TOWARDS ESTABLISHING THE EQUIVALENCE OF THE ENGLISH VERSION OF THE VERBAL ANALOGIES SCALE OF THE WOODCOCK MUÑOZ LANGUAGE SURVEY ACROSS ENGLISH AND XHOSA FIRST LANGUAGE SPEAKERS

Ghouwa Ismail

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

Supervisor: Prof S. E. Koch

September 2010

Keywords: Verbal Analogies Scale, differential item functioning, Woodcock Muñoz Language Survey, bilingualism, cognitive academic language proficiency, threshold theory, verbal reasoning, bias and equivalence theory, secondary data analysis, exploratory factor analysis.

ABSTRACT

In the majority of the schools in South Africa (SA), learners commence education in English. This English milieu poses a considerable challenge for English second-language speakers. In an attempt to bridge the gap between English as the main medium of instruction and the nine indigenous languages of the country and assist with the implementation of mother-tongue based bilingual education, this study focuses on the cross-validation of a monolingual English test used in the assessment of multilingual or bilingual learners in the South African context. This test, namely the Woodcock Muñoz Language Survey (WMLS), is extensively used in the United States in Additive Bilingual Education in the country. The present study is a sub-study of a broader study, in which the original WMLS (American-English version) was adapted into SA English and Xhosa. For this specific sub-study, the researcher was interested in investigating the scalar equivalence of the adapted English version of the Verbal Analogies (VA) subscale of the WMLS across English first-language speakers and Xhosa first-language speakers. This was achieved by utilising differential item functioning (DIF) and construct bias statistical techniques. The Mantel-Haenszel DIF detection method was employed to detect DIF, while construct equivalence was examined by means of exploratory factor analysis (EFA) utilising an *a priori* two-factor structure. The Tucker's phi coefficient was used to assess the congruence of the construct across the two language groups. The sample consisted of 192 English and 193 Xhosa first-language learners, who were selected from "ex Model C" and "previously disadvantaged" schools in the Port Elizabeth and Grahamstown region. The main findings of this study indicated that the adapted English version of the VA scale displayed DIF items across the two language groups. Moreover, construct equivalence could only be established for one factor across the two language groups, as the second factor displayed non-negligible incongruities even after the removal of DIF items.

DECLARATION

I declare that “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 speaking learners” is my own work, that it has not been submitted before for any degree or examination in any other university, and that all sources I have utilized or cited have been indicated and acknowledged as complete references.

Ghouwa Ismail September 2010

Signed: ....

DEDICATION

To my creator, for affording me the opportunities that none of the members of my immediate family has had, and for blessing me with the ability to nurture my dream, the foresight to choose wisely and the wisdom to learn even in failure.

ACKNOWLEDGEMENTS

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

My family, especially my Mom, for being my rock. There is no way I could have coped without knowing they were behind me every step of the way. There are not enough words in the world to express my love and gratitude.

My supervisor, Prof. Elize Koch, for always being available 24/7, especially at the crucial times when I needed assistance, notwithstanding her hectic schedule. Thanks for making the time to read even on the plane! For all the opportunities she has given me, for her confidence in my ability, and for imparting vast amounts of her knowledge to me over the past couple of years. Her passion and drive has inspired me to far exceed my wildest dreams.

My M.A. (Research) colleagues, Naiema, Rizwana, Janine, Jade, Maša and Renier, for the friendships extended and cemented in this brief period of time. I hope our friendships last a lifetime. Remember guys we'll always have Lamberts Bay! A special thanks to Riz who was always there when I needed someone to talk to. Thanks for helping me deal with what life had thrown at me this year. It has been brutal. Remember I am available for baby-sitting duties!

To Serena Williams and Zuhayr Kafaar for selflessly offering up of their valuable time whenever I was in need of assistance.

My mentor Mr Adam for allowing me the space to grow in my own time and constantly causing me to reflect long after our sessions were done.

Last but certainly not least, to the National Research Foundation for funding my studies this year and allowing me financial peace of mind to complete my Masters Degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF FORMULAE

CHAPTER ONE

INTRODUCTION

1.1. Introduction

Across the globe the phenomenon of bilingualism and in particular bilingual education within culturally and linguistically diverse educational contexts, has created a surge of interest in contemporary social research. While bilingualism as a cognitive phenomenon has been researched for quite some time, its nature is both theoretically and empirically elusive, eliciting as many questions as answers. In addition, even though bilingual education provides a platform for the promotion of higher levels of communicative proficiency and addresses the educational needs of language-minority students in many countries (Lindholm-Leary, 2003), there remains a palpable tension regarding what bilingual education is supposed to achieve and what it should consist of.

In South Africa with a language policy adopting 11 official languages into its Constitution, a Department of Education advocating monolingual instruction for primary school learners, and a Language in Education Policy (LiEP) that recognises the critical importance of first-language instruction (Language-in-education-policy, 1997) there is a dire need for the proliferation of literature surrounding bilingual education, particularly surrounding the possible advantages associated with additive bilingual education. Additive bilingual education is a form of education that promotes adding a second language to a student's repertoire while he or she continues to develop conceptually and academically in the first language (Cummins, 2000; Koch, 2009).

Although the policy encourages multilingualism, and the use of first-language instruction, in actuality the majority of the schools in South Africa commence in English and to a lesser degree Afrikaans. The popularity of English-medium education in this country has much to do with the history of apartheid, when Bantu education was used to separate and discriminate (Heugh, 2003).

This form of education was used to limit access and legitimate the linguistic practices of those individuals already in power. The country is still suffering the aftershock of the apartheid era and teachers and learners are still suffering the oppression, ferocity and passivity that characterised apartheid education. Learners are still exposed to prescriptive mediums of instruction, leading to learners not developing adequate literacy proficiency levels in their native language, and subsequently adversely affecting academic success.

The initial instruction in English means that many South African children, even after almost two decades of freedom, still continue to be inadequately educated (Alexander, 2002). The current situation in the classroom still demonstrates subtractive language-in-education practices (using a second language at the expense of the first language) with early-exit mother-tongue education, despite the implementation of the LiEP. One example is where African language learners are educated in their mother tongue as the language of learning and teaching (LoLT) until Grade 3, with an abrupt switch to English as the sole LoLT from Grade 4 (Koch, 2009), while another classroom scenario would see the primary LoLT as Xhosa in the foundation phase, with considerable code-switching in English taking place. Furthermore English is used as the medium of environmental print and the sole LoLT beyond the foundation phase (Koch, 2009).

The reality is that English is taught in almost all the schools either as a primary language or an additional language, and is the generally chosen language for examination at Grade 12 level. Thus regardless of the learners' mother tongue, all subjects in the majority of schools in South Africa are expected to be taught in English, with the exception of one or two languages, depending on the number of additional languages offered by the school.

In an attempt to bridge the gap between English as the main medium of instruction and the nine indigenous languages of the country, this study focuses on cross-validating a monolingual test used in the assessment of multilingual or bilingual learners within the South African context. For the purpose of this paper a distinction is made between monolingual testing and cross-lingual or cross-cultural testing. “Cross-lingual and cross-cultural testing” refers to tests and assessments that have been adapted or translated for use across diverse groups, while “monolingual testing” refers to tests that are available in only one language but are administered across many diverse language groups.

1.2. Background

The current form of “bilingual education” in South Africa has come under fire, and has been regarded as unsuccessful and markedly retrogressive in nature (Alexander, 2000; Alexander, 2002; Heugh, 2003; Kamwangamalu, 2000). The remnants and dominance of the previous colonial and apartheid practices in education provided the impetus to challenge traditional language educational models to meet the increasingly diverse needs of the various indigenous populations within this country in opposition to the prevailing English dominance in education.

Bilingual education is thus regarded as much more than a technique or pedagogy. Bilingual education within the realms of education is viewed as a means of equalising opportunities, and is underpinned by social justice, as well as supporting social practices for learning (García, 2009). The term "bilingualism", on the other hand, remains nebulous in definition and complex in nature, but broadly refers to an individual's ability to process two languages. Williams and Snipper (1990) expand on this definition and typify bilingualism as the ability to process two languages in every one of the four domains in the respective languages, namely listening, speaking, reading and writing. Thus, proficiency in two languages, for example English and Xhosa, would entail understanding the message in both languages, providing a context-appropriate response, and being able to read, write and understand a message in black and white (Williams & Snipper, 1990).

Cummins (2000) is of the opinion that in particular, additive bilingualism is associated with enhanced cognitive, academic and linguistic development. Research indicates that proficiency attained by students via additive bilingual education (i.e. education in their primary language) demonstrates important influences on their academic and intellectual development. The data unequivocally demonstrates the association between additive bilingualism and positive linguistic and academic consequences (Cummins, 2000). However, this only occurs when both languages are encouraged to develop. Cummins (2000) further asserts that a balanced bilingual individual will only achieve positive cognitive advantages when they advance across two distinct thresholds or levels of proficiency in both languages. This refers to the "threshold hypothesis" (Cummins, 2000) which postulates that bilingual children need to achieve certain milestones or threshold levels of linguistic competence in both languages in order to benefit cognitively and avoid cognitive drawbacks (Lenters, 2004).

What this implies is that a bilingual learner who is experiencing difficulties in reading in one of his/her languages, may not have reached an adequate level of language proficiency in both his primary or subsequent language. In other words, the problem may not lie in the learner's reading ability *per se*, but on the language proficiency level acquired. Thus for example in South African classrooms, if Xhosa students experience difficulties in reading ability in English, which is the main language of learning and teaching, then this should not be regarded as cognitive impairment on the part of the learners. According to the threshold hypothesis, in actuality the learners in the South African classrooms have not been given the opportunity to achieve the milestones of linguistic competence in their primary language, and thus experience drawbacks in the acquisition of the second language, which ultimately may have led to the difficulties experienced in reading ability in their second language. Ultimately the more learners know and understand, the easier it is for them to make sense of academic language, since there is internal support for understanding the messages.

1.3. Bilingual Education in South Africa

In South African classrooms, English and to a lesser degree Afrikaans, are the dominant languages of learning and teaching for learners, in spite of their diverse language backgrounds. This prescriptive milieu poses a considerable challenge for the native speakers of African languages (Alexander, 2000; Alexander, 2002; Heugh, 2003). How can African-language learners obtain a firm foundation of knowledge and skills in their first language in order that it can be transferred to the second language when the primary medium of instruction in South Africa (namely English and to some degree Afrikaans) does not accommodate the nine indigenous languages?

Furthermore African language learners are at a disadvantage since the majority of learning resources are available in English only. The barriers they experience to learning because of their limited English proficiency (Nel, 2005), results in their not being skilled enough to learn the complex concepts in mathematics, science, geography or history in their second language. In addition, many African-language educators lack the education, knowledge, tools, and time to assist Xhosa learners to attain their full potential, because of their limited English proficiency levels (Prinsloo, 2005). In an attempt to address this issue, a project called the "Additive Bilingual Education Project" (ABLE) was initiated in 2003. The objectives of the project are: (1) to assess the long-term effects of additive bilingual curriculum delivery of cognitive development, academic achievement, and language proficiency in English and Xhosa by comparing this group with groups from similar rural contexts who have experienced forms of subtractive bilingual education; (2) to describe the form in which additive bilingual curriculum delivery takes place in South Africa, and (3) to describe the effect the additive bilingual model has on the learners, their teachers, their school and the wider community (Koch, Landon, Jackson & Foli, 2009).

As previously mentioned, additive bilingual education is a form of education where the primary language of the learners is used for cognitive and literacy development, while simultaneously learning a second language. In contrast, "subtractive bilingual education" refers to the acquisition of a second language at the expense of the first language (Koch et al., 2009). The first round of results of the ABLE project which used an adapted version of an American language test (adapted from English to Xhosa) provided evidence for the educational advantages of the additive model in the SA context, and for the development of the bilingual literacy of the learners in question, which is in line with global research findings of this nature (Cummins, 2000; Koch, 2009; Lindholm-Leary, 2003; Thomas & Collier, 1997; Thomas & Collier, 2002). Thus, assessing the language proficiency of the learners in English as well as Xhosa is in conformity with one of the project's primary aims, making the measuring instrument of vital importance.

Test development and adaptation in a multicultural, multilingual society in transition is a complex process (Foxcroft, 1997) and advancement in this regard has been more gradual than anticipated. South African society is heterogeneous in terms of factors that are considered to moderate performance on psychological tests, and one such variable is language proficiency. Language proficiency has often been cited as a potential source of bias in relation to ability testing (Foxcroft, 1997).

As a result, the aforementioned broader study has various sub-studies which will be discussed in more detail in Chapter 3. The present study is one of the sub-studies which focus on the equivalence of the adapted English version of the Verbal Analogies Scale of the Woodcock Muñoz Language Survey (WMLS) for use across English and Xhosa first-language learners. More specifically, the present study will evaluate item bias and construct equivalence in order to determine whether this scale is suitable for utilisation across English first-language and Xhosa first-language learners.

It is important to take cognisance of the fact that in using this Verbal Analogies (VA) scale, the assertion will be made that it can be used to make the same statements with regard to verbal reasoning of both English first-language learners and Xhosa first-language learners. In other words this scale was used as an English monolingual scale to assess not only English first-language learners but also Xhosa first-language learners. However, in order for the scores obtained on the scale and the construct of verbal reasoning of the two groups to be comparable, the equivalence of the scale needs to be established. These concepts will be elucidated on in Chapter 2 in the section on Theoretical Framework.

1.4. The Woodcock Muñoz Language Survey (WMLS)

Language proficiency is difficult to measure despite a wide range of instruments currently available. An instrument that is used extensively in the USA for the evaluation of bilingual programmes and learner classification is the WMLS, which is claimed to measure cognitive academic language proficiency (CALP), a construct that will be explained in Chapter 2 (Woodcock & Muñoz-Sandoval, 2001). The WMLS's English version was standardised on approximately 6,300 participants in the USA, and the Spanish version on approximately 2,000 participants in Argentina, Costa Rica, Mexico, Peru, Puerto Rico, Spain, and the USA (Woodcock and Muñoz-Sandoval, 2001). It is not normed for the South African population. At present this test is being implemented in South Africa and, as mentioned above, has been adapted into Xhosa as well as South African English (Koch, 2009).

For the purpose of this study, the focus was specifically on the adapted English version of the WMLS, and more specifically on the Verbal Analogies (VA) scale of the WMLS. In the USA this instrument has proved to be a useful predictor of reading (Laija-Rodríguez, Ochoa & Parker, 2006). However, according to Kao (1998), the WMLS test-makers give insufficient information about validity, and provide little explanation and clarity about the CALP construct, so it is difficult to ascertain exactly what is measured. This sub-study also addresses this issue with regard to the VA scale and explores the construct being measured in the scale.

1.5. Bias and Equivalence

The research of this study will be guided by the theoretical framework of equivalence and bias within psychometric theory. Essentially, “equivalence” refers to the measurement level at which scores can be compared across cultures, while “bias” refers to nuisance factors affecting these test scores across different groups differentially. The aforementioned concepts will be defined and discussed in Chapter 2 under the heading Theoretical Framework.

This study will in particular focus on ascertaining whether the scores on the adapted English items can be compared across two groups, namely, the English first-language group (L1) and the Xhosa first-language group. According to Van de Vijver (1998) and colleagues, equivalence and bias provide slightly different perspectives of the same question, which is the extent to which scores have the same meaning across groups (Van de Vijver, 1998; Van de Vijver & Tanzer, 2004). This is regarded as the fundamental question in cross-cultural and cross-linguistic testing (Koch, 2009), and will be discussed as the thesis unfolds. In addition, it remains important for monolingual tests to be evaluated for equivalence across language groups as it produces the foundation for the secondary objective of obtaining full-scale measurement equivalence, which makes test score comparisons across populations possible.

It is for this very reason (i.e. test score comparisons across populations) that South Africa has implemented a rigorous legal framework to protect its citizens, in order that they receive fair testing across language and cultural groups. This legal framework will be elucidated upon in the following chapter under the heading Theoretical Framework of Bias and Equivalence.

1.6. Rationale and Aims

The original English version of the WMLS was adapted for the South African context into both Xhosa and South African English to incorporate South African words and terms (Koch, 2009). As yet, the WMLS has not been normed for the South African population, and thus a complete psychometric properties dossier of the test for the South African context is not yet available, even though research is currently in progress (Koch, 2009). However, current research indicates that both the adapted English version and the Xhosa version of the WMLS demonstrate promising results on two of the scales of the test, namely the VA and the Letter-Word Identification (LWI) (Arendse, 2009; Haupt, 2009). Results on the adapted English version of the VA Scale indicate that good internal consistency was displayed across the English and Xhosa groups with a Cronbach's Alpha of 0.83 and 0.86 respectively (Haupt, 2009). In addition, a logistic regression differential item functioning (DIF) analysis indicated that only four items displayed DIF on this scale, with two items having large DIF and two items having moderate DIF (Haupt, 2009). The concept of DIF will be expounded in Chapter 2 under the theoretical framework subheading.

Furthermore, even though cross-linguistic validity of assessment measures are viewed as indispensable for multilingual adaptations and comparisons, the construct equivalence across the two language groups of the English version of the VA scale has not yet been established. According to Van de Vijver and Tanzer (2004), a lack of cross-linguistic equivalence may threaten the validity of an entire research study. In Arendse's study (2009) on the equivalence of the two language versions of the test, a factor analysis revealed two factors on the English version of the VA scale, substantiating the findings in Koch (2009), where a weighted multidimensional scaling analysis also displayed two dimensions on this scale. The first factor displayed structural equivalence across the two language versions, while the second factor was found to be inequivalent (Arendse, 2009).

The promising results displayed by the adapted English version of the VA Scale (Haupt, 2009) necessitates increasing focus on this scale in order to cross-validate (using a different DIF technique than was used in the previous research) as well as to refine the research so as to use this measure within the South African context. Cross-validation is of vital importance. It extends prior research, and addresses the methodological limitations of various DIF studies in general, as well as expanding on current information available on its psychometric characteristics. In bilingual education, this specific study is also important to ensure that the monolingual adapted English version is an equivalent test, in order to produce equivalent scores for both English and non-English speakers. This allows the researcher to end up with the same measurement scores across language groups.

This study's overall aim is thus to assess the scalar equivalence of the adapted English version of the VA scale of the WMLS across two language groups, namely an English first-language group and a Xhosa first-language group.

