa speaking learners. Thus, the researcher attempted to ascertain whether scores on this scale could be compared with students who reside in the rural areas of the Eastern Cape and thus are more exposed to standard isiXhosa and students from the urban part of the Eastern Cape who tend to speak the non-standard form of isiXhosa. In other words, the study revolved around the following question: do the scores obtained on the adapted isiXhosa version of the PV scale carry the same meaning for the two groups, namely, is vocabulary being measured? If this is not the case, then construct inequivalence would have occurred.

With regard to the PV scale after the removal of the DIF items, this scale approached congruence. However, we need to exercise caution in the interpretation of the scores across the two dialect groups. As a result of the aforementioned, vocabulary scores obtained on this scale needs to be supplemented with other relevant information about vocabulary following an assessment approach to testing. Just using this score, as an indication of vocabulary knowledge will not be sufficient as it may present with differences across the two dialect groups mainly due to measurement artefacts.

6.5. Recommendations for further research

The promising results obtained in the study has lead the researcher to recommend that more research should be conducted on the dialect differences between rural and urban isiXhosa speaking learners. It is also recommended that research could be conducted where all six items that displayed DIF (Silo, 2010) be excluded in the final analysis with regard to evaluating construct equivalence as this could increase the Tucker phi values. However, this will impact on reliability (as the length of a scale impacts on reliability), and it is thus also recommended that the developers and adaptors of the test use the information regarding non-biased items to add items to the scale. Lastly, the research on this scale would be greatly enhanced if the two dialect groups were matched on ability.

REFERENCES

- Allalouf.,A. & Abramzon, A. (2008). Constructing better second language assessments based on differential item functioning analysis. *Language Assessment Quarterly*, 5, 120-141.
- American Heritage Dictionary of the English Language, (2000). Boston: Houghton Mifflin.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed). Upper Saddle River, New Jersey: Prentice- Hall, Inc.
- Anderson, R.C., & Freebody, P. (1981).Vocabulary knowledge. In J.T.Guthrie (Ed). *Comprehension and teaching: Research reviews* (pp.77-117). Newark, D.E: International Reading Association.
- Apel, K., & Thomas-Tate, S. (2009). Morphological awareness skills of fourth-grade African American students. *Language, Speech and Hearing Services in schools*, 40, 312-324.
- Arnaud, P. (1989). Vocabulary and grammar: A multitrait-multimethod investigation. *AILA Review* 6, 56-65.
- Arendse, D. (2009). Evaluating the structural equivalence of the English and isiXhosa versions of the Woodcock Munoz Language Survey on matched sample groups. Unpublished MA thesis. University of the Western Cape. 
- Babbie, E., & Mouton, J. (2001). *The practice of social research*. Cape Town: Oxford.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bauer, L., & Nation, I.S.P. (1993). Word families. *International journal of Lexicography*, 6, 253-279.
- Beck, I.L., Perfetti, C. A,& McKeown, M.G. (1982). The effects of long-term vocabulary instruction on lexical access and reading comprehension. *Journal of Educational Psychology*, 74, (4), 506-521.

- Bekker, I. (2005). *Language attitudes and Ethnolinguistic Identities in South Africa: A critical review*. Somerville, MA: Cascadilla Press.
- Boyle, R. (2009). The legacy of diglossia in English vocabulary: what learners need to know. *Language Awareness, 18* (1), 19-30.
- Brown, G. (1994). Modes of understanding. In G. Brown, K. Malmkjaer, A. Pollitt, and J. Williams (Eds). *Language and understanding* (pp.10-20). Oxford: Oxford University Press.
- Browne, M. W., Cudeck, R., Tateneni, K, & Mels, G. (2004). CEFA: Comprehensive Exploratory Factor Analysis, version 2.00 [Computer software and manual]. [On-Line]. Accessed 25 August 2011. Available : <http://quantrm2.psy.ohio-state.edu/browne>.
- Caltreux, K.C. (1996). Standard and non-standard language varieties in the urban areas of South Africa. Main report for the STANON research programme. Pretoria: HSRC Publishers.
- Campbell, D. T, & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.
- Campbell, A., Walker, J, & Farrell, G. (2003). Confirmatory factor analysis of the GHQ-12: Can I see that again? *Australian and New Zealand Journal of Psychiatry, 37*, 475-483.
- Chan, A. S., Cheung, M.C., Sze, S.L., Leung, W.W., & Cheung, R.W.Y. (2008). Measuring vocabulary by free expression and recognition tasks: Implications for assessing children, adolescents, and young adults. *Journal of Clinical and Experimental Neuropsychology, 3* (8), 892-902.
- Claasen, N. C.W. (1997). Cultural differences, politics and test bias in South Africa. *European Review of Applied Psychology, 47* (4), 297-307.
- Chapelle, C.A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language research, 10*, 157-187.
- Cain, K., Oakhill, J., & Lemmon, K. (2004) Individual differences in the inference of word meanings from context: the influence of reading comprehension, vocabulary knowledge and memory capacity. *Journal of Educational Psychology, 96*, 671-681.

