Toward Linguistically Fair IQ Screening:

The Multilingual Vocabulary Test

Julian M. Siebert

ACSENT Laboratory

Department of Psychology

University of Cape Town

Dissertation submitted in fulfilment of the requirements for the award of the degree of

Master of Social Science in Psychological Research

Supervisor: Kevin G. F. Thomas

Word Counts:

Body: 29.712

Abstract: 229

COMPULSORY DECLARATION

This work has not been previously submitted in whole, or in part, for the award of any degree. It is my own work. Each significant contribution to, and quotation in, this dissertation from the work, or works, of other people has been attributed, and has been cited and referenced.

Signature: Signed by candidate Date: 03rd September 2018

Acknowledgements

To start with, I owe thanks to my supervisor, Kevin Thomas, for letting me convince him of this idea and for allowing me copious amounts of freedom in the planning and execution of this project. Moreover, I thank him for convincing me to carry on with this project, for making the necessary arrangements, and for believing in my abilities to complete this project on a very ambitious timeline and with no more than the occasionally needed piece of insightful advice.

Then, countless hours of one-on-one data collection in the ACSENT Lab over less than six months in total would not have been feasible without my research assistants; in no particular order, I thank Lelona Booi, Nqabisa Faku, Donna Herr, and Aqeelah Orrie for their time and efforts during the Study 1a proceedings, and I thank Zara Kavalieratos, Buhle Nongxa, Mareldia Rahman, Rebecca Robinson, Keren Shaulov, Teresa Steyn, Gemma Strohbach, Altay Turan, and Suzél Vosloo for helping with Study 2. Rhiannon Changuion deserves a special mention for making herself available flexibly and extensively during the entire project, and for getting involved above and beyond what a researcher could hope for.

Moreover, working cross-linguistically is not possible without language experts and translators, both from within and outside the academy; I thank Nolubabalo Tyam, Anke Theron, Tessa Dowling, Una Petersen for their translation expertise and language advice.

Then, thank you, Throy, for always believing in me, making me work when I did not want to, and distracting me when I needed to be distracted.

Lastly, this research would not be possible without the individuals who offered their time both in the lab and online; hence, thank you to all the participants in the MVT Research Project.

Abstract

Neuropsychological assessment in linguistically heterogeneous populations is fraught with numerous challenges, such as lacking or inappropriate normative data or the unavailability of appropriate tests. Accommodating multilingual individuals exacerbates the issue by adding the question of which language(s) to use when assessing multilingual individuals. Different test-related concepts may be accessible to them via different languages, as their lexicon is spread out over two or more languages. Hence, any monolingual instrument is likely to disadvantage them. The present set of three studies circumvents this question and presents evidence for an inherently multilingual English/Afrikaans/isiXhosa screening tool for intelligence, the Multilingual Vocabulary Test (MVT). I describe the instrument's development from the pilot study to a psychometric analysis of the final, digitally administered version. For an abbreviated 13-item version, Study 3 ($N = 494$) shows an internal consistency of $\omega = .59$ and Study 2 ($N = 101$) produced significant criterion-related validity values of $r = .46$ and $r = .52$ with the KBIT-2 and Shipley-2 VIQ scores respectively. Linear regression analyses show that, while all criterion measures are biased toward E1-speakers, the MVT is largely immune to test-takers' linguistic background. Thus, the MVT paves the way toward more fairness in cognitive assessments, in general, and provides a promising first step toward addressing one of South African neuropsychologists' greatest needs—that of a quick and easy-to-administer, yet linguistically fair screening tool for cognitive impairment.