1.7. Objectives

The specific research aims are as follows:

Research aim 1: To evaluate the differential item functioning (DIF) of the English version of the VA scale across English and Xhosa first-language groups.

Research aim 2: To assess the construct equivalence of the English version of the VA scale across English and Xhosa first-language groups, initially with all the items included and subsequently with the DIF items removed.

The first chapter, namely the Introduction, has focused on the background of this study as well as located it within the context of bilingualism in South Africa, and looked at the Woodcock Muñoz Language Survey measuring instrument in a South African context. It further highlighted the rationale, aim and specific objectives of this study.

The second chapter, namely the Literature Review, will delineate the various studies that have been conducted in the field of verbal analogies or verbal reasoning as well as highlighting the use of verbal analogies as an assessment tool in a bilingual context. This chapter will conclude with a discussion on the theoretical framework within which this study is located.

Chapter 3, Methodology, will outline the research methodology of the main study as well as of the present study. Chapter 4 reports the results of the study, delineated by the various statistical procedures as well as the null hypothesis for objective 2. The fifth chapter, Discussion and Conclusion, provides an in-depth discussion generated by the previous chapter, and concludes with limitations and further recommendations.

CHAPTER TWO

THE MONOLINGUAL ASSESSMENT OF VERBAL REASONING UTILISING VERBAL ANALOGIES AS AN ASSESSMENT TOOL: ISSUES OF BIAS AND EQUIVALENCE

2.1. Introduction

Within the context of bilingual education in multilingual societies such as South Africa, it is important to have tests that produce equivalent scores across groups. Thus monolingual tests that are equivalent across language groups, as well as equivalent tests in more than one language, provide an important platform for comparison of children on the same level between language groups. This chapter develops the main argument for this thesis: scalar equivalence of a verbal analogy scale in English, across two language groups, by placing the study within the context of language proficiency and more specifically verbal reasoning or analogies. This chapter will also elucidate on monolingual testing, as well as the theoretical framework of bias and equivalence as these form an integral part of this thesis, which is set within the realms of psychometric theory.

2.2. Language Proficiency

2.2.1. Introduction

As was discussed in Chapter 1, the LiEP (Language-in-education-policy, 1997) asserts that a minimum of two languages must be taught in classrooms, consisting of a primary language (the native language of the learner) and an additional language (any of the eleven official languages in South Africa). However, even though this policy does not specify the language that must be employed as a medium of instruction, the fact is that the majority of schools in South Africa have adopted English as a medium of instruction. English has become the dominant language of communication, academia, business, and technology. As a result the implementation of the LiEP (Language-in-education-policy, 1997) falls short of creating the platform for equal educational opportunities to second-language English learners in bilingual education.

The Cognitive Academic Language Proficiency (CALP) construct is often cited in the literature as a milestone to second-language (L2) development and as having a significant relationship with academic achievement in the L2 (Cummins, 2000; Lindholm-Leary, 2003). This relates to Cummins's (2000) interdependence hypothesis, which maintains that once an individual has acquired CALP in their first language (L1) this forms the foundation for proficiency in L2 and as a result, as certain thresholds of linguistic competence are met in both languages, the learner experiences positive cognitive effects. CALP refers to decontextualised language skills and practices, and will be further elaborated on in this chapter under the heading Cognitive Academic Language Proficiency. The subsequent section however provides a brief overview of bilingualism.

2.2.2. Bilingualism

The literature on bilingual education and the intellectual abilities of bilingual individuals has often been characterised by contradictory findings. The relationship between bilingualism and cognitive ability at times indicated an advantage, sometimes a disadvantage, and at other times little difference (Bialystok, 2007; Cummins, 2000; Lenters, 2004; Lindholm-Leary, 2003; Williams & Snipper, 1990). According to Cummins (1986), even though specification and operationalisation of the nature of language proficiency of bilingual learners currently goes beyond the realm of empirical validation, evidence continues to mount that bilingualism is associated with enhanced metalinguistic functioning, as well as advantages in other aspects of cognitive performance. Lenters (2004) postulates that through bilingualism children experience important cognitive benefits such as cognitive flexibility, superior language skills, and a higher IQ. However, she is of the opinion that these benefits are only experienced when both languages develop to a certain point of proficiency (Lenters, 2004), thus lending itself to the threshold hypothesis of Cummins as discussed in Chapter 1.

As a result, bilingual proficiency can be considered on a continuum along which bilingual individuals fall at varying intervals, depending on the degree of strengths and cognitive characteristics they exhibit (Solana-Flores & Trumbell, 2003).

The next section deals with the construct of CALP and its centrality within the context of bilingualism in an educational environment.

2.2.3. Cognitive Academic Language Proficiency

Cummins (1986; 2000) has hypothesised that language proficiency, and by extension bilingual language proficiency, can be categorised into two basic constructs, namely CALP and Basic Interpersonal Communicative Skills (BICS). BICS, also referred to as conversational proficiency, can be defined as everyday communicative skills demonstrated in the articulation, vocabulary and grammar that occur in daily dialogue (i.e. contextualised language skills and practices). The more learners know and understand, the easier it is for them to make sense of academic language, since there is an internal support for understanding messages which is inherent in BICS (García, 2009).

Alternatively, CALP relates to proficiency in the academic context, the sort of language that students come across in the classroom. The skills exhibited will entail a semantic and abstract context, in other words, the higher-level language skills (i.e. hypothesising, evaluating, inferring, generalising and predicting) required for literacy and for cognitively demanding content in an academic context (Williams & Snipper, 1990). Students can usually master communicative skills in a L2 in about two years; however, full proficiency in the academic language in the L2 can take up to seven years (Cummins, 2000; Lindholm-Leary, 2003). Cummins (2000) contends that CALP is not superior to BICS, and that developmentally they are not detached but develop jointly within the matrix of social interaction. From the above discussion it is evident that CALP is an intrinsic aspect in academic contexts which assist learners in the development of vital higher-level language skills. Thus, verbal reasoning or analogies as discussed in the subsequent section, and fall within the realm of these higher-level language skills, may play a crucial role in the context of bilingual education and the assessment of CALP in L2.

2.3. Verbal Reasoning

2.3.1. Introduction

The ability to reason by analogy is generally considered a core component in the development of human cognition. It provides an important foundation for learning and classification, as well as for thought and explanation. Many theorists have attempted to explain verbal reasoning in terms of their understanding of the concept, and have produced varying theories (Goswami & Brown, 1990; Piaget & Inhelder, 1958; Piaget, Montangero & Billeter, 1977; Sternberg & Nigro, 1980; Sternberg & Rifkin, 1979). Early theorists espoused the notion that verbal reasoning was a developmentally sophisticated skill, while more recent research has evidence to the contrary (Goldman, Pellegrino, Parseghian & Sallis, 1982; Goswami, 1991; Goswami & Brown, 1990). In other words, proponents of the theory postulate that verbal reasoning is absent in young children and develops with age, while the opponents believe that verbal reasoning is inherent in the individual from a young age, but requires certain cues to access it. A detailed discussion will ensue as this section unfolds. One thing that remains evident throughout these theories despite their opposing views, is the fact that older individuals perform more spontaneously and more competently then younger individuals when utilising verbal reasoning or verbal analogies.

Verbal reasoning is viewed as encompassing higher-order reasoning skills which promote the ability to transfer knowledge to new situations, perform successfully on novel problems, and learn by integrating a variety of information from diverse contexts (Goswami & Brown, 1990). Holyoak and Thagard (1995) coined the term "mental leap" in their definition of verbal reasoning. They postulated that the act of formulating an analogy necessitates perceiving one thing as if it were another, and thus the perceiver is required to make a kind of "mental leap" between domains. This coincides with the generally accepted view of most researchers that verbal reasoning involves reasoning pertaining to relations, in particular with regard to relational similarity, in order that a correlation is ascertained between one set of relations and another (Goswami, 1991; Tagalakis & Keane, 2006). In other words, an individual recognises the relational similarity, for example that a dog is more related to a cat than to a camel. According to Goswami (1991) and Cummins (1992), this definition allows verbal reasoning to encompass problem-solving by using the solution to a known problem to solve a structurally similar problem. Thus, being able to identify these abstract similarities is the underlying attribute of verbal reasoning, which one could argue falls under the over-arching umbrella term CALP as postulated by Cummins. CALP, as previously mentioned, encompasses decontextualised language skills and practices, that is, higher-level language skills.

Suen (2005) on the other hand, refers to verbal reasoning and general intelligence as synonymous constructs and defines them as consisting of vocabulary, ability to determine relationships between words, understanding sentence structures, reasoning ability, and symbol manipulation ability. However, Burton and colleagues (2009) assert that the construct "verbal reasoning" is not synonymous with the construct "intelligence". Verbal reasoning involves a number of logically distinct cognitive operations as well as multiple dimensions such as breadth and depth of understanding, on which individual performance may vary. Burton, Welsh, Kostin and van Essen, 2009 are of the view that verbal reasoning is based on a set of cognitive and metacognitive skills such as planning the way to approach a learning task or evaluating the progress to completion of a task.

This definition of verbal reasoning, one could argue, coincides with Cummins' (2000) definition of cognitive academic language proficiency (i.e. the higher-level language skills required for literacy and for cognitively demanding content within an academic context). Thus verbal reasoning appears to be a complex skill that is essential for higher-order learning and thinking, and its development may be dependent on the correct educational milieu. In other words, exposure to higher-order thinking in L1 while exposed to L2 will ultimately lead to higher-order skills transfer as postulated by Cummins (2000) in his interdependence theory mentioned in the previous section. According to Brown as well as Goswami and Brown (1990; 1989), prominent analogical theorists have postulated that verbal reasoning is the principle means by which cognition develops, even though the way in which it develops (i.e. dependent on context or developmental stages) is still disagreed upon by these very theorists.

2.3.2. Verbal Reasoning in the Context of Developmental Theory

According to the developmental theory, verbal reasoning is a developmentally sophisticated skill. Piaget argues that prior to the stage of formal operations, children do not possess the cognitive capacity to represent the necessary relations to perform classical analogical tasks (Inhelder & Piaget, 1958). According to his theory the prerequisites for verbal reasoning as a function of formal operational thinking, are rooted in a series of abstractions made during the pre-operational and concrete operational periods (Goswami, 1991). Piaget and his colleagues (1977) based these theoretical claims on a series of studies using a form of item analogy. They tested children ranging in ages from 5 to 12 years with a pictorial version of the item-analogy task. This experimental method was based on picture sorting, where children were given sets of pictures to sort into pairs, and subsequently into sets of matching pairs. The intention was that the children would extrapolate analogies based on functional and causal relations.

Piaget found that when he presented children with the A:B:C:? tasks and asked them to select the D term in a pictorial set, younger children frequently relied on lower-order relations to solve the problem (Piaget et al., 1977). In other words, the children selected items that were associated to or resembled the C item. Piaget concluded that this failure to represent the higher-order relations between the A:B and C:D terms demonstrated that children are unable to exercise verbal reasoning before the stage of formal operations. He further argued that true understanding of verbal reasoning did not develop until early adolescence, which emerged during the stage of formal operations. Based on the aforementioned studies, Piaget and his colleagues were able to define three broad stages in the development of reasoning by analogy namely, pre-operational, concrete operational and formal operational (Goswami, 1991).

In a subsequent study Sternberg and colleagues corroborated Piaget's findings, discovering similar limitations in young children's verbal reasoning by observing an over-reliance on lower-order relations during analogical problem solving (Sternberg & Downing, 1982; Sternberg & Nigro, 1980). In addition many other studies have found evidence consistent with the developmental position espoused by Piaget (Gallagher & Wright, 1977; Levinson & Carpenter, 1974; Lunzer, 1965).

From the above discussion it becomes evident that Piagetian theory does not advocate that children perform tasks that are beyond their cognitive capabilities. Thus, in a classroom setting, the teacher merely prepares the environment for the learners' level of mental developmental or motor operations.

Vygotsky, on the other hand, viewed children's cognitive development from a contextual perspective. He asserted that in order to understand a child's cognitive development, one needs to look at the social processes from which their thinking is derived (Papalia, Olds & Feldman, 2004). Cognitive growth is viewed as a collaborative process between children, adults and their environment. Thus, Vygotsky perceived the child as a social being who is able to appropriate new patterns of thinking when learning alongside adults or advanced peers. This guidance is most effective in helping children cross the zone of proximal development (ZPD). ZPD refers to the gap between the child's current level of development and their potential level of development, in collaboration with more competent individuals (Papalia et al., 2004). Social interaction, therefore, supports the child's cognitive development in the ZPD, leading to a higher level of reasoning. According to Papalia et al. (2004), as the child becomes more competent in levels of reasoning, the responsibility for directing and monitoring learning will gradually shift from the adults to the children.

Thus, while Piagetian theory asserts that a learner is limited by their stage of development in a classroom, the Vygotsky approach challenges the child to work beyond their potential. Educators, therefore according to the Vygotskian approach, play a vital role in stimulating developmental processes of the learner in the classrooms, thus encouraging positive cognitive development.

2.3.3. Recent Studies on Verbal Reasoning

Recent studies have, however, revealed that children are capable of effectively reasoning by analogy at younger ages. Findings in these studies in contrast to Piaget and Sternberg's conclusions, demonstrated that children are able to reason analogically prior to the formal operations stage, provided that they possess the relevant domain knowledge used in the task (Goswami & Brown, 1989; Pierce & Gholson, 1994; Singer-Freeman & Goswami, 1999). These studies also indicated, to some degree as did Piaget and Sternberg that improvements in verbal reasoning are because of developmental milestones (Goswami, 1991; Zelazo & Müller, 2002).

The apparent dichotomy of recent theories as opposed to Piaget and Sternberg lies in the subtle nuances of these theories. While all theorists agree that verbal reasoning increases with age, Piaget believed it was completely absent prior to the stage of formal operations. Recent theorists, on the other hand, believe verbal reasoning is present from a very early age but is only activated based on what each theorist's theory espouses. In other words, if one views verbal reasoning from a relational familiarity hypothesis (Goswami, 1991) then analogical development depends critically on the conceptual knowledge of a particular child. "Conceptual knowledge" refers to knowledge rich in relationships and understanding linking discrete bits of information. Therefore, according to the relational familiarity hypothesis, children's performance on analogical reasoning depends on relational knowledge. Relational knowledge involves the underlying relations in tasks recognising commonalities between different domains in higher-order thinking, and forms the foundation to conceptual knowledge (Halford, 1996). As a result, this hypothesis views analogical development as critically dependent on the conceptual knowledge of the learner. Halford (1996) regards relational knowledge as central to mechanisms that are basic to human reasoning, such as analogy and planning.

Thus, as knowledge develops and is associated with knowledge in other domains in different ways, the ability to utilise more sophisticated analogies develops. According to Goswami (1998), the relational familiarity hypothesis is viewed as a core cognitive skill present even in infancy. Younger children revert to analogies by association (i.e. lower-order reasoning) (Piaget et al., 1977; Sternberg & Nigro, 1980) only when this relational knowledge required to reason on the basis of relational similarity is absent. A study by Goswami and Brown (1989) explored item analogies based on physical causal relations. The results of their study indicated that both analogical success and causal relational knowledge increased with age. The study further demonstrated that when children possess the sufficient domain knowledge, they are capable of succeeding on tasks of analogical reasoning. Therefore the data suggested that analogical reasoning in children is highly dependent on relational knowledge, and by extension, conceptual knowledge. From the results of the aforementioned study it can thus be inferred that if analogical reasoning is absent in individuals, it may be due to a lack of conceptual knowledge, and not because it is not inherent in the individual.

This may be an important point of departure in a South African context where many learners in the rural areas not only come from impoverished areas but in fact also receive a lower standard of education and therefore their development of conceptual knowledge, for example, will not be on par with that of learners in the urban areas. This could affect their perceived level of competency in analogical reasoning if measuring instruments are not wary of these differences.

2.4. An Assessment of Verbal Reasoning

2.4.1. Introduction

Analogical reasoning, according to Spearman (1927) and Sternberg (1977), is fundamental in human intelligence, and numerous forms of analogy tests are thus often used for measuring general ability. Ullstadius, Carlstedt and Gustafsson (2008) postulate that verbal analogies are probably the most frequently used form of analogy tests, as the detection of the principle underlying the relationship between two words, is in this case assumed to require reasoning processes that are crucial components in general ability. The items in such a test typically consist of the A:B:C:? tasks as previously discussed under the section Verbal Reasoning in the Context of the Developmental Theory. What this means is that each item consists of a stem of two words and a third word that should be matched to a correct answer among a number of fixed response alternatives in multiple-choice format (Ullstadius et al., 2008; Goswami, 1991; Piaget et al., 1977).

A representative example of a verbal analogy item is: wolf:dog :: tiger:? cat, boar, fish, kitten. The salient relationship and the main source of difficulty of the analogical item is usually considered to be inherent in the first word pair (Goswami, 1991; Ullstadius et al., 2008). Furthermore, the relationship between the first word pair and the third word often provides a clue that is necessary for the solution of a task (Goswami, 1991; Ullstadius et al., 2008). This relations linking is referred to by Piaget as requiring second-order relational understanding which according to him, only emerges or is possible during the stage of formal operations. This relationship is regarded as a key component in the reasoning process. In the above example, the word combination “pig” and “boar” must be compared to “dog” to restrict the field of association in order to arrive at felines.

A study investigating the perception of relationships used in analogical tasks indicated that these analogical tasks were underpinned by eight modes of relationships, such as class, similarities, membership, quantitative and opposites (Whitely, 1977). The study further revealed that what the analogy test in actuality measured was dependent on the composition of relationships. In a subsequent study of analogical reasoning in students, Sternberg and Nigro (1980) found that items involving synonyms, category membership[1], and linear ordering relationships[2] among the word stems were more difficult than those with antonymous and functional relationships[3]. These characteristics corroborate some of the categories found by Whitely. Bejar, Chaffin and Embertson (1991) elaborated on Whitely’s categorisation, formulating a taxonomy encompassing ten classes of relationships. According to their study, these ten classes were further subdivided into two categories, namely intentional relationships and pragmatic relationships (Bejar et al., 1991). Intentional relationships are based exclusively on attributes that are inherent to word meaning, while pragmatic relationships require knowledge about the world that transcend simple word meaning (Ullstadius et al., 2008). Bejar et al. (1991) postulated that analogies pertaining to intentional relationships are in general more complicated than those with pragmatic relationships. What their study also discovered was that verbal analogy tests measured not

[1] How items and categories are related to other categories.