- Catts, H., Adlof, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: a case study for the simple reading view of reading. *Journal of Speech, Language, and Hearing Research, 48*, 1378-1396.
- Coolican, H. (2004). *Research Methods and Statistics in Psychology (4th Ed)*. Great Britain: Hodder and Stoughton Educational.
- Cooper, T., & Van Dyk, T. (2003). Vocabulary Assessment: A look at different methods of vocabulary testing. *Perspectives In Education, 21* (1), 67-79.
- Corrigan, A., & Upsur, J. A. (1982). Test method and linguistic factors in foreign language tests. *IRAL 20*, 313-321.
- Cortina, J. (1993). What is coefficient alpha? An examination of theory and methods. *Journal of Applied Psychology, 78*, (1), 98-104.
- Costello, A.B., & Osborne, J.W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation, 10* (7), 2-9.
- Coyne, M. D., Simmons, D. C., & Kame'enui, E.J. (2004). Vocabulary instruction for young children at-risk of experiencing reading difficulties . In J. F. Baumann and E. J. Kame'enui (Eds). *Vocabulary instruction: Research to practice* (pp.41-58). New York: Guilford Press.
- Cronbach, L. J. (1942). An analysis of techniques for diagnostic vocabulary testing. *Journal of Educational Research, 4*, 181-189.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 22* ,(3), 297-334.
- Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada: A assessment. *Applied Linguistics 2*, 132-149.
- Cunningham, A. & Stanovich, K. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*, 16, 934-945.
- Dale, E. (1965). Vocabulary measurement: techniques and major findings. *English, 42*, (8), 895-901.

- Deumert, A. (2005). The unbearable lightness of being bilingual: English Afrikaans language contact in South Africa. *Language Sciences*, 27, 113-135.
- Doctor, E. A., & Knight, Z. (1993). Language and thought. In D.A Louw and D. J.A. Edwards. *Psychology: An Introduction for students in Southern Africa* (375-400). Johannesburg: Lexicon Publishing.
- Dolch, E. W., & Leeds, D. (1953). Vocabulary tests and depth of meaning. *Journal of Educational Research*, 4, 181-189.
- Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland and H. Wainer (Ed). *Differential Item Functioning* (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum.
- Dougherty Stahl, K. A., & Bravo, M. A. (2010). Contemporary classroom vocabulary assessment for content areas. *The Reading Teacher*, 63, (7), 566-578.
- Elwood, M. I. (1939). A preliminary note on the Stanford-Binet Scale. *Journal of Educational Psychology*, 30, 632-634.
- Farr, R., & Carey, R. F. (1986). *Reading: What can be measured?* (2nd ed.). Newark, DE: International Reading Association.
- Field, A. (2009). *Discovering statistics using SPSS* (3rd edition). Thousand Oaks: Sage Publication Inc.
- Foxcroft, C. (1997). Psychological testing in South Africa: Perspectives regarding ethical and fair practices. *European Journal Psychological Assessment*, 13 (3), 229-235.
- Foxcroft, C., & Roodt, G. (2009). *Introduction to psychological assessment in the South African context* (3rd ed) Cape Town: Oxford University Press Southern Africa (Pty) Ltd.
- Foxcroft, C., Roodt, G., & Abrahams, F. (2009). Psychological assessment: A brief retrospective overview. In C. Foxcroft and G. Roodt (Eds). *Introduction to psychological assessment in the South African context* (3rd ed) pp. 9-26. Cape Town: Oxford University Press Southern Africa (Pty) Ltd.