Keywords: Cross-cultural neuropsychology, assessment, multilingualism, linguistic fairness, South Africa

Acronyms and Abbreviations

AoA	Age of acquisition
APM	Raven's Advanced Progressive Matrices
COWAT	Controlled Oral Word Association Test
CTT	Classical test theory
MVT	Multilingual Vocabulary Test (digital version)
E1/E2	First-/additional-language English-speaker
FSIQ	Full scale intelligence quotient
g	general intelligence (fluid & crystallized intelligence)
ICC	Item characteristic curve
IIC	Item information curve
IRCCC	Item response category characteristic curve
IRF	Item response function
IRT	Item response theory
KBIT-2	Kaufman Brief Intelligence Test (Second Edition)
LAMIC	Low- and middle-income country
LEAP-Q	Language Experience and Profile Questionnaire
L1/L2	First/second language
MoI	Medium of instruction
MVT	Multilingual Vocabulary Test
NVIQ	Nonverbal intelligence quotient
p-MVT	Multilingual Vocabulary Test (pen-and-paper version)
PHQ-9	Patient Health Questionnaire
PIQ	Performance intelligence quotient
SA-WASI	South African-adapted Wechsler Abbreviated Scale of Intelligence
SEM	Standard error of measurement
SES	Socioeconomic status
SRPP	Student Research Participation Programme
TCC	Test characteristic curve
TIF	Test information function
UCT	University of Cape Town
VIQ	Verbal intelligence quotient
WAIS-III	Wechsler Adult Intelligence Scale (Third Edition)
WAIS-IV SA	Wechsler Adult Intelligence Scale (Fourth South African Edition)
WASI	Wechsler Abbreviated Scale of Intelligence

Table of Contents

List of Tables

List of Figures

CHAPTER 1:

INTRODUCTION

Cross-linguistic neuropsychological assessment is fraught with numerous challenges, including those involving creation of parallel test forms with equivalent psychometric properties across different languages, the oft-lacking validity of simply translated tests, the difficulty of working with broad population norms in heterogeneous populations, and the lack of appropriately stratified normative data (see, e.g., Daugherty, Puente, Fasfous, Hidalgo-Ruzzante, & Pérez-García, 2016; Shohamy, 2011; Shuttleworth-Edwards, 2017; Watts & Shuttleworth-Edwards, 2016). Neuropsychologists in low- and middle-income countries (LAMICs), such as South Africa, point especially frequently to the need for quick, economical, and easy-to-administer cognitive screening tools, primarily to detect possible cognitive impairment in clinical settings (Ferrett, 2011; Razani, Murcia, Tabares, & Wong, 2007; Sabanathan, Wills, & Gladstone, 2015). Most measures currently used in South Africa, however, were developed in the global North, and were standardized on Western (predominantly English-speaking, white, urban, and industrialized) samples (Cockcroft, Alloway, Copello, & Milligan, 2015; Foxcroft, Roodt, & Abrahams, 2005). Such mismatches between the sample used to obtain a measure's normative data and the population within which the measure is administered are likely to result in inappropriate interpretations of test-takers' scores. Such misinterpretations are, in turn, likely to produce misdiagnoses (Daugherty et al., 2016; Manly, Byrd, Touradji, & Stern, 2004; Shuttleworth-Edwards & Kemp, 2004).

The broad aim of this research project was to respond to a central challenge of cross-linguistic neuropsychological assessment, and to a critical demand of South African clinicians. Specifically, I sought to fill a significant void by developing a linguistically fair IQ screening

tool for use with multilingual populations: the *Multilingual Vocabulary Test* (MVT).[1] Even though the current study focuses on South Africa's Western Cape province, the concept bears potential for use in any of the growing number of multilingual populations around the world (European Commission, 2007).

The instrument's development is described across three separate empirical studies in Chapters 3-5. Prior to discussion the studies in detail, however, Chapter 2 provides a review of the relevant literature, where I focus on links between multilingualism, cognition, and psychometric assessment; briefly introduce the basic tenets of contemporary item response theory; sketch a picture of the surrounding issues; and outline the history and current state of cognitive assessment in South Africa.

Then, Chapter 3 (Study 1) describes the first version of the MVT and the process of its development from the initial pen-and-paper version (p-MVT), based on the South African-adapted Wechsler Abbreviated Scale of Intelligence (SA-WASI) Vocabulary subtest, to its digitally administered counterpart. Moreover, the chapter presents the results of a pilot study including a psychometric analysis, as well as some preliminary data on the instrument's sensitivity to test-takers' language backgrounds.

Chapter 4 (Study 2) presents an extended evaluation of a revised MVT that builds on the findings from Study 1 and on additional input from language experts. The chapter describes the MVT revision process as well as the rationale behind both the modifications and the extensions to the analytic strategy. It concludes by presenting detailed results of a psychometric analysis of the revised MVT.