[2] Any given change in an independent variable will always produce a corresponding change in the dependent variable (i.e. the relationship of direct proportionality).

[3] Representational systems invented or appropriated by children to represent a generalization of a relationship among quantities (Smith, 2003).

only general ability but also verbal ability (Bejar et al., 1991; Roccas & Moshinsky, 2003). Spearman postulates that general ability provides the key to understanding intelligence (Sternberg, 2006). According to Kuncel, Hezlett and Ones (cited in Aloe & Becker, 2009), verbal ability is regarded as a component of tests of general mental ability or reasoning ability.

From the above discussion one can thus argue that VA is an important construct in the context of bilingual education as it may be used as an important measurement tool in providing vital information to the learners' verbal reasoning development and ultimately their language proficiency in the L1 as well as monitoring whether and when it transfers to the L2.

2.4.2. Verbal Analogies as an assessment tool of verbal reasoning

Verbal reasoning tests and in particular verbal analogies are frequently used as assessments in the field of education. These particular assessments provide measures of verbal reasoning independent of curriculum content, thus providing the opportunity to compare performance of learners from diverse educational backgrounds (Primrose, Fuller & Littledyke, 2000). Whetton (1985) asserts that verbal reasoning tests are highly reliable and relatively good predictors of prospective academic performance. However, according to Primrose and colleagues (2000), these assessments are inherently biased as they are dependent on prior exposure to language, which has implications for socialisation, as well as specific cultural and environmental context. Primrose et al. (2000) contend that the acquired knowledge and skill measured in verbal reasoning assessments are those associated with language and its everyday use. They conducted a study in a preparatory school to investigate the stability of verbal reasoning assessment scores. These verbal reasoning measures were often used as part of the assessment criteria for admission into the school (Primrose et al., 2000).

The objective of the study was to explore the stability of the test scores as a prerequisite for developing a verbal reasoning ipsative reference mode of assessment. As a result this mode of assessment would be a self-comparison either in the same domain over time, or comparative to other domains within the same learner (Foxcroft, 2005). The results indicated that verbal reasoning test scores do not remain constant, but fluctuate over time. Thus, these tests or measures could only be used in assessing the learner's current level of achievement (Primrose et al., 2000), as scores are not fixed, finite measures of future academic potential. From the study Primrose and colleagues (2000) concluded that verbal reasoning assessments provide good measures of the levels of cognitive functioning at a particular point in time, independent of specific subject content.

Cummins (2000) is of the opinion that language proficiency tests seldom consider aspects of language that are crucial to academic success. Thus most such tests limit the construction of language proficiency to grammatical competence. The WMLS, though, which is available in English and Spanish, claims to assess CALP.   The test developers claim that not only does the test predict CALP, which traditional language proficiency test do not, but it will also predict when students will reach a certain level of proficiency in their school careers (Oakley, Urrabazo & Yang, 1998).

The VA scale on the WMLS measures listening and speaking skills[4], and the test developers claim that it assesses an individual's ability to complete oral analogies, which necessitates verbal comprehension and verbal reasoning. It is postulated that these kinds of verbal

[4] The test format provides the child with the written words to assist memory– as a reminder, not to test reading (Woodcock & Muñoz-Sandoval, 2001).

reasoning tasks measure the two stages of concrete[5] and abstract reasoning[6] (Goswami, 1991). These two stages were coined by Piaget in his theory of cognitive development as previously discussed. Piaget's theory postulated that concrete and abstract reasoning emerge during the concrete and formal operations stages of development respectively (Papalia et al., 2004)[7]. From the aforementioned it seems reasonable to posit that concrete reasoning can be viewed as linked to BICS. BICS, as was previously mentioned in the section 2.2.3., relates to contextualised language skills and practices which provide the internal support for the acquisition and understanding of academic language. According to Papalia et al. (2004), concrete reasoning is regarded as the basis of all knowledge that the learner will acquire. They also posit that the acquisition and development of abstract reasoning enables learners to apply the knowledge they acquire in complex ways. Thus one could postulate that abstract reasoning can be linked to CALP.

Can it therefore be postulated that if there is a smooth transition from concrete to abstract reasoning in a supportive educational context, language development, in particular verbal reasoning, will follow the same course? If this is the case, then theoretically one could argue that a measure of verbal reasoning could provide a good measure of the threshold at an oral level that the learner requires to reach in the L2 before he or she can be educated in that language. According to Ushakova (Centeno-Cortés & Jiménez Jiménez, 2004), the L2 is incorporated into the classification system of the L1, and is reliant on the already established semantic system of the L1, and actively deploys L1 phonology.

[5] The ability to analyse information and solve problems on a literal level and includes skills such as basic knowledge of names of objects, places and people etc. (Papalia *et al.* 2004).

[6] The ability to analyse information and solve problems on a complex, thought-based level and includes skills such as formulating theories about nature of objects, ideas, processes, and problem-solving (Papalia *et al.* 2004).

[7] Concrete operations (7-11 years) and formal operations (11 years to adulthood) (Papalia *et al*. 2004).

Analogy has been proved to be a powerful means by which children acquire knowledge. It is, therefore, an essential developmental skill that mediates the progression of children's cognitive abilities. While previous and current research on analogical reasoning have revealed an immeasurably valuable understanding of children's utilisation of analogy, it may very well be of special interest if this research on verbal reasoning is implemented in more applied classroom settings in a South African context. Instruction by analogy has the potential to directly benefit children's education in this country, and can assist in developing children's reasoning and thinking not only in classroom but beyond it, thus developing the mind of the individual holistically in all spheres of learning.

2.5. Cross-linguistic Assessment

2.5.1. Introduction

Internationally there has been a trend of increasing acceptability coupled with positive perception, when it comes to using psychological tests and testing (Foxcroft, Paterson, le Roux & Herbst, 2004). According to Oakland (2004), there are an estimated 5000 standardised tests in English alone, developed predominantly in Western Europe and the USA. Test results provide a wealth of information in a short period of time, and can be used to form the basis for comparisons or evaluations of the test-takers (Paterson & Uys, 2005). On the one hand, these tests are important tools frequently used in the assessment decision-making process both nationally and internationally. On the other hand, though, these assessment tools can act as a disabling factor, if they are inappropriately applied or used in isolation without verifying the results against other measures (Paterson & Uys, 2005).

Thus, within the context of this study, standardised assessments must provide equivalent measurement across different language groups if comparative statements are to have substantive import. Without equivalent measurement, observed scores from the different language groups cannot be directly comparable. Equivalence of a measure is obtained when the relations between observed scores and latent constructs are identical across relevant groups (Van de Vijver & Tanzer, 2004). In other words, learners with the same standing on a latent construct, i.e. verbal reasoning, but sampled from two different language groups, namely an English first-language group and a Xhosa first-language group, should obtain the same expected score on a test of that construct. If the test was to favour the English first-language group, then an English first-language learner would be expected to obtain a higher score than a Xhosa first-language learner with equal verbal reasoning skills. The concept of equivalence will be further expounded upon in section 2.6.

2.5.2. Monolingual Assessment Across Language Groups

“Monolingual assessment” refers to the use of assessment tools which are available in only one language across two or more language groups (Koch, 2005). In following global trends, South Africa has become one of the countries that use monolingual tests to measure individuals on a particular trait. The dilemma in using these tests is threefold: (1) None of the tests have been either developed or adapted in a multicultural and multilingual context; (2) Some of the tests (e.g., the Bender and the Beery VMI) have been imported from overseas and full-scale national normative studies have never been carried out in an attempt to provide practitioners with appropriate norms, and (3) A number of the tests developed in South Africa are outdated as they were only developed for specific groups of South Africans e.g., SSAIS-R or JSAIS (Foxcroft et al., 2004). In addition, monolingual measures are oblivious to the fact that bilingual individuals may prefer using different words depending on the setting, interlocutor, and context (Iglesias, 2001) as well as their cultural experiences (Peña, 2001).

Allalouf and Abramzon (2008) assert that the use of a monolingual test across two language groups is problematic, because a single test form cannot assess proficiency where there is a large variation in the nature of language ability between the two groups. Furthermore, they are of the opinion that the use of different test forms, i.e. a special test form for each group rather than a single test form for both groups, has adverse implications for fairness and standardisation (Allalouf & Abramzon, 2008).

Bredell and colleagues (cited in Paterson & Uys, 2005) assert that culture is an important moderator in test performance because it affects test behaviour and consequently the construct being measured, and language is closely related to culture. Language can cause complications on three levels: (1) the language in which the test is constructed; (2) the difficulty level of the test language, in particular if the test is administered in the test-taker's second or third language (Van de Vijver and Rothmann, 2004), and (3) the language competence of the test-taker (Paterson & Uys, 2005). Additionally, Huysamen (2002) contends that consideration should be given to language deficiencies and cultural contexts when using monolingual assessment tools, as this may account for the poor standing of test-takers on the construct measured, and not owing to poor performance on their part. Huysamen (2002) refers to this as "construct-irrelevance", which occurs when a construct being measured may be relevant to one group and not to another. He further asserts that irrelevant variance may not be restricted to language proficiency only, but could extend to cultural differences that the test is not designed to measure (Huysamen, 2002). Thus, it is important to determine whether the performance on the test reflects the test-taker's ability, and not the level of competence in the test language (Foxcroft, 2004). In other words, one has to ensure that the same construct is measured across groups of different languages.

Pearson, Fernández and Oller, (1993) proposed using conceptual scoring as a more meaningful measure of the bilingual child's conceptual knowledge. The system entails computing all the concepts demonstrated by the test-taker, either through constructed or selected responses in both languages, and correcting for concepts shared in the two languages. This approach results in a more valid or accurate representation of a bilingual child's knowledge of concepts. In other words, a response on a test is scored regardless of the language in which it is produced. Thus, when describing a ball, if a learner said, "It's red and blue *ibenomgca nenkwenkwezi* (it's red and blue and has a stripe and a star), the learner would achieve a monolingual score of two in English, but a conceptual score of four because he or she expressed unique concepts in each language (Bedore, Peña, García & Cortez, 2005) namely English and Xhosa. As a result, when testing concepts, a bilingual child's conceptual system as a whole should be considered, rather than as two language-specific systems.

2.5.3. Research on Monolingual Assessment

There is new awareness about the limitations of monolingual instruments and their use in multilingual and multicultural contexts. This has sparked renewed interest globally by numerous researchers in identifying various issues surrounding the use of monolingual assessments. However, in South Africa, research in this domain is in its early stages and thus further research in this field is sorely required.

In a study on the Hebrew Proficiency Test (HPT) Allalouf and Abramzon, (2008) studied participants who were Arabic and Russian first-language speakers. This study was unique in that it examined differences in performance on L2 test items between groups from different L1 backgrounds. The results of this study indicated that the Arabic speakers performed better than the Russian speakers. Proficiency differences between the two groups were small, and so the accuracy of DIF detection was increased. The results revealed that vocabulary and grammar items usually favoured the Arabic speakers because of the similarity between Arabic and Hebrew and because of the presence of cognates in the test. Thus in light of the study's results, it was concluded that the HPT functioned differently across the two groups.

Research conducted by Rossier (2004) reported on the cross-cultural equivalence of a number of personality inventories in frequent use. He investigated personality traits in Burkina Faso, a sub-Saharan African country, and in Switzerland. His results indicated that the structural equivalence of tests is affected by the theoretical differences on which the tests are based. In particular, Rossier (2004) postulated that when tests are based on theories that are sensitive to cultural context and environmental influences, structural equivalence is less likely to be observed. Rossier concluded that tests that are more dependent on cultural contexts are less stable across cultures. According to Paterson and Uys (2005), cultural context in particular becomes a problem when performing personality assessments, as constructs have different meanings and are experienced differently across cultures. Van de Vijver and Rothmann (2004) contend that, in general, personality tests require high levels of language proficiency.

A South African study conducted by Abrahams and Mauer (1999) investigated the impact of home language on response to items of the Sixteen Personality Factor Questionnaire (16PF). They found that anomalies existed as far as the comparability of items across groups were concerned. The factor structures of the African languages groups and the English groups differed, and as a result Abrahams and Mauer (1999) queried the cross-cultural use of this measure.

In a recent South African study, Koch (2007) used a reading comprehension test to evaluate equivalence across three language groups, namely, English, Afrikaans and African-language speakers. The analysis of data revealed differential item functioning (DIF) of item difficulties across groups, and construct bias across the groups was also found. "Construct bias" refers to the question of whether the same underlying construct is being measured in each language group (the concept of DIF will be discussed in greater detail under the later heading in section 2.6.). The test displayed unacceptable levels of item-bias, measuring different constructs across English L1 and L2 groups. What was concluded from the study was that the scores of the English L1 and L2 students on the reading comprehension test could not be used to make equivalent statements regarding the construct measured across the groups, and the scores of the different groups could not be placed on a common scale for comparison.

Further research was conducted by Koch and Dornbrack (2008) evaluating bias in the South African context, particularly with regard to monolingual assessment. Their study evaluated the utilisation of language criteria for admission to higher education in the SA context (Koch & Dornbrack, 2008). Despite the fact that higher education institutions have adopted a multilingual language policy which includes only one or two African languages as additional languages of teaching and learning, the predominant languages of learning and teaching have remained English and Afrikaans (Koch & Dornbrack, 2008). As a result Xhosa first-language students from disadvantaged educational backgrounds would suffer major repercussions. According to Koch and Dornbrack (2008), these criteria for admission will prejudice the African-language-speaking students, as the educational backgrounds of students applying to these institutions are not being taken into consideration. Furthermore, the study revealed that evaluating students' performance in a single language as representative of their academic literacy in the language of teaching and learning can be viewed as biased and problematic (Koch & Dornbrack, 2008).

However, according to Van de Vijver and Tanzer (2004), a test that is biased in one context may not be biased in another, and as a result, tests need be evaluated in the context of their usage. Research regarding bias and equivalence of assessment tools in South Africa is still in its infancy stage. Van de Vijver and Rothmann (2004) contend that much more research is required on bias and equivalence of assessment tools used in a South African context before tests and testing can live up to the demands implied in the Equity Act. This Act will be discussed in the next section, namely Theoretical Framework of Bias and Equivalence. According to Watson (2004), this Act demonstrates a zero-tolerance approach towards inferior and culturally inappropriate tests. The next section considers the methodological framework of bias and equivalence for multicultural and multilingual assessment.

2.6.Theoretical Framework of Bias and Equivalence

2.6.1. Introduction

In South Africa, the Equity Act 55 of 1998 stipulates the prohibition of psychological testing and other assessment measures, unless scientific validity and reliability, fairness, and non-bias against participants or groups can be validated (Van de Vijver & Rothman, 2004). As a result, the concept of bias and the attainment of equivalence are of fundamental significance in cross-linguistic research and measurement, in multilingual and multi-cultural contexts (Van de Vijver & Leung, 1997) as is the case in South Africa. These concepts are associated with the validity of a measure and are intrinsic in the characteristics of an instrument in cross-linguistic comparison (Van de Vijver, 1998).

Even though Van de Vijver (1998) considers bias and equivalence as interrelated, they provide different perspectives of the same question, namely the extent to which scores have the same meaning across groups (in the case of tests that are available only in one language i.e. monolingual tests), or different language versions of a test (i.e. multilingual tests). Bias and equivalence form the theoretical framework of the current study in order to cross-validate results found by Haupt (2009) and Arendse (2009) on the Verbal Analogies (VA) scale of the WMLS. More specifically this framework provides the foundation for ascertaining whether the scores on the adapted English items could be compared across two groups, namely, the English first-language group and the Xhosa first-language group. For test scores to be comparable (i.e. establish equivalence) they have to demonstrate that the test is not biased.

The concepts of the taxonomy of equivalence and the three categories of bias will now be explained in order to gain a more acute understanding of the theoretical link between test equivalence and test bias.

2.6.2. The Taxonomy of Equivalence

According to Hambleton and Kanjee (1995), for any comparison between different language groups to be valid, the test used must demonstrate equivalence. “Equivalence” refers to the implications of bias on cross-linguistic score comparisons to be made (Van de Vijver, 1998; Van de Vijver & Lueng, 1997; Van de Vijver & Rothman, 2004). This indicates the measurement level at which scores of tests that are available in more than one, or only one language but are administered to participants of different languages, can be comparable (Van de Vijver, 1998). Various theorists distinguish between various levels of equivalence and regard these levels as forming a hierarchical pyramid. These levels of equivalence are, according to Van de Vijver (1998), divided into four ranking categories namely: construct inequivalence, construct equivalence, measurement unit equivalence and at the top of the hierarchy scalar equivalence. Van de Vijver and Tanzer (2004) view construct inequivalence as the polar opposite of construct equivalence, and thus categorize equivalence into only three ranking categories. For the purpose of this thesis, Van de Vijver's (1998) four ranking categories will be used.

At the lowest level of the hierarchical pyramid is construct inequivalence, which refers to the incomparability of constructs across language groups and is tantamount to "comparing apples and oranges" (Van de Vijver, 1998). Construct equivalence or structural equivalence falls at the next level of the equivalence hierarchy and occurs when the instrument measures the same construct across different language groups (Van de Vijver, 1998; Van de Vijver & Rothman, 2004). In other words there is a link between scores obtained in one language group and scores obtained in another group. Construct equivalence postulates that the same construct be measured across all researched groups, and that it be measured with equal reliability in all groups (Sireci, Bastari, Xing, Allalouf, & Fitzgerald, 1998). A comparison of nomological networks across language groups is one avenue of addressing construct validity in each language group (e.g. convergent/discriminant validity studies) (Van de Vijver, 1998). The aforementioned studies investigate whether a measure shows a pattern of high correlations with the related measures (convergent validity) and low correlations with measures of other constructs (discriminant validity) (Van de Vijver & Tanzer, 2004) as would be expected from the scale measuring VA.

Construct equivalence is said to be reached when the dimensional structure of a test is found to be consistent across the different language versions or language groups (Sireci et al., 1998). This means that both the conceptual definition and structure of the construct can be generalised across all groups of interest, and that scores are comparable across groups (Tanzer & Sim, 1999). A lack of construct comparability and item bias can lead to test-wide bias or scalar inequivalence, which implies that inferences from test scores are not equivalent across groups (Sireci et al., 1998). Item bias will be discussed in the next section of "categories of bias". Construct equivalence is a shared feature of the construct and the groups of interest, and does not depend on the test used to measure the construct (Tanzer & Sim, 1999). In other words, construct equivalence is inherent in the construct and the groups assessed, rather than in the particular instrument used to measure the construct under investigation.