- 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 practioners. Unpublished final report, South Africa.
- Garvin, P. L., & Machiot, M. (1956). *The urbanisation of the Gurani language: A problem in language and culture*. In Wallace (Ed): Men and cultures. (Reprinted in Fishman, 1968).
- Gravetter, F. J., & Forzano, L.B. (2009). *Research methods for the behavioural sciences*. Canada: Wadsworth Cengage Learning.
- Grau, E. (2007). Using factor analysis and Cronbach's Alpha to ascertain relationships between questions of a dietary behaviour questionnaire. Proceedings of the Survey Research Methods Section, ASA. Accessed on 2 December 2011 at <http://www.amstat.org/section/srms/proceedings/y/2007.html>.
- Gray,S., Plante, E., Vance, R. & Henrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests administered to pre-school age children. *Language, Speech and Hearing Services in Schools, 30*, 196-206.
- Green, L. (2002). A descriptive study of African American English. *Research in linguistic and education, 15* (6), 673-690.
- Haitana, T., Pitana, S. & Rucklidge, J. J. (2010). Cultural biases in the Peabody Picture Vocabulary Test-III: Testing Tamariki in a New Zealand sample. *New Zealand Journal of Psychology, 39*, (3), 24-34.
- Hair, J. F., Black, W.C., Babin, B.J. & Anderson, R.E. (2010). *Multivariate data analysis* (7th ed). New Jersey: Prentice Hall.
- Halliday, M. A.K. & Hasen, R. (1989). *Language context and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.
- Hambleton, R.K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptations. *European Journal of Psychological Assessment, 11* ,(3) 147-157.
- Hambleton, R. K., Marendia, P. F., & Spielberger, C. D. (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Mahwah, NJ: Lawrence Erlbaum.

- Haupt, G.R. (2010). The evaluation of the group differences and item bias of the English version of a standardised test of academic language proficiency use across English and Xhosa first-language speakers. Unpublished MA thesis. University of the Western Cape.
- Henning, G. (1991). A study of the effects of contextualization and familiarization on responses to the TOEFL vocabulary test items. *TOEFL Research Reports*, 35. Princeton, NJ: Educational Testing Service.
- Henriksen, B. (1999). Three dimensions of vocabulary development. *SSLA*, 21, 303-317.
- Hoff, E. (2009). *Language Development*. Belmont, CA: Wadsworth, Cengage Learning.
- Hudson, R. A. (1996). *Sociolinguistics (2nd ed)*. Cambridge: Cambridge University Press.
- Huysamen, G. K. (2002). The relevance of the new APA Standards for educational psychological employment testing in South Africa. *South African Journal of Psychology*, 23 (2), 26-33.
- Ismail, G. (2010). Towards establishing the equivalence of the English version of the verbal analogies scale of the Woodcock Munoz Language Survey across English and Xhosa first language speakers. Unpublished MA thesis. University of the Western Cape.
- International Test Commission (2000). Guidelines for adapting educational and psychological tests. [Online] Available: http://www.intestcom.org/adapt_test.htm.
- Jalongo, M. R., & Sobolak, M. J. (2011). Supporting young children's vocabulary growth: the challenges, the benefits and evidence-based strategies. *Early childhood education Journal*, 38, 421-429.
- Jensen, A. R. (1975). Test bias and construct validity. Paper presented at the Annual meeting of the American Psychological association (83rd, Chicago, Illinois. Accessed 15 July 2011.
- Johnson, V.E. (2005). Comprehension of third person singular/s/ in AAE speaking children. *Language, Speech and Hearing Services in schools*, 36, 116-124.
- Kanjee, A., & Foxcroft, C. (2009). Cross-cultural test adaptation, translation and tests in multiple languages. In C. Foxcroft and G. Roodt (Eds). *Introduction to psychological assessment in the South African context (3rd edition pp.77-89)*. Cape Town: Oxford University Press Southern Africa (Pty) Ltd.
- Klare, G. R. (1975). Assessing readability. *Reading Research Quarterly*, 10, 62-102.