Chapter 5 (Study 3), in a logical continuation, is informed by the results of Study 2. The study uses a greater and more diverse sample in order to obtain a more powerful reliability

[1] Although here the abbreviation *MVT* refers to the instrument in all its versions, generally the term refers to the digital version.

analysis of the revised MVT. Moreover, Study 3 examines the influence of select sociodemographic and linguistic factors on MVT performance.

Together, the three studies make a case not only for measures akin to the revised MVT as linguistically fair IQ screening tools, but for an increased consideration of multilingualism in assessments of overall cognitive functioning. Moreover, the procedures described in Studies 1 to 3 suggest a feasible way of creating inherently multilingual—and hence linguistically fair—instruments. The General Discussion in Chapter 6 provides a summary of, and elaboration upon the lessons learnt about multilingual test development, the influence of various linguistic variables on overall cognitive performance, and other salient issues, and is followed by some concluding comments in Chapter 7.

CHAPTER 2:

LITERATURE REVIEW AND STUDY RATIONALE

Multilingualism and Cognition

Language is one of the most important factors influencing cognitive performance in both everyday settings and on standardized tests, as well as its underlying neural architecture (see, e.g., Abutalebi & Clahsen, 2016; Friederici & Gierhan, 2013; Thierry, 2016). Verbally-based cognitive measures, in particular, are heavily reliant on the test-taker's proficiency in the language of assessment (Blumenfeld, Bobb, & Marian, 2016; S. V. Sanchez et al., 2013; Schwartz et al., 2014). Performance on some of the world's most widely used intelligence tests, such as, for example, the Wechsler family of tests, strongly correlates with test-takers' verbal skills, and hence with their proficiency in the language of test administration (Wechsler, 2008).

One linguistic factor that is relatively rarely examined in the neuropsychological literature is multilingualism, the focus of the present study. Most neuropsychological research in this area is superseded by debates over the existence of a multilingual cognitive advantage in task-shifting, inhibition, and other executive control tasks (see, e.g., Bialystok & Craik, 2010; Bialystok, Craik, & Luk, 2012; Higby, Kim, & Obler, 2013). Instead, because that strand of research is not directly relevant here, the focus of this review is, first, to define and understand the construct of *multilingualism*, and to then describe research investigating effects of multilingualism on cognition, as gauged by neuropsychological test performance. Moreover, although I acknowledge the dependency of (verbal) cognition and neuropsychological test performance on developmental stage, and the existence of research suggesting conflicting findings across different developmental stages, this thesis is explicitly focused on verbal cognitive assessment in multilingual adult populations. Hence, the review does not focus on such

assessment in multilingual child/adolescent populations and does not address work on age as a moderator of cognition and test performance in multilinguals (for reviews of research in those areas, see, e.g., Barac, Bialystok, Castro, & Sanchez, 2014; Bialystok, 2001).

Defining Multilingualism

Traditionally, the psychological literature has used the terms *multilingualism* and *bilingualism* interchangeably. However, in recent years, increasing numbers of papers have begun distinguishing the terms, and multilingualism became the focus of more intense research interest than before. In this thesis, I understand *multilingualism* in the way Aronin and Singleton (2008) use it; for them, it constitutes a generic umbrella term, referring to two or more languages, hence incorporating bilingualism, trilingualism, and so on. More sociolinguistic definitions of multilingualism focus on the distinction between social and individual dimensions of the construct. Such definitions emphasize, for instance, the occurrence of a set of languages within a given society, often on the policy level, and the interaction with a set of languages by an individual (Cenoz, 2013). The latter, stressing "the individual as the locus and actor of contact" (Moore & Gajo, 2009, p. 138) is sometimes referred to as *plurilingualism*, while multilingualism in the societal sense is used more broadly and, for example, encompasses descriptions of monolingual individuals living in multilingual regions.