At the third level of the hierarchy is measurement unit equivalence, which occurs when instruments have the same units of measurement across language groups but the origin differs, such as the Kelvin and Celsius scales in temperature measurement (Van de Vijver & Rothman, 2004). In other words, the units of measurement are identical but there is a constant difference of the measure. Thus, no direct score comparisons can be made across the different groups unless the size of the difference in scale origin is known (Meiring, Van de Vijver, Rothmann & Barrick, 2005). This would result in the origins of the scales across the groups being affected, but the measurement unit would remain the same (Van de Vijver, 1998).

Scalar equivalence is the highest level of equivalence and assumes that identical interval or ratio scales apply to measures in the language groups compared, and when the scale has the same origin (Van de Vijver, 1998; Van de Vijver & Rothman, 2004). Thus, scalar equivalence can only be demonstrated when the same construct is measured, with no item and measurement bias. This is the only type of equivalence that allows the researcher to make valid conclusions when averages are compared across language groups, for example, by using t-tests and analysis of variance (Van de Vijver, 1998; Van de Vijver & Rothman, 2004).

Equivalence is always challenged when bias, at any level, occurs, and thus to maintain the utmost level of equivalence, the adapted measure and its subsequent application must be as free from bias as possible (Van de Vijver, 1998). The subsequent sub-section considers the concept of bias with particular focus on its various categories.

2.6.3. Categories of Bias

"Bias" is a generic term which refers to the presence of all nuisance factors (superfluous but systematic sources of variations) in cross-linguistic score comparisons – thus with tests used across language groups, whether they are monolingual or multilingual (Van de Vijver, 1998). It alludes to unintended sources of variation that represent alternative explanations of intergroup differences (Van de Vijver, 1998). Thus if bias is present, cross-linguistic score differences are not engendered by the target construct (such as intelligence) but by some other characteristic (such as social desirability). According to Van de Vijver and Tanzer (2004), when score differences in the indicators of the target construct do not correspond with differences in the underlying trait or ability, then bias has occurred. Bias can occur for many reasons, some of which include poor item translation, inappropriate item content, and lack of standardisation in administration procedures. Since equivalence is evaluated by assessing bias and bias has to do with the characteristics of an instrument in a specific cross-linguistic comparison which is indicative of this study, rather than with its intrinsic properties (Van de Vijver & Tanzer, 2004), the following section will explore the three sources of bias in cross-linguistic testing, namely construct, method, and item bias (Van de Vijver & Rothman, 2004).

Construct bias occurs when the construct measured, in this case verbal reasoning, is not identical across groups (Van de Vijver & Rothman, 2004). This can stem from a lack of, or overlap in, behaviours associated with the construct in the groups studied (Van de Vijver, 1998). For example, research into Western and non-Western countries has revealed that everyday concepts of intelligence in non-Western countries are more all-encompassing than the domain covered by most Western intelligence tests (Van de Vijver and Leung, 1997). This indicates that certain concepts or constructs are context-specific and cannot be used all over across cultures or language groups.

Ho (1996, cited in Van de Vijver, 1998) provides a fitting example. He studied the concept of filial piety, in other words being "a good son or daughter" in Chinese society. What he discovered was that this concept in a Chinese context as opposed to the individualistic Western society, encompassed a much broader array of behaviours, such as taking care of one's parents, conforming to their requests, and treating them well. Thus, if this construct were to be measured, a Western-based measure would insufficiently cover the Chinese concept, while a Chinese questionnaire would be over-inclusive according to Western standards. A means of offsetting the aforementioned shortcoming is to clearly define the behaviours included in the measure (Van de Vijver, 1998). However, it is important to note that even though a construct is clearly defined in a measure, there is no 100% guarantee that the scores will not display bias.

Method bias consists of sample bias, administration bias, and instrument bias, and refers to all sources of bias emanating from a methodological-procedural aspect which includes factors such as sample incomparability, instrument differences, tester and interviewer effects, and the mode of administration (Van de Vijver, 1998; Van de Vijver & Rothman, 2004). In other words, intergroup differences in social desirability, response sets such as acquiescence and extremity ratings, familiarity with stimuli, and response formats (e.g. multiple choice or Likert scales) etc. can all constitute method bias. Even communication problems such as poor mastery of the testing language by one of the parties involved, interviewer characteristics such as gender or cultural preference, can trigger method bias. This will lead to differences in scores between groups that are not attributed to any intrinsic differences of the groups on the construct researched. According to Van de Vijver and Leung (1997), method bias in general affects scores at the level of the whole instrument.

The last category of bias is item bias, which refers to anomalies of an instrument at an item level (Van de Vijver, 1998) such as poor wording, inappropriateness of item content in a cultural group, and inaccurate translations (Van de Vijver & Hambleton, 1996). According to Hambleton (1994), an example of inaccurate translation comes from a Swedish-English comparison of educational achievement. The item read: "Where is a bird with webbed feet most likely to live? (a) in the mountains; (b) in the woods; (c) in the sea; (d) in the desert. The Swedish translation of "webbed feet" was "swimming feet", thus providing a clear cue about the correct answer.

The term "item bias" is synonymously used with the term "differential item functioning" (DIF). DIF is a statistical analysis procedure that has been used widely for comparison of adapted measures between language groups (Gierl & Khaliq, 2001). DIF analysis is a procedure to identify items that function differently across two different groups, and is based on the underlying assumption that examinees with similar ability should perform similarly (Sireci & Allalouf, 2003). DIF might affect test performance in favour of one or another particular group, which occurs when an item is significantly more difficult for one group than for another when the group's ability is taken into consideration (Allalouf, Hambleton & Sireci, 1999; Sireci & Allalouf, 2003). In other words, DIF occurs when an item functions differently across groups of participants of equal ability but from different groups, for example different L1 groups, do not have equal probability of responding correctly to that item (Allalouf & Abramzon, 2008; Allalouf, Hambleton & Sireci, 1999).

In general, bias will always lower the level of equivalence and increase with the cultural or linguistic distance to be bridged by the measure. This, according to Van de Vijver (1998), is more likely to occur when a measure displays more cultural saturation. In other words, the more culture-specific a measure is, the more likely it is to display lower levels of equivalence. A final point with regard to item bias is the distinction made between uniform and non-uniform bias. An item is regarded as uniformly biased if the main effect of culture or language is significant (Van de Vijver, 1998). In other words, for each observed total score level, the item is consistently easier or more endorsed in one group than in another. An item displays non-uniform bias if the interaction of score level and language or culture is significant (Van de Vijver, 1998). Thus, the cross-linguistic score difference varies with the observed total test score.

According to Van de Vijver and Leung (1997), uniform and non-uniform bias may be harmless for construct equivalence since numeric score comparisons across language groups are not permitted. In addition, uniform bias will not threaten measurement unit equivalence, since unbiased scores at this level of equivalence cannot be directly compared across each group. In other words scores in centimetres cannot be directly compared to scores in metres even though they are both units of measurement. Both these scores need to be converted into either centimetres or metres to be directly comparable. Furthermore, adding a constant to all scores in either a single group or one group and not the other, does not affect equivalence at this level (Van de Vijver & Leung, 1997).

In contrast, if non-uniform DIF occurred in the two groups, it would drastically eliminate equivalence as the measurement units would no longer be the same because one group would be favoured over the other group (Van de Vijver, 1998). As a result, when several items display the category of bias, cross-linguistic score comparisons are likely to produce inaccurate findings (Van de Vijver & Leung, 1997). In addition, the introduction of uniform bias to scores that display scalar equivalence will lead to a loss of scalar equivalence (Van de Vijver & Leung, 1997). Van de Vijver and Leung (1997) regard DIF as "dangerous" for equivalence and results in compromising scalar equivalence and thus the comparability of test scores across groups.

Numerous techniques have been developed to identify item bias. They include the delta plot, analysis of variance method, the Mantel-Haenszel method, and the logistic regression approach (Kamata & Vaughn, 2004). The Mantel-Haenszel procedure is the most popular technique used to detect bias in dichotomously scored items (Van de Vijver & Tanzer, 2004) and is the method of choice for the present study.

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

CHAPTER THREE

METHODOLOGY

3.1. Introduction

This study falls under the umbrella of a larger study consisting of numerous phases concerning the adaptation of the WMLS: (1) adaptation of the original WMLS (English version), into South African English and Xhosa; (2) evaluation of the equivalence of the two language versions of the WMLS across English first-language and Xhosa first-language groups; (3) evaluation of the predictive, construct and content validity of both these adapted versions across the two groups in the South African context.

This sub-study was located within a broader study. More specifically, it falls within the second phase of evaluation of the equivalence of the WMLS across English first-language and Xhosa first-language groups, and draws upon its existing dataset. Participants and data were thus not selected specifically for the aims of this study, but rather for the aims of the broader study. Accordingly, the research design, reported sampling procedures, and sample characteristics are those of the broader study. This sub-study utilises secondary data analysis (SDA), which can be described as the analysis of data that has been collected previously by another researcher. It concentrates primarily on the equivalence of the adapted English version of the WMLS, specifically focusing on the VA scale, across the two language groups, as stated in the aims.

The SDA addresses methodological limitations in the previous research (Haupt, 2009) on the scale, by using a different method for differential item functioning (DIF) to cross-validate findings in that research, and by extending the previous research by assessing the effect of

DIF on construct equivalence. The ultimate aim is thus to evaluate the scalar equivalence of the adapted English version of the VA scale across English and Xhosa first-language groups. These analyses will be explained in more depth in this section.

The study thus falls within the ambit of a quantitative research methodology that is informed by psychometric test theory, in particular the theory and methodology dealing with bias and equivalence. A distinct characteristic and strong point of quantitative research is that it is very structured, and attempts to control for various forms of error through different intricate measures of control (Babbie & Mouton, 2001).

3.2. Design of the Study

In the current study the researcher evaluated the scalar equivalence of the VA scale across two language groups. The study utilised comparative and correlational statistical techniques to conduct comparisons between the two language groups on the English version of the VA scale. A differential research design (Gravetter & Forzano, 2008) was used since the researcher did not actively manipulate the assignment of participants to groups; instead, the participants were automatically assigned to groups based on pre-existing characteristics, namely their first language as well as their grade level at school. Furthermore, the method of sampling used, as will be delineated in the next section, allowed the researcher to control for confounding variables.

3.3. Sampling

The sampling procedure used in the main study consisted of convenience purposive sampling. The aim of this type of sampling was to select a sample on the basis of the researcher's knowledge of the population, its elements, and the nature of the objectives

(Babbie & Mouton, 2001; Welman, Kruger & Mitchell, 2008). In other words, the learners in the two language groups were purposefully selected in an attempt to maintain as far as possible an equal number of male and female Grade 6 and 7 learners in both rural and urban areas across groups. This type of sampling allowed the researcher to control for confounding variables such as gender, grade and educational background.

3.4. Participants

Since the researcher was using SDA, the participants of the larger study were retained for the present study. The participants consisted of 198 English first-language learners and 197 Xhosa first-language learners, who were tested on the English version of the WMLS during the second half of 2006 and the second half of 2007.

The English and Xhosa first-language speakers were selected from ”ex-model C” and “previously disadvantaged” schools in the Port Elizabeth and Grahamstown regions. The learners were selected from these specific schools with the aim of maintaining the validity of the learners’ different educational levels of English as well as the differing levels of their teaching through the medium of English.

Tables 1 to 5 below represent the distribution of the sample in terms of the two language groups, the English first-language-speaking learners and Xhosa first-language-speaking learners, gender and grade.

Table 1:

Distribution of participants per language group

Language Group	**Sample Size (n)**	**Percentage**
English	192	49.90
Xhosa	193	50.10
Total	**385**	**100**

The above table indicates the number of participants from each language group. There is only one more participant in the Xhosa first-language-speaking group.

Table 2:

Distribution of participants per gender

Gender	**Sample Size (n)**	**Percentage**
Male	174	45.20
Female	211	54.80
Total	**385**	**100**

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

Table 3:

Distribution of participants per grade

Grade	**Sample Size (n)**	**Percentage**
Grade 6	177	46.00
Grade 7	208	54.00
Total	**385**	**100**

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

Table 4:

Distribution of language group per gender

Language	**Gender**					
Group	**Male**	**%**	**Female**	**%**	**Total**	**%**
English	98	49.75	99	50.25	192	100
Xhosa	76	39.58	112	60.42	193	100
Total	174		211		385	

Table 4 provides a view of the each language group disaggregated by gender. The above table indicates that the sample consisted of more males in the English first-language group and more females in the Xhosa first-language group.

Table 5:

Distribution of language group per grade

Language	**Grade**					
Group	**Grade 6**	**%**	**Grade 7**	**%**	**Total**	**%**
English	82	42.70	110	57.30	192	100
Xhosa	95	49.22	98	50.78	193	100
Total	177		208		385	

Table 5 indicates the language groups disaggregated by grade (Grades 6 and 7). The English first-language group consisted of more Grade 7 learners in the sample, while the Xhosa first-language group had more Grade 6 learners.

It is evident from the above tables that despite the use of convenience purposive sampling in order to control for various confounding variables as discussed above, gender and grade numbers of learners could not be equally maintained across the two language groups.

A t-test (Table 6 below) demonstrated that the mean scores of 14.11 and 8.97 for the Grade 7 learners and 13.06 and 8.73 for the Grade 6 learners on the verbal analogies scale across the English and Xhosa first-language groups respectively, indicated an overall performance that favoured the English first-language group. However, across both grades and across both groups, the standard deviations were relatively small and clustered around the mean. However, this does not indicate whether the differences between the English and Xhosa first-language groups were significant.

Table 6:

Mean score and standard deviations across the two language groups on the VA scale

Language Group	Grade	Mean	Standard Deviation
English	6	13.06	4.42
	7	14.11	4.62
Xhosa	6	8.73	4.33
	7	8.97	4.59

In Haupt’s study (2009), a Hotellings’ T^2-test was conducted to identify whether the differences between the two language groups were significant. The results of the Hotellings’ T^2-test indicated that there were significant overall differences between the English and Xhosa first-language groups on the adapted English version of the VA scale (Table 7 below).

Table 7:

Hotellings'T^2-test results for the English

and Xhosa first-language groups

Subscale	**Mean differences**	**Df**	**p**
Verbal Analogies	2168.40	384	0.00

The discrepancies of an unequal number of males and females in each group, as well as an unequal numbers of learners in the two grades, could very well impact the results of the findings of this study. This is especially pertinent with regard to the two grades, since it is assumed that Grade 7 learners would be better equipped academically, and thus would fare better on the test that their fellow learners in Grade 6. However, as the DIF analysis will be conditioned on ability as measured by the total score on the VA scale (as will be explained in the data analysis section) and factor analysis results are not influenced by differences in ability (Sireci & Khaliq, 2002), the differences in scores across the two language groups are not expected to impact negatively on the results of this study. The sampling incongruencies therefore are not regarded as seriously affecting the internal validity of the study.

3.5. Data Collection Instrument

This study focused specifically on the adapted English version of the WMLS, and even more specifically on the VA scale of the WMLS. Thus, this section will: (1) provide a brief overview to the WMLS and the various subscales; (2) discuss the psychometric properties of the WMLS in the American context with a specific focus on the VA scale, and (3) conclude with discussing the psychometric properties of the VA scale in the South African context.

3.4.1. Woodcock Muñoz Language Survey (WMLS)

The WMLS is a test used to measure academic language proficiency of learners, and has been extensively used in the USA to evaluate Additive Bilingual Education. The original WMLS is available in English and Spanish (Woodcock & Munoz-Sandoval, 2001). It consists of sets of individually administered scales designed to measure a broad sampling of proficiency in four critical areas of oral language, listening, reading, and writing.

The four subscales are: Picture vocabulary, Verbal Analogies (forming the oral language cluster), Letter -Word Identification, and Dictation (forming the reading-writing cluster). The test requirements, as well as what each test measures, are given in Table 8 below. The content was selected to represent important skills needed for language proficiency for a diverse population, covering a broad range of development from ages 3 to adulthood (Woodcock & Muñoz-Sandoval, 2001). These scales are primarily measures of language skills predictive of success in situations characterised by CALP requirements. In other words, the instruments provide an overall measure of language competence as well as CALP (see a discussion of CALP in Chapter Two).

Table 8:

Test Requirements and Test Measurement of the WMLS

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

This study will utilise one of the scales of the adapted English version of the WMLS, namely the VA Scale. This 35-item scale is used to measure listening and speaking skills, either individually or collectively, and purports to assess an individual's ability to complete oral analogies, which necessitates verbal comprehension and verbal reasoning, such as "A bird flies; a fish swims". The vocabulary remains simple throughout, but the relationships become increasingly complex. The items of the WMLS are not made available in an appendix of this thesis, as it is a commercially purchased test, and items are therefore confidential as well as copyright.

3.4.2. Psychometric Properties of the WMLS

When selecting an instrument for utilisation, two key issues need to be taken into consideration, namely the reliability and validity of the instrument. The WMLS was standardised on populations in the USA, central America, South America and Spain. The reliability of the WMLS, internal consistency reliability coefficients (r*n*) and standard errors of measurement (SEMs) were calculated for all English forms and clusters across their range of intended use (Woodcock & Muñoz-Sandoval, 2001). The reliability of the WMLS was calculated using split-half reliability as well as odd and even raw scores (Woodcock & Muñoz-Sandoval, 2001). The corrected reliability coefficient was calculated by means of the Spearman-Brown formula. The median reliabilities were found to range from 0.80 to 0.93 for the scales and 0.88 to 0.96 for the clusters. The median reliabilities for the VA scale were found to be 0.81.

Furthermore, the validity of the WMLS was evaluated on content, concurrent, and construct validity (Woodcock and Muñoz-Sandoval, 2001). "Content validity" refers to the extent to which the content of a test represents the domain of content that it is designed to measure. "Concurrent validity" refers to the extent to which scores on a test are related to scores on a certain criterion measure, which is typically expressed as a correlation coefficient between the test and the criterion (Woodcock & Muñoz-Sandoval, 2001). The more similar the test is to the criterion measure, the higher the validity coefficient will be, and vice versa.

The above-mentioned reliability and validity are based on the original American version of the WMLS. The current study forms part of a larger study investigating the psychometric properties of the adapted South African versions of the WMLS, and thus will add to the psychometric information currently being collected for the South African population.