- Koch, E. (2007). The Monolingual testing of competence: acceptable practice or unfair exclusion. In P. Cuvelier, T. Du Plessis, M. Meeuwis and L. Teck (Ed). *Multilingual and exclusion: Policy, Practice and prospects* (pp. 79-103). Pretoria: Van Schaik.
- Koch, E. (2009). The case for bilingual language tests: a study of test adaptation and analysis. *Southern African Linguistics and Applied studies*, 27 (3), 301-317.
- Koch, E. & Dornbrack, J. (2008). The use of language criteria for admission to higher education in South Africa: Issues of bias and fairness investigated. *Southern African Linguistics and Applied Language Studies*, 26(3), 333–350.
- Koch, E., Landon, J., Jackson, M.J. & Foli, L. (2009). First brushstrokes initial comparative results on the Additive Bilingual Education Project (ABLE). *Southern African Linguistics and Applied Language Studies*, 27 (1), 93-111.
- Kubinger, K.D. (2003). On artificial results due to using factor analysis for dichotomous variables. *Psychology Science*, 45 (1), 106-110.
- Laufer, B. (1992). Reading in a foreign language: How does L2 lexical knowledge interact with the reader's general academic ability? *Journal of Research in Reading* 15, 95-103.
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics* 16, 307-322.
- Lewinski, R. J. (1948). Vocabulary and mental measurement: A quantitative investigation and review of research. *Journal of Genetic Psychology* 72, 247-281.
- Mahan, H. C., & Witmer, L. (1936). A note on the Stanford-Binet vocabulary test. *Journal of Applied Psychology* 20, 258-263.
- Mansour, G. (1993). Introduction to the symposium on linguistic imperialism. *World Englishes*, 12 (3), 335-336.
- Meara, P. (1996a). The dimension of lexical competence. In G. Brown, K. Malmkjaer and J. Williams (Eds). *Performance and competence in Second language acquisition* (pp.35-53). Cambridge: Cambridge university Press.

- Meiring, D., Van de Vijver, A.R.J., Rothmann, S., & Barrick (2005). Construct, item, and method bias of cognitive and personality tests in South Africa. *South African Journal of Psychology*, 31 (1), 1-8.
- Melka, F. (1997). Receptive vs productive aspects of vocabulary. In Schmitt and McCarthy (eds), *Vocabulary, description, acquisition and pedagogy* (pp.84-102). Cambridge: Cambridge University Press.
- Millett, J., Atwill, K., Blanchard, J. & Gorin, J. (2008). The validity of receptive and expressive vocabulary measures with Spanish-speaking kindergarteners learning English. *Reading Psychology*, 24, 534-551.
- Messick, S., (1989). Validity. In R. L. Linn (ed). *Educational Measurement*. New York: Macmillan.
- Nagy, W., & Anderson, R. C. (1984). How many words are there in printed school English? *Reading Research Quarterly* 19, 304-330.
- Nation, K., Clarke, P., Marshall, C. M., & Durand, M. (2004). Hidden language impairments in children: parallels between poor reading comprehension and specific language impairment? *Journal of Speech, Language and Hearing Research*, 27, 377-391.
- Nation, P. (1993b). Using dictionaries to estimate vocabulary size: essential but rarely followed procedures. *Language testing* 10, 27-40.
- Nunnally, J. C. (1967). *Psychometric theory* (2nd ed). New York: McGraw-Hill.
- Oakland, T. (2004). Use of educational and psychological tests internationally. *Applied Psychology: An International Review*, 53 (2), 157-172.
- O'Rourke, J. P. (1974). *Toward a science of vocabulary development*. Netherlands: The Hague.
- Palmer, S. (2004). Homai te Waiora kiAhou: A tool for the measurement of wellbeing among Maori-the evidence of construct validity. *New Zealand Journal of Psychology*, 33, 2, 50-58.
- Pawley, A., & Syder, F. H., (1983). Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J. C. Richards and R. W. Schmidt (Eds). *Language and Communication*. London: Longman.