One widely cited definition of multilingualism is from Li Wei, who states that "anyone who can communicate in more than one language, be it active (through speaking and writing) or passive (through listening and reading)" (2008, p. 4) is multilingual. Another is from the European Commission, which defines multilingualism as "the ability of societies, institutions, groups and individuals to engage, on a regular basis, with more than one language in their day-to-day lives" (2007, p. 6). For the purpose of this thesis, I accept a societal dimension to multilingualism (e.g., I refer to South Africa as a multilingual society and address the resulting

challenges to cognitive assessment on a structural level), but for the most part I use the term with a focus on individuals and their use of multiple languages.

Multilingualism as an Individual Phenomenon

Complicating matters, however, is that linguists have not yet developed a general model of how multilingualism manifests within the individual. For many years, multilingual individuals were considered to be what Grosjean, somewhat disdainfully, calls "two monolinguals in one person" (1989, p. 4). This viewpoint is, by its nature, limited, and hence I subscribe to Grosjean's notion of a holistic view of multilingualism, one that is more nuanced and accounts for a qualification of each individual's unique pattern of multilingualism. Within such a theoretical framework, multilinguals can be considered balanced or unbalanced, depending on how similar their knowledge of their languages is, and how similarly they use them (Bialystok, Craik, Green, & Gollan, 2009; Cenoz, 2013; Grosjean, 1989).

Often, multilingualism research produces inconsistent results, likely due to the many factors that influence the ways in which multilingualism manifests within the individual (Blumenfeld et al., 2016; Grosjean, 2008). Generally, the literature suggests that monolinguals and multilinguals differ with regard to cortical organization (see, e.g., Perani & Abutalebi, 2005), lexical processing (see, e.g., Higby et al., 2013; Kroll, Gullifer, & Rossi, 2013), orthographic and phonological processing (see, e.g., Marian & Spivey, 2003), and other neural and cognitive structures and processes. Moreover, another strand of research suggests that multilinguals differ from one another with regard to these structures and processes. For instance, the level of proficiency, age of acquisition (AoA), degree and length of exposure to the various languages, and the overall number of languages one speaks all influence linguistic processing in multilingual individuals (see, e.g., Bialystok & Craik, 2010; Blumenfeld et al., 2016; D. Klein, Mok, Chen, & Watkins, 2014; Marian, Blumenfeld, & Kaushanskaya, 2007; Wei et al., 2015). In

other words, there is no one kind of multilingualism and, following from that, multilingualism does not occur on a binary scale, but rather on a spectrum. Most importantly, it is not the dichotomous opposite of monolingualism. What all multilinguals have in common, though, is their use (either active, passive, or both) of more than one language.

This fluid account of multilingualism substantiates the argument that the use of monolingual verbal measures in linguistically diverse (and particularly multilingual) populations increases the likelihood of inaccurate performance interpretations (Barac et al., 2014; Sabanathan et al., 2015). While this problem is difficult to conceive in the monolingual mindset, which assumes a perfectly balanced 'two-monolingual' setup (Grosjean, 1989; Shohamy, 2011), the great challenge multilingualism poses to assessment becomes clearer when subscribing to the more recent and better substantiated notion of fluid and individual multilingualism, which is more likely to manifest in an unbalanced manner.

Multilingual Language Processing

Most scholars in the field agree that the cognitive processes involved in language comprehension and production differ between multilingual and monolingual individuals. There is, however, no consensus as to the nature and extent of those differences. On the one hand, there is some evidence for neuro-architectural differences between monolinguals and (early) bilinguals, usually backed up by functional neuroimaging studies suggesting greater grey and white matter integrity in the frontoparietal network and an increased involvement of the basal ganglia in implicit L1 processing (see, e.g., Perani & Abutalebi, 2005; Singh et al., 2017). On the other hand, however, Wong et al.'s (2016) meta-analysis concluded that, for the most part, the brain networks involved in all aspects of language processing overlap between the two groups.