3.4.3. Psychometric Properties of the Verbal Analogies Scale

As yet, the WMLS has not been normed for the South African population. Therefore a complete psychometric properties dossier of the test for the South African context is not yet available, even though research is currently in progress (Koch, 2009). The research in progress indicates that both the adapted English version and the Xhosa version of the WMLS demonstrate promising results on two of the scales of the test, namely the VA and the Letter-Word Identification (LWI) (Arendse, 2009; Haupt, 2009). According to results on the adapted English version of the VA Scale, in particular, good internal consistency was displayed across the English first-language and Xhosa first-language groups, with a Cronbach's Alpha of 0.83 and 0.86 respectively (Haupt, 2009).

Furthermore, a logistic regression differential item functioning (DIF) analysis across English and Xhosa first-language groups on the English version of the scale indicated that only six items (1, 5, 8, 9, 14 and 18) displayed DIF on this scale, two items having large DIF, two items having moderate DIF and two items displaying negligible DIF (Haupt, 2009). The study only rejected the null hypothesis of "no DIF" for the items displaying moderate and large DIF. In addition, the findings revealed that items 8 and 9 were uniform DIF items that favoured the English first-language group, while items 5 and 18 were non-uniform DIF and favoured the Xhosa first-language group.

In Haupt's study (2009) on the same data that was utilised for the current study, the mean scores of 13.66 and 8.94 on the VA scale across the English and Xhosa first-language groups respectively, indicated that the overall performance of the Xhosa first-language group was lower than that of the English first-language group. The standard deviations of 4.56 and 4.47 across the English and Xhosa first-language groups respectively were relatively small, indicating that scores clustered around the mean. The mean item correlations of 0.32 (English first-language group) and 0.34 (Xhosa first-language group) were similar across both language groups, while the item difficulty for the two language groups on the VA scale were 0.39 (English first-language group) and 0.25 (Xhosa first-language group). The standard deviation of the mean item difficulty values for the English first-language group is 0.31 and 0.26 for the Xhosa first-language group. From the aforementioned values it is evident that both groups displayed a satisfactory mean item discrimination level on the items, but the VA scale was easier for the English first-language group than for the Xhosa first-language group, even though both language groups' standard deviation clustered around the mean.

In Arendse's study (2009) on the equivalence of the two language versions of the test, a factor analysis revealed two factors on the English version of the VA scale, corroborating the findings in Koch (2009), where a weighted multidimensional scaling analysis also displayed two dimensions on this scale. The first factor displayed structural equivalence across the two language versions, while the second factor was found to be inequivalent (Arendse, 2009).

The promising results displayed by the adapted English version of the VA scale (Haupt, 2009; Koch, 2009) necessitates increasing focus on this scale in order to cross-validate the previous findings (using a different DIF technique than was used in the previous research). Furthermore the current study aims at refining previous research to include construct equivalence in order to gain absolute certainty in the scalar equivalence of this measure, so as to use it in the South African context.

3.6. Procedure

The researchers of the larger study received ethical clearance from the Nelson Mandela Metropolitan University (NMMU), then known as the University of Port Elizabeth's (UPE), ethics committee in 2006, as well as permission from the Eastern Cape Department of Education (Appendix A and B). Subsequently, contact was made with the principals of the various schools selected for the study. Information sheets as well as informed consent forms (Appendix C and D) in both English and Xhosa were given to the principals of these schools to forward to the parents of the learners. Only those children whose parents completed these forms were allowed to participate in the study.

The data was collected by trained UPE Psychology Honours students, and captured in Excel spreadsheets and combined by the main researcher. The current researcher received permission from the main researcher to use and re-analyse the data collected from the main study.

3.7. Data Analysis

3.7.1. Introduction

Owing to the use of Secondary Data, the researcher will use the existing data of the main study to conduct various statistical tests using the statistical programmes of SPSS (Statistical Package for the Social Sciences) and CEFA (Comprehensive Exploratory Factor Analysis) Version 3.04 (Browne, Cudeck, Tateneni & Mels, 2004). SDA allows the researcher to repeat analyses in order to address methodological issues or to augment previous data with current findings (Babbie & Mouton, 2001). Table 9 below represents an overview of the analysis conducted in this study.

Table: 9

Overview of the steps and techniques utilised in the analysis process

Step	Technique	Procedure
1	Exploratory factor analysis	**a.** Run a factor analysis for the English group on the VA scale first to find a stable factor structure using an oblique rotation method. **b.** Impose the same factor structure of the English group on the Xhosa group. **c.** Estimate factorial congruence utilizing the Tucker's phi and a scatterplot of the factor loadings per group.
2	Mantel-Haenszel DIF procedure	**d.** Run a Mantel-Haenszel DIF procedure to identify the DIF items.
3	Exploratory factor analysis	**e.** Run a subsequent factor analysis on both groups without DIF items using the same procedure as in step 1.

A detailed description of the analysis conducted to test the two specific research aims in order to achieve the overall aim of this study, follows below.

3.7.2. Differential Item Functioning (DIF)

Research Aim 1: To evaluate the differential item functioning (DIF) of the English version of the VA scale across English and Xhosa first-language groups.

Null Hypothesis: The probability of scoring 1 on item *i* on the subscale will be the function of ability alone, in other words there will be no items functioning differentially across the English and Xhosa first-language groups on the adapted English version of the VA scale.

Statistical Test: The statistical analysis that was employed to detect DIF was the Mantel-Haenszel DIF detection method using the Mantel-Haenszel chi-square statistic. The Mantel-

Haenszel DIF technique is a commonly used procedure to detect bias in dichotomously scored data (Sireci & Allalouf, 2003).

An item is regarded as exhibiting DIF when individuals from the focal and reference groups differ in the probability of answering the item correctly, after controlling for ability (Kamata & Vaughn, 2004). The “reference group” is the group to which performance on the item is being compared, while the “focal group” is the group in which an item is suspected to function differentially (Kamata & Vaughn, 2004). The Mantel-Haenszel DIF procedure matches individuals on ability (usually total test score) to determine whether comparable individuals from different populations perform the same on particular items. In the current study one would expect the English and Xhosa first-language groups who have the same total test score to perform in an equivalent manner on each VA item.

The MH chi-square statistic is calculated as:

$$\chi^2_{MH} = \frac{\left(\left|\sum_j A_j - \sum_j E(A_j)\right| - 0.5\right)^2}{\sum_j var(A_j)} \qquad (1)$$

where the variance of A$_j$ (var A$_j$) equals:

$$var(A_j) = \frac{n_{Rj} n_{Fj} m_{1j} m_{0j}}{T_j^2 (T_j - 1)} \qquad (2)$$

Field Code Changed

The expected value of A_j ($E(A_j$) is calculated from the margins, as in a typical chi-square analysis (Sireci & Allalouf, 2003). The items with significant Mantel-Haenszel chi-squared statistics are identified as biased, and thus the null hypothesis on these items should be rejected. The MH chi-square was computed using the statistical software SPSS. The significance of the Chi-square was assessed using a very stringent criterion p value of 0.0001 ($\rho < 0.0001$). Items that met this criterion were flagged as displaying DIF. Furthermore, a “constant odds ratio” was used to provide an estimate on the magnitude of the DIF (Sireci & Allalouf, 2003).

This ratio (α_{MH}) is computed as:

$$\text{[illegible]} = \frac{\sum_{\text{[illegible]}} \frac{\text{[illegible]}}{\text{[illegible]}}}{\frac{\text{[illegible]}}{\text{[illegible]}}} \qquad (3)$$

Field Code Changed

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

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

$$MH\_D - DIF = - 2.35 \ln [\alpha_{MH}] \qquad (4)$$

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

3.7.3. Evaluation of construct equivalence

Research Aim 2: To assess the construct equivalence of the English version of the VA scale across English and Xhosa first-language groups with the DIF items removed

Statistical Tests and Steps Utilised in the Analysis: The method used for analysing the construct equivalence on the two groups was the statistical technique of exploratory factor analysis (EFA) of dichotomous items at an item-level, using tetrachoric correlations to extract the factors (Kubinger, 2003). The Tucker's phi coefficient was used to assess the congruence of the construct(s) across the two language groups. The motivation behind using EFA, is to identify a latent subset of characteristics or factors, that underlie a specific domain (Schaap & Vermeulen, 2008). This is the most frequently employed technique to ascertain construct equivalence.

The Tucker's phi coefficient is commonly used to evaluate the similarity of factors across different groups (Zumbo, Sireci & Hambleton, 2003). In other words, the Tucker's phi makes known how similar the pattern of high and low factor loadings (i.e. factor loading patterns) are, across different groups (Zumbo et al. 2003). The Tucker's phi formula can be presented as follows:

$$\phi_{xy} = \frac{\sum X_i \ Y_i}{\sqrt{\sum X_i^2 \sum Y_i^2}} \qquad (5)$$

(Van de Vijver & Leung, 1997).

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

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

3.7.4. Factor Analysis

Factor analysis is essentially a multivariate, linear reduction, statistical technique that is used to explore the observed and empirical relationships between variables. This process permits the minimising of variables that the researcher has to contend with, while at the same time increasing the conceptual understanding of the domains measured by the instrument (Hair, Black, Babin & Anderson, 2010). Thus, according to Thompson (2004), factor analysis provides a holistic means of extrapolating a parsimonious set of underlying dimensions from an unfathomable mass of variables.

According to Hair et al. (2010), factor analysis is in essence a procedure used for reducing the complexity of the data by attempting to identify an underlying set of relationships between variables. Factor analysis has only gained popularity since the advent of computer-based computation. This was due to the size and complexity of the calculations that needed to be undertaken. As a result, two broad approaches to data reduction came about using factor analytic techniques: exploratory factor analysis and confirmatory factor analysis (CFA). The exploratory approach is the more popular approach and is employed when the data under investigation is to be analysed from a theoretical perspective, and the various factors to be extrapolated are identified and labelled *post facto* (Campbell, Walker & Farell, 2003). Thus, in exploratory factor analysis, the researcher has little or no knowledge about the factor structure. While CFA on the other hand, assumes that the factor structure is known or hypothesised *a priori*.

Zumbo et al. (2003) postulate that the investigation of construct equivalence is typically explored by the utilisation of a pair-wise comparison of factors, in other words latent variables or dimensions, across groups. They further contend that even though various methods such as cluster analysis and multidimensional scaling (MDS) have been utilised in exploring construct equivalence, CFA has become the standard and commonly recommended approach (Zumbo et al., 2003). However according to Van de Vijver and Leung (1997), EFA or CFA can be used in order to examine construct equivalence. Cartell (cited in Thompson, 2004) is of the opinion that factor analysis is the reigning queen of correlational techniques and is the most advanced logical development.

3.7.4.1 Exploratory Factor Analysis

EFA is by nature and design exploratory. There are no inferential statistics (i.e. testing of hypothesis and making decisions regarding the acceptance or the rejection of hypothesis on the basis of probability). Thus its design as contended by Costello and Osborne (2005) is most appropriate for utilisation in exploring a data set, since it was not designed to test hypotheses or theories. Thus, it is an approach that quantifies or measures the similarity of the factor loadings across groups (i.e. the two language groups) by rotating the two factor solutions to be most advantageously similar, and then computing some sort of similarity index. The principal components factoring (PCF) and principal axis factoring (PAF) are the most popular estimation techniques for exploratory factor analysis (Hair et al., 2010). In EFA there are numerous available alternatives in order to characterise the relationships between the variables. The Pearson correlation matrix is the most conventional correlation matrix employed in exploratory factor analysis. The Pearson correlation matrix necessitates interval scaled data, as opposed to the Spearman correlation matrix which calculates correlations for ordinal scaled data (Thompson, 2004). However according to Kubinger (2003), EFA comprising dichotomous variables, often leads to artificial factors and thus he recommends the use of factoring tetrachoric correlations as opposed to Pearson correlations, to generate more valid results.

Tetrachoric correlation is a distinctive instance of polychoric correlation for dichotomous variables (Wuensch, 2007). Polychoric correlation is based on the assumption that the response categories are actually proxies for unobserved, normally distributed variables. In the more general sense, the measurement variables are ordinal groups. The means and variances of the latent variables are not identified, but the correlation of the dichotomous items can be estimated from the joint distribution of each pair of variables, which results in the tetrachoric correlation coefficient (Edwards & Edwards, 1984).

3.7.5. Executing the Factor Analysis

Factor analysis follows a linear process structure. The first step is to decide on the method of extraction. The current study utilised a Common Factor analysis in order to ascertain whether the variables shared underlying latent factors. Common factor analysis only considers shared common or shared variance, and is suitable for data reduction (Hair et al., 2010). The next step would entail selecting the number of factors to retain. Since an *a priori* factor structure was employed, the use of a scree-plot and its eigenvalues to determine how many factors to retain, was excluded.

The subsequent step would be to decide which rotation method to select. An oblique rotation was decided on for this study, as it produces correlated factors facilitating easy interpretation (Hair et al., 2010). An oblique rotation was employed in the current study seeing that literature suggests that one is likely to discover a relationship between factors (Cummins, 2000). According to Field (2009), oblique rotation requires an examination of the Pattern Matrix table. This is the next step in the Factor Analysis process. In order to consider the relative contribution of each item to a factor, a strict critical value of 0.40 was used to evaluate the factor loadings on the two factors. Items that loaded on more than one factor were regarded as poor items, as at least three items should load on a factor in order for it to be considered a stable factor.

The factor analysis was run separately for the English and Xhosa first-language groups, and the results were compared. The first phase of the factor analysis required the selection of a two-factor solution using the data of the English first-language speaking group first. The other steps that were followed in this analysis will be described in the results chapter. Subsequently, the analysis of the data for the Xhosa first-language group was specified to include the same items, as well as using a two-factor solution.

3.7.6. The Reporting of the Factor Analysis

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 presented and discussed.
5. Steps 2, 3 and 4 will be repeated with the DIF items removed.

The results will be presented for the two phases of the factor analysis separately.

3.2. Ethical Considerations

3.2.1. The Overall Study

The researcher of the primary study administered the Adapted English version as well as the adapted Xhosa version of the WMLS to English and Xhosa first-language learners respectively. The data obtained from these learners was used to evaluate the psychometric properties of both the adapted English and the adapted Xhosa version of the WMLS test in a South African context. This study, however, focused on the adapted English version of the

WMLS, more specifically on the VA scale of the WMLS in order to cross-validate the scalar equivalence of this version to be used across English first-language learners and Xhosa first-language learners.

All research procedures and data collection were done by the researcher of the primary study strictly in accordance with the Ethical regulations of the Nelson Mandela Metropolitan University (NMMU) previously known as UPE. These considerations have been delineated in the above section of this chapter under “Procedure”.

3.2.2. Ethics of this Study

As this study uses SD and forms a sub-study to the primary study, it falls within the ambit of this primary study. The present researcher received permission from the main researcher to utilise and re-analyse the data collected for the primary study (Appendix E). Furthermore, the voluntary and informed consent was obtained by the researchers of the primary study. All data used was handled strictly by the present researcher herself, and the data was stored in a safe and secure place at all times when not in use. All information was completely anonymous since the only identifying data was gender, age, grade and school.

CHAPTER FOUR

RESULTS

4.1 Introduction

This chapter focuses on the overall aim, to assess the scalar equivalence of the adapted English version of the VA scale of the WMLS across the English first-language group and the Xhosa first-language group in a South African context, by evaluating differential item functioning (DIF) and construct equivalence. The two specific aims was analysed by means of either descriptive statistics or inferential statistics. The statistical procedures utilised were the Mantel-Haenszel DIF technique, exploratory factor analysis and the Tucker's phi coefficient. The findings and results from these statistical techniques are summarised into tables and graphs in this chapter in order to facilitate analysing and interpreting the data. This will form the basis for the subsequent chapter in which the implications of these results will be discussed.

Since this study used SDA, the researcher will not be analysing the group differences, namely mean score, mean item characteristics and Cronbach's Alpha, as this was previously explored in Haupt's study (2009) as discussed in Chapter 3 under the section Psychometric Properties and Sampling.

4.2. Differential item functioning displayed on the subscale across the two language groups

Specific Research Aim 1: To evaluate the differential item functioning (DIF) of the English version of the VA scale across English and Xhosa first-language groups. The null hypothesis tested for this aim is:

The probability of scoring 1 on item *i* on the subscale will be the function of ability alone, in other words there will be no items functioning differentially across the English and Xhosa first-language groups on the adapted English version of the VA scale.

The specific research aim was evaluated by means of the Mantel-Haenszel DIF analysis which was conducted across the two groups. The null hypothesis of "no DIF" will be rejected for the items displaying moderate to large DIF.

Tables 10-11 present the results of the DIF analyses of the English version of the VA subscale of the WMLS across the two language groups in the study. Table 10 illustrates the summary of the Mantel-Haenszel DIF procedure across the two language groups for each item on the VA scale, while Table 11 represents the specific DIF items flagged on the VA scale, their estimates as well as their DIF size across the two language groups.

Table : 10

Summary of Mantel Haenszel DIF procedure: Verbal Analogies

Items	MH chi-square	df	Significance	DIF
VA1	0.255	1	0.614	No DIF
VA2	1.582	1	0.208	No DIF
VA3	3.434	1	0.064	No DIF
VA4	0.096	1	0.757	No DIF
VA5	5.510	1	0.019	No DIF
VA6	4.595	1	0.032	No DIF
VA7	0.000	1	0.999	No DIF
VA8	16.044	1	0.000	Large DIF
VA9	26.417	1	0.000	Large DIF
VA10	1.437	1	0.231	No DIF
VA11	6.266	1	0.012	No DIF

VA12	6.673	1	0.010	No DIF
VA13	0.006	1	0.936	No DIF
VA14	9.029	1	0.003	No DIF
VA15	0.483	1	0.487	No DIF
VA16	0.003	1	0.959	No DIF
VA17	0.032	1	0.858	No DIF
VA18	15.095	1	0.000	Large DIF
VA19	0.882	1	0.348	No DIF
VA20	0.196	1	0.658	No DIF
VA21	0.012	1	0.911	No DIF
VA22	0.006	1	0.940	No DIF
VA23	0.654	1	0.419	No DIF
VA24	0.012	1	0.885	No DIF
VA25	0.289	1	0.591	No DIF
VA26	1.235	1	0.266	No DIF
VA27	0.078	1	0.780	No DIF
VA28	2.413	1	0.120	No DIF
VA29	1.997	1	0.158	No DIF
VA31	0.333	1	0.564	No DIF
VA32	0.000	1	0.996	No DIF
VA33	0.333	1	0.564	No DIF

***VA30, VA34, VA 35 displayed no variance**

Using a strict significance level of 0.0001 (p ≤ 0.0001) to detect DIF items, the Mantel-Haenszel DIF procedure identified 3 items all displaying large DIF. The above table indicates that 3 items display DIF (8, 9, 18) corroborating 3 out of the 4 DIF items identified by Haupt's study (2009), in which 4 items (VA5, 8, 9 & 18) were flagged as having large and moderate DIF by means of the logistic regression procedure.