- Pearson, P., Hiebert, E., & Kamil, M. (2007). Vocabulary assessment: what we know and what we need to learn. *Reading Research Quarterly* 42, 2, 282-296.
- Perfetti, C. A., & Hart, L. (2001). The lexical quality hypothesis . In L. Verhoeven, C. Elbro and P. Reitsma (Eds). *Precursors of functional literacy* (pp. 189-214). Oxford: oxford university Press.
- Pride, J. B., & Holmes, Y. J. (1979). *Sociolinguistics*. London: Penguin.
- Phillips, G., McNaughton, S., & MacDonald, S. (2004). Managing the mismatch: enhancing early literacy progress for children with diverse language and cultural identities in mainstream urban schools in New Zealand, *Journal of Educational Psychology*, 96, (2), 309-323.
- Poortinga, Y. H. (1989) Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology*, 24, 737-756.
- Quian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52,(3), 513-536.
- Raptis, H. (1997). Is second language reading vocabulary best learned by reading? *The Canadian Modern Language review*, 53, (3), 565-580.
- Raven, J. C. (1948). The comparative assessment of intellectual ability. *British Journal of Psychology*, 39, 12-19.
- Read, J. (1997). Vocabulary testing. In N. Schmitt & M. McCarthy (Eds). *Vocabulary description, acquisition and pedagogy* (pp. 303-321). UK: Cambridge University Press.
- Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.
- Reese, E., & Cox, A. (1999). Quality of adult book reading affects children's emergent literacy. *Developmental Psychology*, 35, (1), 20-28.
- Reese, E., & Read, S. (2000). Predictive validity of the New Zealand MacArthur Communicative Development Inventory: Words and sentences. *Journal of Child Language*, 27, 255-266.
- Restrepo, M.A., Schwanenflugel, P.J., Blake, J. Neuharth-Pritchett, S., Cramer, S.E., & Ruston, H. P. (2006). Performance on the PPVT-III and the EVT: Applicability of the measures with African American and European American preschool children. *Language Speech, and Learning services in schools*, 37, 17-27.

- Ricketts, J., Nation, K. & Bishop, P.V.M. (2007). Vocabulary is important for some, but not all reading skills. *Scientific Studies of Reading*, 11 (3), 235-257.
- Rock, D.A. & Stenner, J. (2005). Assessment issues in the testing of children at school entry. *The Future of Children*, 15 (1), 15-34.
- Roodt, G. (2009). Reliability: Basic concepts and measures. In C. Foxcroft and G. Roodt (eds). Introduction to psychological assessment in the South African context. (pp.45-53). Cape Town: Oxford University Press.
- Roomaney, R. (2010). Towards establishing the equivalence of the isiXhosa and English versions of the Woodcock Munoz Language Survey: An item and construct bias analysis of the verbal analogies. Unpublished MA thesis. University of the Western Cape.
- Rossier, J. (2004). An analysis of the cross-cultural equivalence of some frequently used personality inventories. International perspectives on career development. Symposium conducted at a joint meeting of the International Association for Educational and Vocational Guidance and the National Career Development Association, San Francisco.
- Schaap, P., & Vermeulen, T. (2008). The construct equivalence of the PIB/SPEEX motivation index for job applicants from diverse cultural backgrounds. *SA Journal of Industrial Psychology*, 29, 2, 49-59.
- Schedl, M., Thomas, N., & way, N. (1995). An investigation of proposed revisions to Section 3 of the TOEFL test. *TOEFL Research Reports*, 47, Princeton, NJ: Educational testing Service.
- Schmitt, N. (1996). Uses and abuses of coefficient Alpha. *Psychological Assessment*, 8, (4), 350-353.
- Schmitt, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language testing* 16, (2), 189-216.
- Schmitt, N., & Meara, P. (1997). Research vocabulary through a word knowledge framework word associations and verbal suffixes. *Studies in Second language Acquisition* 19, 17-36.
- Silo, U.L. (2010). An evaluation of group differences and items bias, across rural isiXhosa learners of urban isiXhosa learners of the isiXhosa version of the Woodcock Munoz Language Survey (WMLS). Unpublished MA thesis, University of the Western Cape.

- Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
- Sireci, S. G., & Khaliq, S. N. (2002). Comparing the psychometric properties of monolingual and dual language test forms. *Centre for Education Research No.458*. Amherst, MA: School of Education, University of Massachusetts.
- Spache, G. (1943). The vocabulary tests of the revised Stanford-Binet as independent measures of intelligence. *Journal of Educational Research*, 36, 512-516.
- Spolsky, B. (1995). *Measured words*. Oxford: Oxford University Press.
- Stahl, S.A., & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, 56, (2), 72-110.
- Stahl, S. A., & Nagy, W. E. (2006). *Teaching word meanings*. Mahwah, NJ: Lawrence Erlbaum.
- Stalnaker, J. M., & Kurath, W. (1935). A comparison of two types of foreign language vocabulary test. *Journal of Educational Psychology* 26, 435-442.
- Streiner, D. L., (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, 80, (1), 99-103.
- Stockman, I. (2000). The new PPVT-III: An illusion of unbiased assessment? *Language, speech, and hearing services in schools*, 31, 340-353.
- Sutarsyah, C. P., Nation & Kennedy, G. (1974). How useful is EAP vocabulary for ESP? A corpus based case study. *RGCC Journal*, 25, 34-50.
- Takala, S. (1984). Evaluation of student's knowledge of English vocabulary in the Finnish comprehensive school. Reports from the Institute for Educational Research, 350. Jyvaskyla: University of Jyvaskyla.
- Tannenbaum, K.R., Torgesen, J.K., & Wagner, R.K. (2006). Relationship between word knowledge and reading comprehension in third grade children. *Scientific Studies of reading*, 10 (4), 381-398.
- TenBerge, J. M. F. (1986) Rotation to perfect congruence and the cross validation of component weights across populations. *Multivariate Behavioural Research*, 21 (1), 41-64.

- Terman, L. M. (1918). Vocabulary test as a measure of intelligence. *Journal of Educational Psychology*, 9, 452-466.
- Thipa, H. M. (1989). The difference between rural and urban Xhosa varieties: A sociolinguistic study. D. Phil. Thesis. Pietermaritzburg: university of Natal.
- Van de Vijver, F. J. R. (1998). Towards a theory of bias and equivalence. *ZUMA-Nachrichten Special*, Retrived 04 March 2011.
- Van de Vijver, F.J.R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks: Sage.
- Van de Vijver, F.J. R., & Poortinga, Y. H. (2002). Structural Equivalence in multilevel data. *Journal of Cross-Cultural Psychology*, 33 (2), 141-156.
- Van de Vijver, A. J. R., & Rothmann, S. (2004). Assessment in multicultural groups: The South African case. *South African Journal of Industrial psychology*, 30 (4), 1-7.
- Van der Oord , S., Van der Meulen, E. M., Prins, P. J. M., Oosterlaan, J., Buitelaar J. K., & Emmelkamp P.M.G. (2005). A psychometric evaluation of the social skills rating system in children with attention deficit hyperactivity disorder. *Behaviour Research and Therapy*, 43, 733–746.
- Verhoeven, L., & Van Leeuwe, J. (2008). Prediction of the development of reading comprehension: A longitudinal study. *Applied Cognitive Psychology*, 22, 407-423.
- Washington, J. A., & Craig, H. K. (1999). Performance of at-risk, African American preschoolers on the Peabody Picture Vocabulary Test-III. *Language, Speech and Hearing services in Schools*, 30, 75-82.
- Webb, V. (2002). *Language in South Africa: The role of language in national transformation, reconstruction and development*. Pretoria: University of Pretoria.
- Weshe, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: depth vs breadth. *The Canadian Modern Language Review*, 53, (1), 13-40.
- Woodcock, R.W., & Munoz-Sandoval, A.F. (2001). *Comprehensive Manual: Woodcock-Munoz Language Survey Normative Update*. Itasca, IL: Riverside Publishing.