Multilingual Lexical Access

A well-research area of difference between monolinguals and multilinguals is that of vocabulary size (see, e.g., Bennett & Verney, 2018; Bialystok et al., 2012; Portocarrero, Burright, & Donovick, 2007). Generally, such research suggests that even though multilingual individuals typically possess a greater overall vocabulary size (summed across all their languages), their per-language vocabulary is, on average, smaller than that of their monolingual peers. In other words, their vocabulary is partially distributed across their languages, so that some words will be encoded (and will thus be accessible only) in their first language (L1), while others will be encoded (and will thus be accessible only) in their second or other additional language (L2; Bialystok, 2009; Oller, Pearson, & Cobo-Lewis, 2007). Further, while some studies using narrative context tasks suggest that multilinguals' smaller vocabulary does not constitute a performance disadvantage (Barbosa, Nicoladis, & Keith, 2016), Verhallen and Schoonen (1998) and Oller et al. (2007) showed that individuals' greater L1 knowledge cannot compensate for lacking L2 knowledge in other lexical tasks, such as vocabulary knowledge tasks.

Assessment of General Intellectual Functioning in Multilingual Individuals

The concept and measurement of *intelligence* is controversial, especially due to its racialized history. Given the space limitations of a thesis and the circumscribed focus of this research project, I neither weigh in on this debate nor discuss the benefits and shortfalls of defining and/or measuring intelligence in any particular way. Rather, in light of the widespread use of the concept of intelligence across different subfields of psychology and beyond, this thesis takes a practical approach to eliminating one of the shortfalls in the measurement of the construct: its susceptibility to language of administration. Hence, the way I refer to the construct of intelligence relies on Wechsler's open and widely accepted definition: "Intelligence is the

aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with his environment" (1944, p. 3). This ability, commonly termed general intelligence, or *g* (Spearman, 1904), is considered to be comprised of crystallised (verbal) and fluid (non-verbal) intelligence, a distinction underlying the majority of popular IQ scales (see, e.g., A. S. Kaufman & Kaufman, 2004; Shipley, Gruber, Martin, & Klein, 2009; Wechsler & Zhou, 2011).

Returning to the issue of multilingualism, the evidence summarised in the previous section, particularly that relating to differing vocabulary sizes in monolinguals and multilinguals, suggests that multilingual individuals are likely to be disadvantaged when tested using monolingual verbal measures—regardless of the language of administration. With these individual implications of multilingualism in mind, I address the core matter of this dissertation: the linguistically fair assessment of overall cognitive functioning and the detection of cognitive impairment in multilingual individuals. In response to the challenges multilingualism poses to neuropsychological assessment, scholars have attempted to find an answer to the question of which language might be best to use when assessing multilingual individuals. There is currently no generally agreed upon answer to that question, and the debate is ongoing. Although many argue that the most obvious choice is an individuals' home language, Griessel (2005) and Nell (1999) warn that individuals may have acquired many test-relevant concepts, or even individual words, via their medium of educational instruction. Such concepts and words, then, remain potentially inaccessible to multilinguals who are tested in their home language—even if these individuals know the underlying concept in another of their languages (Oller et al., 2007). This factor is especially important to consider when using tests, such as the Wechsler Vocabulary subtests, whose latter halves often feature words that are not part of everyday language and that are primarily used in formal, academic settings.

How do clinicians deal with these issues? Often, English measures are simply translated (sometimes on the spot) or an interpreter is used during the testing session (Brickman, Cabo, & Manly, 2006). Some measures, such as the Bilingual Verbal Ability Test (which is, strictly speaking, a linguistic/cognitive test, rather than a neuropsychological one; Muñoz-Sandoval, Cummins, Alvarado, & Ruef, 2005), go one step further by allowing responses in the test-taker's native language if (and only if) the default English prompt does not elicit a response. Hence, these tests allow second-language English-speakers to resort to their L1. However, the BVAT and other instruments of that ilk fail to truly accommodate multilingualism because, for the purpose of the test, they treat an individual's various languages as hierarchically ranked. Overall, these courses of action, although offering practical alternatives to difficult challenges, are not ideal and can certainly be improved upon—but only if clinicians are offered empirically tested and easily implemented means of testing multilinguals.

Regardless, however, of whether one chooses to test multilingual individuals in their L1 or in their medium of instruction, and given the important moderating role of language in assessment, the decision to administer an instrument in one language only might deny multilinguals access to parts of their knowledge—those parts accessible to them only via their additional language(s). As a consequence thereof, the need for inherently multilingual assessment tools covering all cognitive domains, tailored to the set of languages an individual draws on, is undeniable (Menken & Shohamy, 2015).