The Mantel-Haenszel DIF procedure cannot identify non-uniform DIF and this could possibly be the reason why it did not flag the fourth item.

Three items (30, 34, 35) displayed no variance in the Xhosa first-language group while only one item (35) displayed no variance in the English first-language group and were thus not included in the analysis. The null hypothesis of “no DIF” was rejected for the 3 items (8, 9, 18) displaying DIF.

Table 11: Verbal Analogies

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

Items 8 and 9 identified in Table 11 above indicate that the English first-language group is favoured on two of the three items. This is in corroboration with the findings of Haupt’s study (2009) where similar results were found using a logistic regression, indicating an overlap in the direction of bias in the two methods used.

Furthermore, according to Haupt (2009) it was argued that these items disadvantaged the Xhosa first-language group as their cultural background was not taken into account. Since the majority of the learners in this group hailed from the rural areas in the Eastern Cape, where certain descriptive words used were not very easily identified, as opposed to the English first-language group, the Xhosa group were inevitably at a disadvantage. An example is item 8 which reads: “train is to track as car is to ______” (answer: road, highway, street or lane). In rural areas where there are not many roads or highways, the Xhosa first-language group was disadvantaged and this item could be viewed as cultural irrelevance (Haupt, 2009).

Item 18 (viz. “Movie is to actor as game is to ______”) on the other hand favoured the Xhosa first-language group. This item was more difficult than the preceding items (Haupt, 2009) and requires higher-order verbal reasoning (Arendse, 2009). This item 18 in comparison to item 8 and 9 might be more appropriate for the Xhosa first-language group since the concepts used were more relatable.

Thus, it could be postulated that the Xhosa group was favoured on a higher-order verbal reasoning item because their reasoning is based on relational similarity, as discussed in Chapter 2 under section 2.3.3. Relational knowledge involves the underlying relations in tasks recognising commonalities between different domains in higher-order thinking. When relational similarity is used then lower-order thinking is absent or abandoned, which could be an alternative explanation for their lack of performance on items 8 and 9. However it could also be a chance finding where the content of the item was familiar to the learners.

4.3. Construct equivalence of the VA scale across the two groups

Specific Research Aim 2: To assess the construct equivalence of the English version of the VA scale across English and Xhosa first-language groups with the DIF items removed.

The factor analysis of the current study was executed in two phases, namely the initial factor analysis with DIF items and a subsequent factor analysis with DIF items removed. As a result, the findings will be presented per analysis, first presenting the results of the factor analysis with the DIF items and next presenting the factor analysis with the DIF items removed.

4.3.1. Steps in conducting the factor analysis

The first phase of the factor analysis required the selection of a two-factor solution using the data of the English first-language-speaking group first. The following steps were followed:

1. A two-factor solution was specified based on a previous study conducted by Arendse (2009) across two language versions of the VA scale of the WMLS, namely an English version and a Xhosa version. This study revealed a stable structure for a two-factor solution across both language versions.
2. Items displaying no variance in either language group were removed (VA 30, 34 & 35).
3. Given the sample size of 192 and 193 respectively for the English first-language and the Xhosa first-language groups, a strict cut off score of 0.40 was used for determining the factor loadings (Hair et al., 2010). In pursuit of an acceptable factor solution, items 1, 6 and 9 were removed as they did not load on either factor 1 or factor 2.
4. This resulted in 29 items ranging from VA2 to VA33 being used for the final solution. This solution provided a stable structure for the final analysis.

The final analysis on the English first-language group thus consisted of a two-factor solution and a total number of 29 items were retained. Subsequently, the analysis on the data for the Xhosa first-language group was specified to include the same items as well as using a two-factor solution.

The Tucker's phi coefficient and scatterplots per factor were used to examine factor congruence.

The final phase consisted of a factor analysis being administered on both groups with the DIF items removed. The factor analysis procedure was discussed in the previous chapter under section 3.7.5. Again, the Tucker's phi coefficient and scatterplots per factor were used to examine factor congruence.

4.3.2. Results of the factors with the DIF items included

4.3.2.1. The pattern matrix results for the two language groups

The results of the pattern matrix for the adapted English version of the VA scale are illustrated in tables 12 and 13 across the two language groups. The names of the two factors (as named in Arendse, 2009) are provided to assist with the interpretation. The naming of the factors is discussed in more detail in section 4.3.4.

Table 12 below indicates the loadings on factor 1 (higher-order reasoning) and factor 2 (concrete reasoning). The two factors are distinguished by their high factor loadings and the sufficient number of items loading on a particular factor and the loadings are as to be expected with the easier items loading on factor 2 (concrete reasoning) and the more difficult items loading on factor 1 (higher-order reasoning).

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

Table : 12

The pattern matrix loadings for the English first language group

English First Language Group		
Item	**Higher-order reasoning**	**Concrete reasoning**
2	-0.32	**0.58**
3	0.09	**0.47**
4	0.00	**0.79**
5	-0.34	**0.79**
7	0.15	**0.68**
10	0.28	**0.53**
11	0.28	**0.61**
12	0.32	**0.46**
8	**0.42**	0.29
13	**0.63**	0.26
14	**0.46**	0.22
15	**0.44**	0.17
16	**0.58**	0.17
17	**0.64**	0.20
18	**0.53**	0.11
19	**0.81**	-0.07
20	**0.93**	-0.14
21	**0.82**	-0.22
22	**0.57**	0.32
23	**0.94**	-0.01
24	**0.85**	-0.01
25	**0.84**	0.07
26	**0.78**	-0.13
27	**0.58**	0.11
28	**0.82**	0.25
29	**0.56**	0.18
31	**0.63**	-0.31
32	**0.62**	0.13
33	**0.68**	-0.28

Table 13 below represents the two-factor solution for the Xhosa first-language group of the adapted English version of the VA scale.

Table : 13

The pattern matrix loadings for the Xhosa first language group

Items	**Higher-order reasoning**	**Concrete reasoning**
3	**-0.40**	**0.57**
4	0.12	**0.47**
5	-0.36	**0.73**
7	0.08	**0.84**
8	0.24	**0.65**
10	-0.01	**0.72**
11	-0.02	**0.90**
12	0.35	**0.50**
13	0.27	**0.77**
14	0.24	**0.71**
15	0.37	0.22
16	0.02	**0.65**
17	0.04	**0.70**
18	-0.31	**0.87**
19	0.33	**0.58**
22	0.31	**0.41**
20	**0.46**	**0.47**
2	**-0.61**	0.36
21	**0.78**	0.38
23	**0.94**	0.01
24	**0.86**	0.16
25	**0.81**	0.13
26	**0.90**	-0.05
27	**0.69**	0.12
28	**0.74**	**0.41**
29	**0.50**	**0.46**
31	**0.57**	0.16
32	**0.43**	**0.43**
33	**1.01**	-0.22

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

4.3.2.2. The Tucker's phi of these factors

The following table represents the Tucker's Phi coefficients on the factor analysis results with the DIF items included.

Table : 14

The Tucker's Phi coefficient per factor

Factor 1	**Factor 2**
0.74	0.79

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

4.3.2.3. The scatterplots of the factors

The following diagrams illustrate the factor pattern coefficients for factor one (figure 1) and factor two (figure 2) of the adapted English version of the VA scale across the two language groups namely, the English first-language group and the Xhosa first-language group respectively.

The diagrams below are used in order to confirm the results obtained in the two-factor solution tables (12-13) as well as the results of the Tucker's phi (table 14). These figures illustrate the relation of items towards an identity line.

As is evident in figure 1 and 2 below the items are not closely aligned across the two groups for both factor one and two. This alludes to a lack of structural equivalence across the two groups, corroborating the findings of the Tucker's phi (table 13) for the factors where the DIF items were not removed.

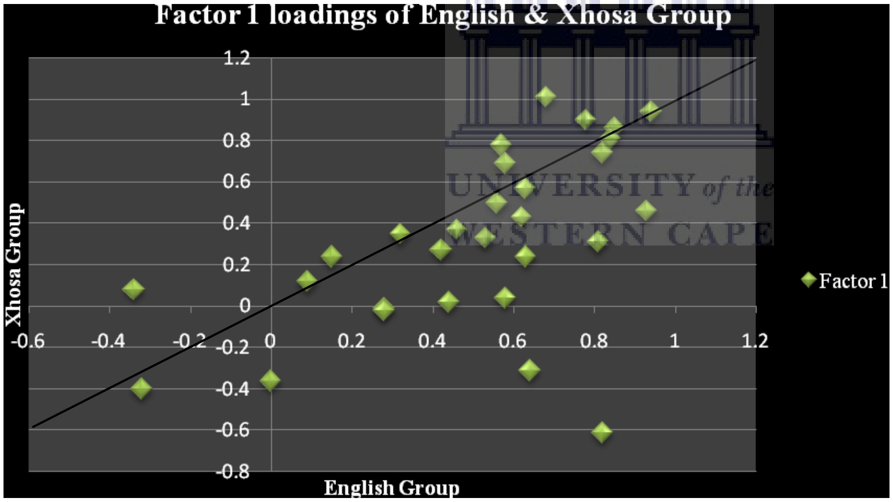


Figure 1: A scatter plot of the factor pattern coefficients for the VA subscale for factor 1 across the English and Xhosa first-language groups with the DIF items

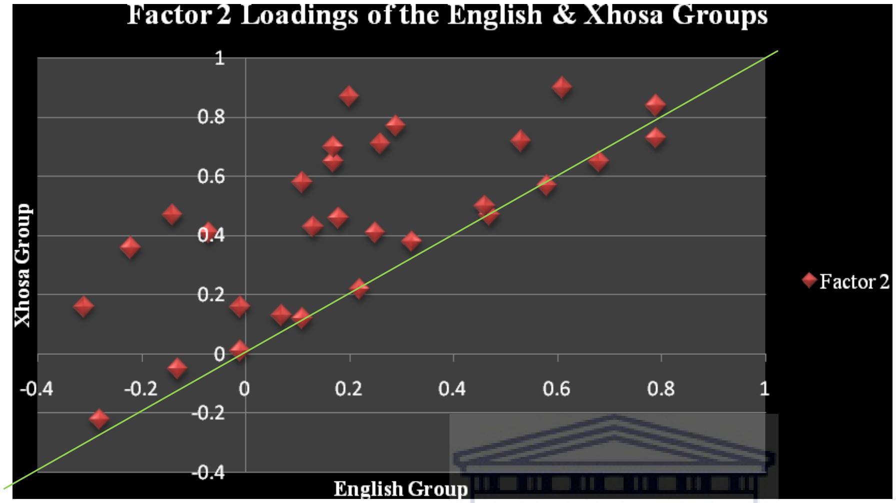


Figure 2: A scatter plot of the factor pattern coefficients for the VA subscale for factor 2 across the English and Xhosa first-language groups with the DIF items

4.3.3. Results of the factors with the DIF items excluded

4.3.3.1. The pattern matrix results across the two language groups

The following tables (15-16) represent the two-factor solution for the English and Xhosa first-language group of the adapted English version of the VA scale with the DIF items removed. Again, the names of the two factors (as named in Arendse, 2009) are provided to assist with the interpretation.

Table : 15

The pattern matrix loadings for the English first-language group with the DIF items removed

English First-Language Group		
Item	**Higher-order reasoning**	**Concrete reasoning**
2	-0.34	**0.58**
3	0.10	**0.46**
4	0.00	**0.80**
5	-0.35	**0.82**
7	0.15	**0.67**
10	0.29	**0.55**
11	0.27	**0.58**
12	0.32	**0.46**
13	**0.63**	0.25
14	**0.47**	0.22
15	**0.43**	0.14
16	**0.57**	0.14
17	**0.65**	0.19
19	**0.81**	-0.10
20	**0.93**	-0.17
21	**0.82**	-0.25
22	**0.57**	0.30
23	**0.95**	-0.02
24	**0.86**	-0.01
25	**0.85**	0.06
26	**0.79**	-0.14
27	**0.58**	0.09
28	**0.81**	0.24
29	**0.56**	0.19
31	**0.66**	-0.31
32	**0.61**	0.13
33	**0.68**	-0.28

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

Table 16 below indicates that the pattern of loading in the Xhosa first-language groups changed when the DIF items were removed. More items, namely 7, 10, 11, 16, 17, 23, 26 and 33 simultaneously loaded on both factors. Only 2 items (4 and 12 –higher-order reasoning as opposed to concrete reasoning) loaded on a different factor compared to the English group.

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

Table : 16

The pattern matrix loadings for the Xhosa first-language group with the DIF items removed

Xhosa First-Language Group		
Item	**Higher-order reasoning**	**Concrete reasoning**
2	-0.37	**0.52**
3	-0.01	**0.62**
4	**0.44**	0.29
5	0.13	**0.60**
7	**0.63**	**0.46**
10	**0.47**	**0.41**
11	**0.57**	**0.58**
12	**0.67**	0.12
13	**0.77**	0.35
14	**0.70**	0.30
15	**0.50**	0.03
16	**0.45**	**0.43**
17	**0.50**	**0.51**
19	**0.71**	0.30
20	**0.77**	0.10
21	**1.01**	-0.10
22	**0.57**	0.11
23	**0.93**	**-0.41**
24	**0.94**	-0.29
25	**0.86**	-0.31
26	**0.83**	**-0.44**
27	**0.74**	-0.20
28	**1.00**	-0.03
29	**0.79**	0.04
31	**0.65**	-0.09
32	**0.70**	0.09
33	**0.84**	**-0.66**

4.3.3.2. The Tucker's phi of these factors

The following table represents the Tucker's Phi coefficients on the factor analysis results with the DIF items removed.

Table : 17

The Tucker's Phi coefficient per factor

Factor 1	**Factor 2**
0.95	0.75

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

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

4.3.3.3. The scatterplots of the factors

The following diagrams illustrate the factor pattern coefficients for factor one (figure 3) and factor two (figure 4) of the adapted English version of the VA scale across the two language groups namely, the English first-language group and the Xhosa first-language group respectively.

Figure 3 below illustrates that the item loadings are fairly closely aligned around the identity line across the two language groups for factor 1 after the DIF items had been removed. This alludes to an indication that factor one with the DIF items removed is structurally equivalent, which corroborates the results of the Tucker's phi illustrating a value of 0.95 (table 17). However, there were a number of item loadings scattered far from the identity line, indicating some problems in line with the discussion in the previous section.

Figure 4 below continues to illustrate items that are not closely aligned even after the removal of the DIF items and thus confirms the results of the Tucker's phi (table 17) indicating that the structural equivalence of factor 2 (concrete reasoning) across the English first-language group and the Xhosa first-language group remains problematic even with the removal of the DIF items.

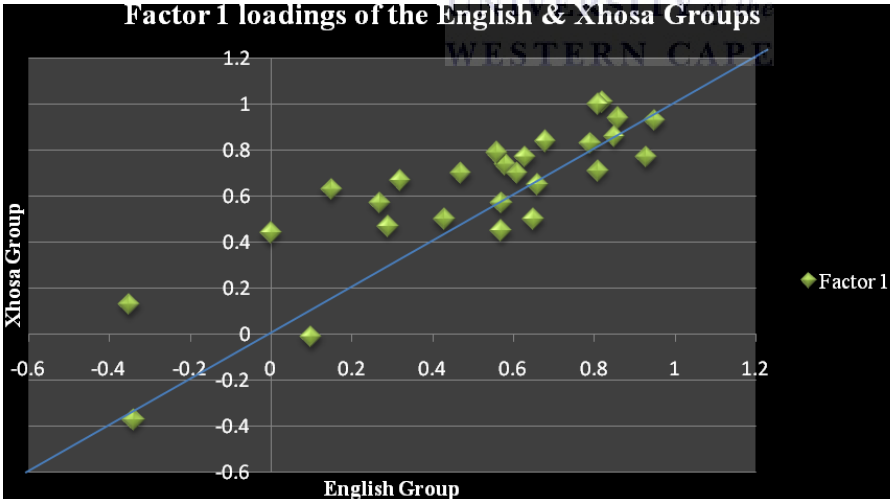


Figure 3: A scatter plot of the factor pattern coefficients for the VA subscale for factor 1 across the English and Xhosa first-language groups with the DIF items removed

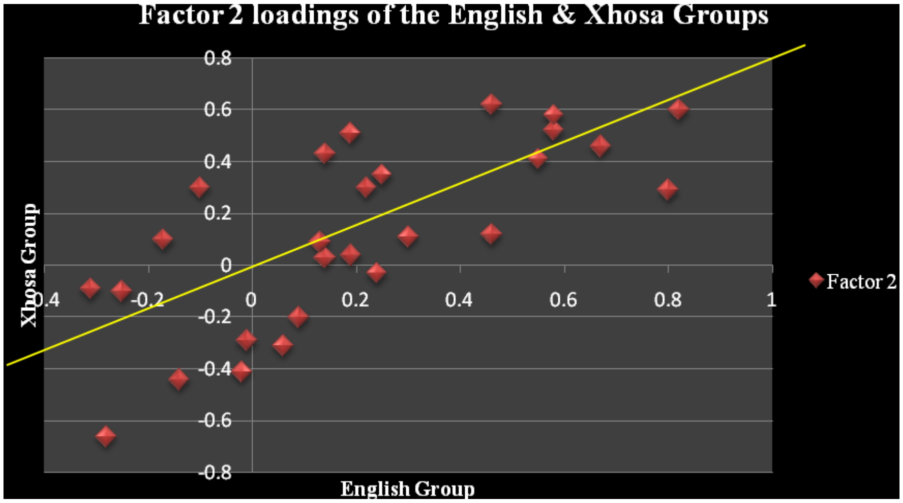


Figure 4: A scatter plot of the factor pattern coefficients for the VA subscale for factor 2 across the English and Xhosa first-language groups with the DIF items removed

4.3.3.4. The Cronbach's Alpha of the two factors across the two groups

Table 18 below represents the findings of the Cronbach's Alpha for both factors across both language groups.