Yiakoumetti, A. (2006). A bidialectal programme for the learning of standard Modern Greek in Cyprus, *Applied Linguistics*, 27 (2), 265-317.

Zumbo, B. D. (1999). A handbook on the theory and methods of dif: logistic regression modelling as a unitary framework for binary and Likert type (ordinal) item score. Ottawa: Directorate of Human research and evaluation, Department of National Defense. Assessed 23/02/2011 at <http://www.educ.ca/faculty/Zumbo/DIF/handbook>. Pdf.

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



APPENDIX A

PERMISSION LETTER FROM PUBLISHERS IN TERMS OF ADAPTING THE WWLS FROM ENGLISH TO ISIXHOSA



December 8, 2005

Via electronic mail transmission

Ms. Maria Hansford, Copyright Officer
Legal Services
Summerstrand Campus (South)
Nelson Mandela Metropolitan University
Private Bag X6058
Port Elizabeth 6000

Dear Ms. Hansford:

This will serve as written confirmation that The Riverside Publishing Company has granted Beverley L. Burkett, Project Leader: Language Education of the Academic Development Unit of Nelson Mandela Metropolitan University (the "University") permission to translate and adapt the *Woodcock-Muñoz Language Survey – Revised, English Version (WMLS-R)* ("the material") into isiXhosa, a South African indigenous language, for research on language development of Xhosa-speaking children. Use of the material referenced herein may be used by the University only and may only be used under the following conditions:

1. There will be no deletions, additions, or other changes to the material without the prior written permission of The Riverside Publishing Company ("Publisher").
2. The permission granted is non-exclusive and is not transferable to other persons or institutions.
3. The permission is limited to the material as identified above for the purpose as stated. The translated test may not be used for any other purpose or otherwise reproduced, used, published, or distributed. Under no circumstances may the University receive any remuneration of any kind in consideration of the translated test or the material.
4. In using the material, it is understood that the Publisher protects the material as valuable, confidential and proprietary information of Publisher. You shall take all reasonable steps to ensure that the material and other Publisher confidential information embodied therein are protected and used only for evaluation purposes.
5. Credit will be given as follows:

"Copyright © 2005 by The Riverside Publishing Company. All rights reserved. Translated and adapted from the *Woodcock-Muñoz Language Survey – Revised, English Version* by Richard Woodcock, Ana F. Muñoz-Sandoval, Mary Ruff, and Criselda G. Alvarado. No part of this work may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, or by any information

425 Spring Lake Drive

Itasca, Illinois

601 43-2079

800.323.9540
customer service

800.767.8420
general

www.riverpub.com

APPENDIX B

NMMU ETHICS APPROVAL LETTER

4-APR-2007 08:39

NC SE KOCH

241 5661 773

PAGE



**Nelson Mandela
Metropolitan
University**

for tomorrow

• PO Box 77000 • Nelson Mandela Metropolitan University
• Port Elizabeth • 6031 • South Africa • www.nmmu.ac.za

Summerstrand South Campus

Human Ethics Committee

Tel . +27 (0)41 504-2354 Fax. +27 (0)41 583-3152

Yvonne.smith@upe.ac.za

Contact person: Y Smith

12 May 2005

Dr E Koch & Ms B Burkett
NMMU
Bldg 07, Ground Floor

Dear Dr Koch

RESEARCH PROJECT FOR ETHICS APPROVAL

The proposed project entitled *A longitudinal study on the effect of additive bilingual education on the academic achievement, cognitive development and language proficiency of rural Xhosa children* was submitted for approval in April 2005.

The proposal was accepted without any amendments.

We wish you well with the study.

Sincerely

A handwritten signature in black ink, appearing to read 'Bent Potgieter'.