Table: 18

The Cronbach's Alpha for the two factors across the two language groups

	Cronbach's Alpha	
Language Groups	**Factor 1**	**Factor 2**
English first-language group	.83	.64
Xhosa first-language group	.84	.63

The alpha coefficient for factor 1 (higher-order reasoning) of the English first-language group and the Xhosa first-language group is .83 and .84 respectively, suggesting that the items comprising this factor have relatively high internal consistency. The alpha values compare well with the values of the total scale reported in Haupt (2009) of .83 (English first-language group) and .86 (Xhosa first-language group) even though the number of items in the factor are fewer than with the total scale. According to Hair et al. (2010) a reliability coefficient of .70 or higher is considered acceptable. The Cronbach's Alpha coefficient of factor 2 (concrete reasoning) for both groups is below the acceptable value and thus is not regarded as displaying internal consistency.

4.3.4. Naming the factors

The adapted English version of the VA scale items' names were based on the questions of each item. The factors were named based of the content of these individual items. Factor 1 was labelled higher-order verbal reasoning as the items tapped into this domain of reasoning. These items require an individual to display a clear understanding of concepts as well as a conceptual understanding of the items used. The analogies in these items are more indirect and involve more advanced verbal reasoning. Factor 2 was labelled direct verbal reasoning since it involves a direct understanding of the concepts covered in these items. These items involve simple analogies and rely on the individuals using their general verbal reasoning.

The table below presents the names of the items and the two factors as indicated in Arendse's study (2009) of the adapted English version of the VA scale. Due to the confidentiality of the instrument, only the highest loading will be named, as listing the various items names would compromise the test material. The remaining items will be presented with their respective loadings.

Table : 19

The factor names and item names for the English group

Factor 1	**Higher-Order Verbal Reasoning**		**Factor 2**	**Direct Verbal Reasoning**	
	Factor Loading	**Name**		**Factor Loading**	**Name**
13	0.63	*	2	0.58	*
14	0.47	*	3	0.46	*
15	0.43	*	4	0.80	Run-walk
16	0.57	*	5	0.82	Sky-tree
17	0.65	*	7	0.67	*
19	0.81	Shampoo-hair	10	0.55	*
20	0.93	Horse-walk	11	0.58	*
21	0.82	Water-boat	12	0.46	*
22	0.57	*			
23	0.95	Dig-shovel			
24	0.86	Finger-elbow			
25	0.85	Circle-ball			
26	0.79	*			
27	0.58	*			
28	0.81	Scissors-cut			
29	0.56	*			
31	0.66	*			
32	0.61	*			
33	0.68	*			

*****Only items with a factor loading of .80 or above will be named as the test is confidential and enumerating these names would compromise the integrity of the test material.

4.4. Summary

This chapter concentrated on statistical analysis in order to evaluate two specific aims, outlined in Chapter 1 of this study. The researcher has outlined the two aims with both statistical and descriptive analyses.

The first aim assessed the differential item functioning (DIF) of the English version of the VA scale across English and Xhosa first-language groups. The specific aim was evaluated by means of the Mantel-Haenszel DIF technique and three DIF items (8, 9 & 18) were identified across the English first-language group and the Xhosa first-language group on the adapted English version of the VA subscale of the WMLS. As a result the null hypothesis of "no DIF" was rejected for the 3 items displaying DIF.

Aim 2 examined the construct equivalence of the English version of the VA scale across the English and Xhosa first-language groups with the DIF items removed. This aim was evaluated by means of CEFA where a two-factor solution structure based on the English first-language group was used across the two language groups. This two-factor solution provided stable factors across the English first-language group, where items divided into two factors and no factors loaded on both factors simultaneously. When the two-factor structure was applied to the Xhosa first-language group certain items loaded simultaneously on both factors, other factors loaded on the opposite factor, while the remaining items loaded similarly to the English first-language group with both. Factors 1 and 2 both displayed non-negligible factorial incongruence across both language groups.

When the DIF items were removed, factors continued to cross-load as well as load on a different factor in the Xhosa first-language group in comparison to the English first-language group. However, only two items (4 & 12) loaded on the opposite factor as opposed to 12 items that loaded prior to the DIF items being removed. Eight items cross-loaded on both factors in the Xhosa first-language group and only 3 items loaded on factor 2 (concrete reasoning) compared to the eight items loading on this factor in the English first-language group.

The Tucker's Phi coefficient indicated structural equivalence of factor 1 after the DIF items were removed in the English first-language group, while factor 2 continued to display non-negligible incongruities across the two language groups. This is in line with the findings by Arendse's study (2009) where only factor one displayed structural equivalence across the two language versions of the VA scale of the WMLS. However, the fact that so many items in the Xhosa first-language cross-loaded on the two factors (they were included in the calculation of the Tucker's phi for the first factor) remains problematic for the factor congruence of factor 1 (higher-order reasoning).

In the ensuing chapter, Discussion and Conclusion, the results of this study will be discussed in light of the overall aim of the study, which is assessing the scalar equivalence of the adapted English version of the VA scale of the WMLS across the English first-language group and a Xhosa first-language group.

CHAPTER FIVE

DISCUSSION AND CONCLUSION

5.1. Introduction

The overall aim was to assess the scalar equivalence of the adapted English version of the VA scale of the WMLS across English and Xhosa first-language speaking groups -- in other words, to state whether the scores on the VA scale can be utilised across the two language groups. The overall aim was evaluated by means of two objectives: (1) to evaluate the differential item functioning (DIF) of the English version of the VA scale across English and Xhosa first-language groups, and (2) to assess the construct equivalence of the English version of the VA scale across English and Xhosa first-language groups with the DIF items removed.

The current chapter will focus on the major findings of the study, and a comprehensive discussion will be given to identify the implications of these results as well as identify the limitations of the study. Recommendations for future research will be discussed based on these results, as well as concluding remarks on the present study.

5.2. Discussion of the Results

The following results will be discussed in terms of the two sub-aims of this study in order to evaluate the main aim of the study of scalar equivalence.

5.2.1. Results of the Differential Item Functioning

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

Van de Vijver and Tanzer (2004) assert that item bias (DIF) can emanate from a host of sources, especially in cross-linguistic comparison. The most common causes of DIF are poor item translation due to translation errors, ambiguities in the original item, low familiarity or appropriateness of the item content in certain language groups, or influence of cultural specifics such as nuisance factors or connotations associated with the item wording (Van de Vijver & Tanzer, 2004). In light of this, it is evident that the DIF items identified in this subscale could have arisen from numerous causes.

In the present study results indicated that items 8 and 9, identified as having DIF, favoured the English first-language group, while item 18 favoured the Xhosa first-language group. What was interesting in these results was that items 8 and 9 that favoured the English first-language group were among the easier items on the VA scale that required concrete reasoning, while item 18 that favoured the Xhosa first-language group, required higher-order reasoning (Arendse, 2009). Moreover, it was postulated that many of the Xhosa first-language speakers came from the rural areas in the Eastern Cape, where there were not many roads or highways, and thus items 8 and 9 were said to form part of cultural irrelevance as the group might have displayed low familiarity with the item content, as previously discussed under section 4.2. in Chapter Four. In the current study the researcher postulated that item 18 could possibly have favoured the Xhosa first-language group because reasoning on this specific item is based on relational similarity. Relational similarity involves the underlying relations in tasks recognising commonalities between different domains in higher-order thinking (Halford, 1996). Halford (1996) regards this knowledge as central to mechanisms that are basic to human reasoning, such as analogy and planning, as was discussed in Chapter Two.

Van de Vijver and Leung (1997) maintain that DIF can be viewed as providing important information about cross-cultural differences. As a result the unbiased items within the VA scale could define the cultural commonalities of the construct, while the biased items could denote cultural idiosyncrasies.

A conventional approach in dealing with DIF is to deal with it as a distortion at an item level that should be removed (Van de Vijver & Tanzer, 2004). Therefore DIF analysis is used in order to identify and remove biased items and subsequently retain the unbiased items in order that the scale under investigation should entail a sound grounding for comparison across groups. One should remember, though, that even an unbiased measure may not work equally well for different language groups.

5.2.2.   Results of Evaluating Construct Equivalence Across the Two Groups

The second sub-aim of construct equivalence was evaluated by means of an exploratory factor analysis. The results obtained from the factor analysis are formulated at two levels. The first level is in terms of the factor analysis with the DIF items included in the analysis. The second level deals with the general implications of these results with the DIF items removed from the analysis, as well as the use of the Tucker's phi and scatterplots to cross-validate the results found in both factor analyses.

5.2.2.1. Phase one

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

The next step was to evaluate construct equivalence after the DIF items had been removed. According to Sireci and Khaliq (2002), it is considered important to assess construct equivalence after the DIF items have been removed in order to assess their contribution to construct inequivalence, should it occur.

5.2.2.2. Phase two

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

With the Xhosa first-language group, even after the elimination of the DIF items, construct differences continued to occur on both factors. The Tucker's Phi of 0.95 and 0.75, on the two factors across the two groups, indicated that factor 1 was structurally equivalent, with factor 2 displaying structural inequivalence. The VA scale thus continued to appear to measure different constructs across the English and Xhosa first-language groups. As a result, scalar equivalence remains a problem, as it was not shown that the VA scale measures invariant constructs across the English and Xhosa first-language groups.

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

5.3. Implications of the findings

In the current study it was speculated that the detection and removal of DIF items would contribute towards establishing scalar equivalence across the two language groups on the adapted English version of the VA scale. It was assumed that structural differences were due to the presence of differential item functioning, and that DIF would be enough of an explanation to explain these structural differences if they occurred. As a result, if the DIF items were removed, the study would then find that the VA scale was actually measuring the same construct across the two language groups. Even though the Tucker's Phi value improved, providing construct equivalence for factor 1, the same could not be said for factor 2.

The detection and removal of the DIF items for factor 2 did not achieve the desired outcome, and as a result, construct equivalence was not established. Since construct equivalence was not displayed even after the DIF items were removed, differential item functioning is not enough of an explanation for the construct inequivalence found in factor 2. Even though we did not identify a large number of DIF items, evidence still points to the fact that two different constructs were being measured. Thus it could be speculated that because construct equivalence was not attained even after the removal of the DIF items, all the DIF items present might not have been identified. This is in line with previous studies which postulated that the Mantel-Haenszel DIF procedure is not effective in identifying non-uniform DIF (Sireci, Patsula & Hambleton, 2005). Previous research demonstrated that tests available in only one language cannot be utilised across groups in order to compare an underlying construct (Allalouf & Abramzon, 2008; Foxcroft, 2004; Huysamen, 2002; Paterson & Uys, 2005; Rossier, 2004).

In the current study, certain items of factor 2 (concrete reasoning) of the Xhosa first-language group started to load on a different factor (higher-order reasoning). Thus, some of the easier items (the more "direct" items) started loading on the more "indirect" items even after the removal of the DIF items. In other words for the Xhosa first-language group, because English is not their first language, concrete analogy items became higher-order reasoning analogy items. According to Goswami and Brown (1989), Pierce and Gholson (1994), Singer-Freeman and Goswami (1999), analogies become increasingly more difficult if the learner is not familiar with the domain knowledge. The question thus arises, how do Xhosa first-language learners access the appropriate domain knowledge if they lack the language proficiency to understand English instruction in the first place? Poor performance on these items could thus be due to a lack of domain knowledge and not due to inadequate verbal reasoning skills.

In light of the aforementioned discussion it is evident that structural or construct equivalence has not been attained. If construct equivalence was reached after the removal of the DIF items, it would have been possible to continue using the VA scale without the DIF items and as a result scalar equivalence would have been established. The VA scale would have been a step closer to becoming applicable in the South African context. However, the adapted English version of the VA scale continued to display scalar inequivalence and thus the primary aim of this study has not been met.

Concluding remarks on the study are presented in the next section.

5.4. Conclusion

The central focus of this study was to establish scalar equivalence of the adapted English version of the VA scale across English and Xhosa first-language groups. Thus, the researcher wanted to ascertain whether score comparability was possible on this scale in learners whose first language was English and children whose second language was English. In other words, do the scores obtained on the adapted English version VA scale mean the same thing in the two groups, namely that verbal reasoning is being measured? If this is not the case, then the researcher would in actuality be comparing apples and pears, as discussed in Chapter Two under the section Theoretical Framework of Bias and Equivalence (Van de Vijver, 1998). This would in effect constitute construct inequivalence.

Chapter Two explored the construct of verbal reasoning, in terms of its development and some of the issues surrounding monolingual testing. Previous studies have found that monolingual tests are quite problematic when using one test across two language groups (Allalouf & Abramzon, 2008; Foxcroft, 2004; Huysamen, 2002; Paterson & Uys, 2005; Rossier, 2004). The results obtained and presented in Chapter Four of this study confirm previous studies as the adapted English version of the VA scale continues to display scalar inequivalence and thus is not measuring the same construct across the English and Xhosa first- language groups.

Thus until scalar equivalence is established on this scale, it cannot be utilized with confidence in the broader study.

The limitations of the current study are discussed in the next section.

5.5. Limitations

The primary limitation of the current study and the previous sub-study, as well as the overall broader study, is that the sampling procedure was used without considering generalisability, which has implications of affecting the external validity of the adapted English version of the VA scale. In other words, no attempt was made at this stage to explore the applicability of the VA scale for use across diverse language groups in the South African context. The primary researcher, as well as the current researcher, did not deem generalisability crucial at this stage of the research, as focus centred on the internal validity in order to perform the various psychometric procedures. Attention to this will be important in future research.

A subsequent limitation is the rather modest sample size used in the current study. Even though the sample of the study did adhere to the minimum sampling criteria specified for the different statistical methods employed, a larger sample size could have yielded more significant results.

The third limitation is the sole use of factor analysis as this may have yielded spurious factors based on item difficulties instead of on real latent constructs.

A final limitation is that the two groups were matched on the total score of the VA scale. It might seem counter-intuitive to include the studied item itself when calculating a scale score for the matching criterion, especially if the scale is shown to measure two different constructs, as in the case of this study. Furthermore, using the total score of the scale for condition in the DIF analysis, may be the reason why few DIF items were detected as mentioned above. From the construct equivalence analysis, it is clear that we were dealing with test-wide bias (i.e. different constructs were probably measured across the two groups), and this would have affected the results on the DIF analysis leading to the under-detection of DIF items (Koch, 2009).

5.6. Recommendation for Future Research

The researcher would like to recommend further investigation into the construct equivalence of the adapted English version of the VA scale in order that full scalar equivalence could be obtained, so as to use this measure across the two language groups.

Since the study used a Mantel-Haenszel DIF procedure and the groups were matched on ability (as discussed under the section “Limitations”) the current researcher would like to recommend running another DIF procedure using an external conditioning variable in order to alleviate the pitfalls of being matched on the total score of the test. Furthermore, a larger sample size might yield more significant results.

The current researcher recommends the use of the Rasch modelling technique to cross-validate the factor analysis results of this study in order to identify the latent construct with confidence and prevent the identification of spurious factors.

In conclusion, the use of different assessment measures in the South African context creates many challenges today. Thus, obtaining equivalent measures that may be used across a diversity of linguistic and cultural backgrounds is possibly most central in cross-cultural and cross-linguistic comparative research (Huysamen, 2002). With this in mind it is important to note that this study in conjunction with Haupt's study (2009) fall under the umbrella of the broader study, and appear to be in uncharted waters, as these studies constitute one of the first few studies of their kind regarding monolingual language tests and their utilization across language groups. When a measure is biased towards a group, the scores for the group consistently underestimate or overestimate the true values, and as a result becomes a vicious cycle of groups constantly being biased. This study has thus contributed to the need to cross-validate data in the attempt to evaluate the scalar equivalence of the monolingual language measure for use across different language groups in the South African context.

REFERENCES

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

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

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

Alexander, N. (2000). *English unassailable but unattainable: The dilemma of language policy in South African education*. University of Cape Town: PRAESA Occasional Papers No. 3.

Alexander, N. (2002). *An ordinary country. Issues in the transition from apartheid to democracy in South Africa.* Pietermaritzburg: University of Natal Press.

Aloe, M.A. & Becker, B.J. (2009). Teacher verbal ability and school outcomes: Where is the evidence? *Educational Researcher*, *38*(8), 612–624.

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* (9th impression). Cape Town: Oxford University Press.

Bedore, L.M., Peña, E.D., García, M. & Cortez, C. (2005). Conceptual versus monolingual scoring: When does it make a difference? *Language, Speech and Hearing Services in Schools, 36*, 188–200.

Bejar, I.I., Chaffin, R., & Embertson, S. (1991). *Cognitive and psychometric analysis of analogical problem solving.* New York: Springer-Verlag.

Bialystok, E. (2007). Acquisition of literacy in bilingual children: A framework for research. *Language Learning*, *57*(1), 45–77.

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 21 February 2010. Available: http://quantrm2.psy.ohio-state.edu/browne/.

Burton, N.W., Welsh, C., Kostin, I., & van Essen, T. (2009). *Toward a definition of verbal reasoning in higher education* (*ETS RR-09-33)*. Princeton, NJ: ETS.

Campell, 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.

Centeno-Cortés, B. & Jiménez Jiménez, A.F. (2004). Problem-solving tasks in a foreign language: The importance of the L1 in private verbal thinking. *International Journal of Applied Linguistics, 14*(1), 7–35.

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 & Evaluation, 10*(7), 2–9.

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

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

Cummins J. (1986). Empowering minority students: A framework for intervention. *Harvard Educational Review, 56*, 18–36.

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

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

Edwards, J.H. & Edwards, A.W.F. (1984). Approximating the tetrachoric correlation coefficient. *Biometrics*, *40*, 563.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage Publications.

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

Foxcroft, C. (2005). Developing a psychological measure. In C.D. Foxcroft & G. Roodt (Eds.), *An introduction to psychological assessment in the South African context* (2nd ed.) (pp 46–56). South Africa: Oxford University Press.

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

Gallagher, J. M. & Wright, R. J. (1977). *Children' solution of verbal analogies: Extension of Piaget's concept of reflexive abstraction.* Paper presented to the Society for Research in Child Development, New Orleans.

García, O. (2009). *Bilingual education in the 21st century: A global perspective*. London: Wiley-Blackwell.

Gierl, M.J. & Khaliq, S.N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of*

Educational Measurement, 38, 164–187.

Goldman, S.R., Pellegrino, J.W., Parseghian P., & Sallis, R. (1982). Developmental and individual differences in verbal analogical reasoning. *Child Development*, *53*, 550–559.

Goswami, U. (1998). *Cognition in children*. Hove: Psychology Press.

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

Goswami, U. (1989). Relational complexity and the development of analogical reasoning. *Cognitive Development, 4*, 251–268.

Goswami, U. & Brown, A. L. (1989). Melting chocolate and melting snowmen: Analogical reasoning and causal relations. *Cognition*, *35*, 69–95.

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

Gravetter, F.J. & Forzano, L.B. (2008). *Research methods for behavioral sciences* (3rd ed.). Belmont, CA: Thomas Wadsworth.

Hair, J.F., Anderson, R.E., Babin, B., & Black, B. (2010). *Multivariate data analysis.* Upper Saddle River, NJ: Pearson.

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

Hambleton, R.K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, *10*, 229–244.

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

Haupt, G.R. (2009). *The evaluation of the group differences and item bias of the English version of a Standardized Test of Academic Language Proficiency for use across English and Xhosa first language speakers.* Unpublished MA thesis, University of the Western Cape.

Heugh, K. (2003). *Language policy and democracy in South Africa: The prospects of equality within rights-based policy and planning*. Stockholm: Stockholm University Centre for Research on Bilingualism.

Holyoak, K.J. & Thagard, P. (1995). *Mental leaps*. Cambridge, MA: MIT Press.

Huysamen, G.K. (2002). The relevance of new APA standards for Educational and Psychological testing for employment testing in South Africa. *South African Journal of Psychology*, *32*, 26–33.

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

Kamata, A. & Vaughn, B. K. (2004). An introduction to Differential Item Functioning Analysis. *Learning Disabilities: A Contemporary Journal*, *2*(2), 49–69.

Kamwangamalu, N. M. (2000). A new policy, old language practices: Status planning for African languages in a multilingual South Africa. *South African Journal of African Languages, 20*(1), 50–60.

Kao, C. (1998). *Review of the Woodcock-Muñoz Language Survey. The thirteenth mental measurements yearbook*. Lincoln, NE: University of Nebraska Press.

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

Koch, E. (2007). The monolingual testing of competence: Acceptable practice or unfair exclusion. In P. Cuvelier, T. Du Plessis, M. Meeuwis & L. Teck (Ed.), *Multilingual and exclusion. Policy, practice and prospects* (pp79-103). Pretoria: Van Schaik.

Koch, E. (2009). The case for bilingual language tests: A study of test adaptation and analysis. *South African Linguistic and Applied Language Studies*, *27*(3), 67–83.

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, C. (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.

Laija-Rodríguez, W., Ochoa S.H., & Parker R. (2006). The cross-linguistic role of cognitive academic language proficiency on reading growth in English and Spanish. *Bilingual Research Journal, 20*(1), 87–106.

Language in education policy. 1997. Government Gazette 17997(383). Pretoria: Government Printer.

Lenters, K. (2004). No half measures: Reading instruction for young second language learners. *International Reading Association*, 328–336.

Levinson, P.J. & Carpenter, R.L. (1974). An analysis of analogical reasoning in children *Child Development, 45*, 857–861.

Lindholm-Leary, K.J. (2003). *Dual language education*. Clevedon: Multilingual Matters Ltd.

Luzner, E.A. (1965). Problems of formal reasoning in test situations. In P. H. Mussen (Ed.), European research in child development. *Monographs of the Society for Research in Child Development*, *30*(2), Serial No. 100.

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

Nel, N. (2005). Second language difficulties in a South African context. In E. Landsberg, D. Krüger & N. Nel (Eds.), *Addressing barriers to learning. A South African perspective* (pp.149-168). Pretoria: Van Schaik.

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

Oakley, C., Urrabazo, T., & Yang, H. (1998). *When can LEP students exit a BE/ESL Program: Predicting academic growth using a test that measures cognitive language proficiency*. San Diego, CA: American Educational Research Association.

Papalia, D.E., Olds, S.W., & Feldman, R.D. (2004). *A child's world: Infancy through adolescence* (9th ed.). New York: McGraw-Hill.

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

Pearson, B.Z., Fernández, S. & Oller, D.K. (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language and Learning, 43*(1), 93–120.

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

Piaget, J. & Inhelder, B. (1958). *The growth of logical thinking from childhood to adolescence*. New York: Basic Books.

Piaget, J., Montangero, J., & Billeter, J. (1977). Les 22 Child Development correlate. In J. Piaget (Ed.), *L'abstraction reflechusante* (pp. 115–129). Paris: Presses Universitaires de France.

Pierce, K.A., & Gholson, B. (1994). Surface similarity and relational similarity in the development of analogical problem solving: Isomorphic and nonisomorphic transfer. *Developmental Psychology, 30*, 724–737.

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

Prinsloo, E. (2005). Second language difficulties in a South African context. In E. Landsberg, D. Krüger & N. Nel (Eds.), *Addressing barriers to learning. A South African perspective* (pp. 449-465). Pretoria: Van Schaik.

Roccas, S. & Moshinsky, A. (2003). Factors affecting the difficulty of verbal analogies.

Applied Measurement in Education, *16*(2), 99–113.

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

Schaap, P. & Vermeulen, T. (2008). The construct equivalence and item bias of the PIB/SpEEx Conceptualization- Ability Test for members of the five language groups in South Arica. *SA Journal for Industrial Psychology, 34*(3), 29–38.

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

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

Sireci, S. G., Bastari, B., Xing, D., Allalouf, A. & Fitzgerald, C. (1998). *Evaluating construct equivalence across tests adapted for use across multiple languages*. Unpublished paper delivered at Annual meeting of American Psychological Associatio*n*, San Francisco, CA.

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

Sireci, S.G., Patsula, L., & Hambleton, R.K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R.K. Hambleton, P.F. Merenda & C.D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment.* Mahwah, New Jersey: Lawrence Erlbaum Associates Inc.

Solano-Flores, G. & Trumbell, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English-language learners. *Educational Researcher*, *32*(2), 3–13.

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

Smith, E. (2003). Representational thinking as a framework for introducing functions in the elementary curriculum. In J. Kaput, D. Carraher, & M. Blanton (Eds.), *The Development of Algebraic Reasoning in the Context of Elementary Mathematics.* (in press).

Sternberg, R.J. (2006). *Cognitive psychology* (4th ed.). Belmont, CA: Thomas Wadsworth.

Sternberg, R.J. (1977). Component processes in analogical reasoning. *Psychological Review*, *84*, 353–378.

Sternberg, R.J. & Downing, C. (1982). The development of higher-order reasoning in adolescence. *Child Development, 53*, 209–221.

Sternberg, R.J. & Nigro, G. (1980). Developmental patterns in the solution of verbal analogies. *Child development*, *51*, 27–38.

Sternberg, R.J. & Rifkin, B. (1979). The development of analogical reasoning processes. *Journal of Experimental Child Psychology, 27*, 195–232.

Suen, H.K. (2005). *Review of the verbal reasoning tests.* [On-Line]. Accessed 11 January 2010. Available: http://www.BUROS10%20verbal%20reasoning.pdf

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

Tanzer, N.K. & Sim C.Q.E. (1999). Adapting instruments for use in multiple languages and cultures: A review of the ITC guidelines for test adaptations. *European Journal of Psychological Assessment*, *15*, 258–269.

Thomas, W. P. & Collier, V. P. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement: Final report.* Santa Cruz, CA & Washington, DC: Center for Research on Education, Diversity & Excellence (CREDE).[On-Line]. Accessed 11 February 2010. Available: http://crede.berkeley.edu/research/llaa/1.1_final.html/

Thomas, W.P. & Collier, V.P. (1997). *School effectiveness for language minority students.* Washington, DC: National Clearinghouse for Bilingual Education.

Thompson, B. (2004). *Exploratory and confirmatory analysis: Understanding concepts and* *applications* (1st ed.).Washington, DC: American Psychological Association.

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.

Ulstadius, E., Carlstedt, B. & Gustafsson, J. (2008). The multidimensionality of verbal analogy items. *International Journal of Testing, 8,* 166–179.

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

Van de Vijver, F.J.R. & Hambleton, R.K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, *1*, 89–99.

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. *SA Journal of Industrial Psychology*, *30*(4), 1–7.

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

Watson, M.B. (2004). *Career assessment in a multicultural context: A South African perspective*. In International perspectives on career development. Symposium conducted at a joint meeting of the International Association for Educational and Vocational Guidance and the National Career Development Association, San Francisco.

Welman, C., Kruger, F., & Mitchell, B. (2005). *Research methodology* (3rd ed.). Cape Town: Oxford University Press.

Whetton, C. (1985). Verbal reasoning tests. In T. Hussen & N. Postleth-Waite (Ed.), *The International Encyclopedia of Education*. Oxford: Pergamon.

Weunsch, K. L. (2007). Tetrachoric correlation. [On-Line]. Accessed 28 July 2010. Available: http://core.ecu.edu/psyc/wuenschk/sas/Tetrachoric.doc

Whitely, S.E. (1977). Relationships in analogy items: a semantic component of a psychometric task. *Educational and Psychological Measurement*, *37*, 725–739.

Williams, D. & Snipper, G.C. (1990). *Literacy and bilingualism*. New York: Longman.

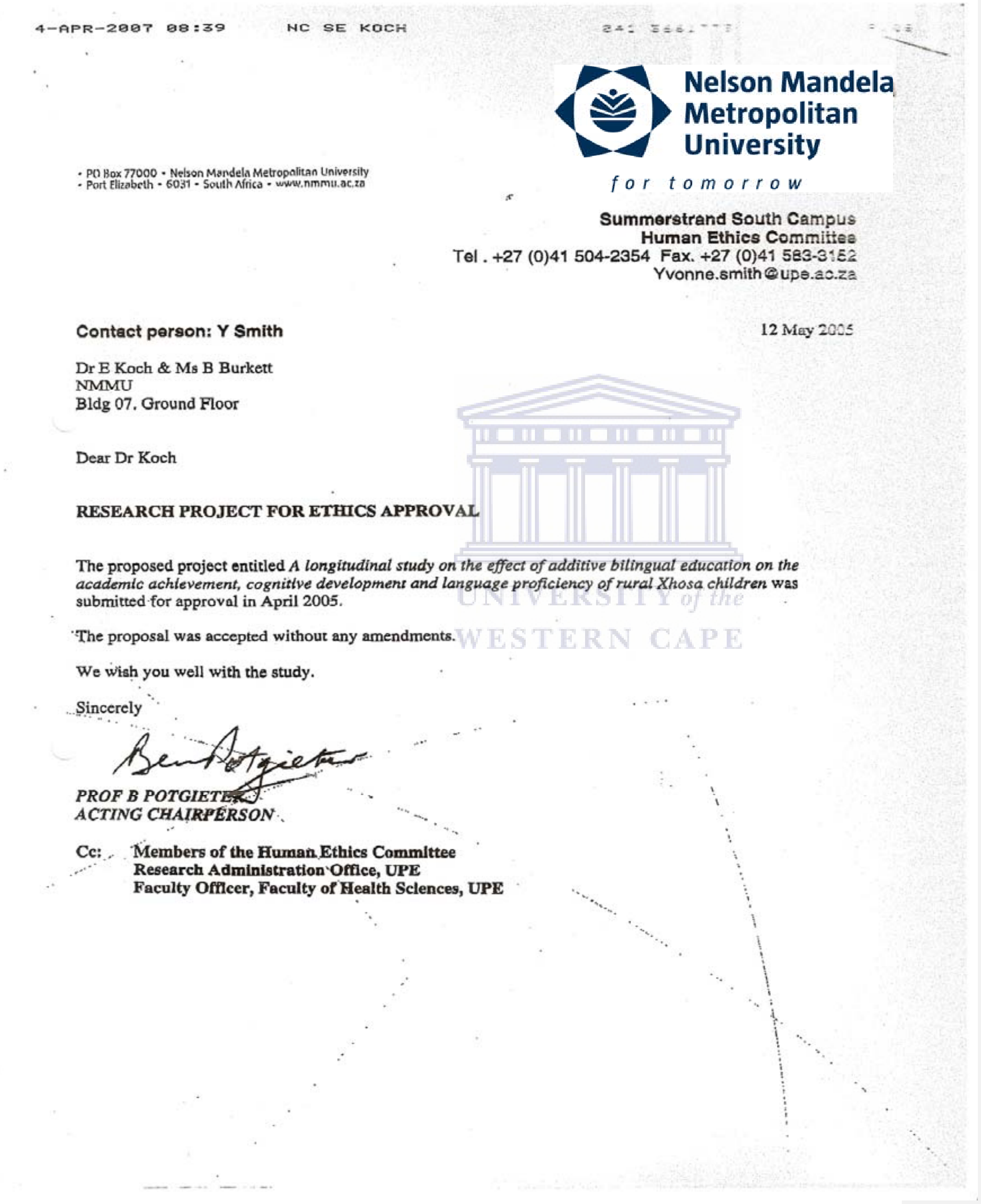
Woodcock, R. W. & Muñoz-Sandoval A. F. 2001. *Woodcock-Muñoz Language Survey: Normative update*. Itasca: Riverside Publishing Company.

Zelazo, P.D., & Müller, U. (2002). The balance beam in the balance: Reflections on rules, relational complexity, and developmental processes. *Journal of Experimental Child Psychology, 81*, 458–465.

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

NMMU Ethics Approval Letter



4-APR-2007 08:39 NC SE KOCH 241 [illegible]

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

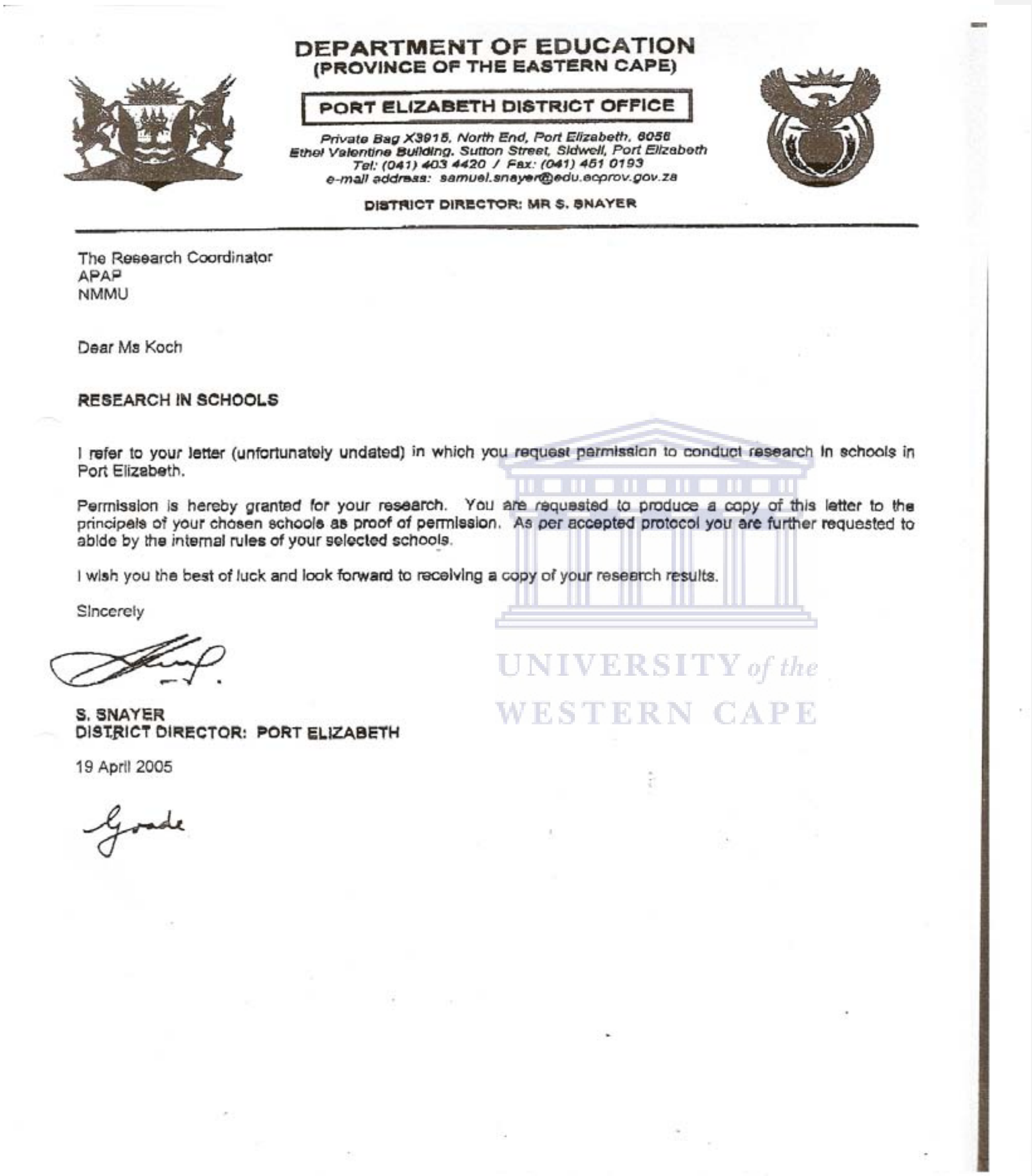
Ben Potgieter

PROF B POTGIETER
ACTING CHAIRPERSON

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

APPENDIX B

Permission from the Eastern Cape Department of Education



DEPARTMENT OF EDUCATION
(PROVINCE OF THE EASTERN CAPE)

PORT ELIZABETH DISTRICT OFFICE

Private Bag X3915, 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

Goade

APPENDIX C

Information Sheet

Nelson Mandela
Metropolitan
University

for tomorrow

2008

Dear Parent

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

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

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

Regards

Dr. Elize Koch

Main Researcher.

APPENDIX D

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."

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

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

Participant name:……………………………………………………………………………

Participant signature (parent):……………………………………Date…………………………

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

Signature of researcher…………………………………..Date…………………………

APPENDIX E

Permission Letter from Main Researcher to Use 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

3/3/

2010

TO WHOM IT MAY CONCERN

I hereby give Ghouwa Ismail permission to use the data originally collected on the English and 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 Verbal Analogies 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

Prof Elize Koch

Principal investigator.