**PROF B POTGIETER
ACTING CHAIRPERSON**

**Cc: Members of the Human Ethics Committee
Research Administration Office, UPE
Faculty Officer, Faculty of Health Sciences, UPE**

APPENDIX C

PERMISSION FROM THE EASTERN CAPE EDUCATION DEPARTMENT



**DEPARTMENT OF EDUCATION
(PROVINCE OF THE EASTERN CAPE)**

PORT ELIZABETH DISTRICT OFFICE

Private Bag X3015, North End, Port Elizabeth, 6056
Ethel Valentine Building, Sutton Street, Sidwell, Port Elizabeth
Tel: (041) 403 4420 / Fax: (041) 451 0193
e-mail address: samuel.snayer@edu.ecprov.gov.za



DISTRICT DIRECTOR: MR S. SNAYER

The Research Coordinator
APAP
NMMU

Dear Ms Koch

RESEARCH IN SCHOOLS

I refer to your letter (unfortunately undated) in which you request permission to conduct research in schools in Port Elizabeth.

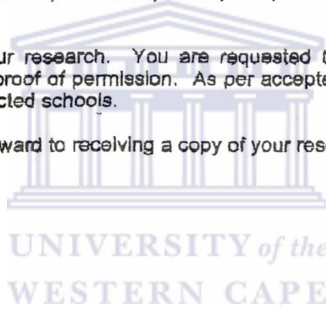
Permission is hereby granted for your research. You are requested to produce a copy of this letter to the principals of your chosen schools as proof of permission. As per accepted protocol you are further requested to abide by the internal rules of your selected schools.

I wish you the best of luck and look forward to receiving a copy of your research results.

Sincerely

**S. SNAYER
DISTRICT DIRECTOR: PORT ELIZABETH**

19 April 2005



APPENDIX D

INFORMED CONSENT (PARENTS ENGLISH VERSION)



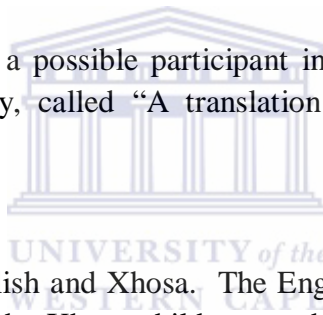
**Nelson Mandela
Metropolitan
University**

f o r t o m o r r o w

E-mail: elize.koch@nmmu.ac.za

Dear Parent

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



The test is available in both English and Xhosa. The English children will be tested on the English version of the test, and the Xhosa children on the Xhosa version of the test. The testing will take about one hour, and will be conducted at the school. Permission for this research project has been obtained from both the district manager and the school principal.

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

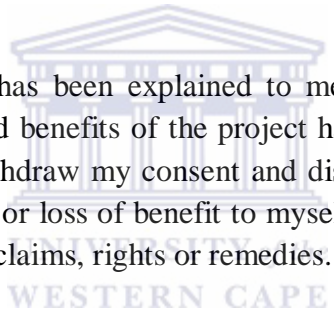
Regards

Dr. Elize Koch

Main Researcher.

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.”
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.”
5. “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.”
6. “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 (parent):.....Date.....signature

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.....

APPENDIX E

INFORMED CONSENT (PARENT'S ISIXHOSA VERSION)



**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.

IFOMU YESIVUMELWANO YAKWA

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.....

APPENDIX F

PERMISSION TO RE-ANALYSE THE DATA



UNIVERSITY *of the* WESTERN CAPE

DEPARTMENT OF PSYCHOLOGY

Private Bag X 17, Bellville 7535, South Africa, Telephone: (021) 959-2283/2453
Fax: (021) 959-3515 Telex: 52 6661

4/3/2011

TO WHOM IT MAY CONCERN

I hereby give Qunita Brown permission to use the data originally collected on the Xhosa versions of the Woodcock Munoz Language Survey for a bigger research study, called "Adapting a test of academic language proficiency from English into isiXhosa" for the purposes of a secondary data analysis. The data that she may use will be limited to the Picture Vocabulary scale, and will be available for re-analysis only for her MA thesis study. Any articles or presentations flowing from this thesis will be co-authored by the principal investigator.

Regards

A handwritten signature in black ink, appearing to read 'E. Koch', is written over a horizontal line.

Prof Elize Koch
Principal investigator

A faint, light blue watermark of the University of the Western Cape logo is visible in the background. It features a classical building with columns and the text 'UNIVERSITY of the WESTERN CAPE' below it.