South Africa's Linguistic Landscape

In South Africa, language has been a contentious issue for a long time—not only in the domain of cognitive assessment. This is largely because language and race are deeply intertwined in the county's history. The remnants of Apartheid policies still drive the societal divide across the mostly parallel lines of race, class, and language (Alexander, 2013). Regardless

of whether the focus of a law was race or language, both served the Apartheid ideal of a racially segregated society. The Bantu Education Act, for example, maintained the already existing linguistic divide between White and Black South Africans by preventing Black South Africans from learning English (then and now an educational and labour market asset, as well as an elite marker indicating higher social status) and, at the same time, by the banning of African languages from most public spheres. An important manifestation of this policy in the psychological discipline was the disregard of African languages in psychometric test development (Alexander, 2012; Desai, 2013; Foxcroft, 1997).

Since the advent of a new democratic dispensation in 1994, the South African government has recognized 11 official languages (see Table 1), although many more local and officially unrecognized varieties are spoken by its people. Moreover, whereas English and isiZulu function as lingua francas across the country, many languages are more strongly associated with particular regions. For example, the Western Cape province—the setting for the current study—is home to one such distinct, linguistically heterogeneous and predominantly multilingual, population. The official and predominant languages in the Western Cape are English, Afrikaans, and isiXhosa, spoken as a first language only by a combined 39.5% of the population on a national level, but by 20.2%, 49.7%, and 24.7% of the population in the Western Cape province, respectively (Statistics South Africa, 2012a). In other words, most South Africans—and most residents of the Western Cape—speak a language other than English as their L1.

Table 1
Population by First Language Spoken in South Africa and the Western Cape Province

	South Africa		Western Cape	
First language	Frequency	%	Frequency	%
Afrikaans	**6,855,082**	**13.5**	**2,820,643**	**49.7**
English	**4,892,623**	**9.6**	**1,149,049**	**20.2**
isiNdebele	1,090,223	2.1	15,238	0.3
isiXhosa	**8,154,258**	**16.0**	**1,403,233**	**24.7**
isiZulu	11,587,374	22.7	24,634	0.4
Sesotho	3,849,563	7.6	64,066	1.1
Sesotho sa Leboa	4,618,576	9.1	8,144	0.1
Setswana	4,067,248	8.0	24,534	0.4
South African Sign Language	234,655	0.5	22,172	0.4
siSwati	1,297,046	2.5	3,208	0.1
Tshivenda	1,209,388	2.4	4,415	0.1
Xitsonga	2,277,148	4.5	9,152	0.2
Other	828,258	1.6	127,117	2.2
Total	50,961,443	100.0	5,675,604	100.0

Notes. All data from Statistics South Africa (2012a). Where percentages do not add up to 100, it is due to rounding. Unspecified and inapplicable responses are excluded. Official languages in the Western Cape province are in boldface font.

To further complicate matters of multilingual assessment, the three relevant languages are typologically different. Afrikaans has its origin in the first contact of Dutch settlers in the 18th century with the native residents of the region known as the Western Cape province today (Roberge, 2002). Despite having common origins, a particularly pertinent (for current purposes) difference to English is the language's ability to form compound nouns, which makes the meaning of a (compounded) word more accessible than in other languages lacking that ability. isiXhosa, a Bantu language of the Nguni family of languages, is an agglutinating language, which means that information conveyed using a separate word in other languages (such as tense, aspect, or manner) is conveyed by means of a suffix, prefix, or infix (Herbert & Bailey, 2002). This circumstance, as well as the fact that isiXhosa has many borrowings from English (especially in its modern variety), makes it difficult to produce equivalent translations of a given concept while still maintaining equal difficulty and word length standards.

The aforementioned circumstance of domain-specific language use is aggravated by what Grieve (2005) terms a *double disadvantage* experienced by many Coloured and Black residents of the Western Cape (who are mostly speakers of Afrikaans and isiXhosa, respectively). These individuals frequently use their various languages in very distinct and isolated domains. In South Africa, this manifests in may public institutions, such as courts, which—despite being legally obliged to offer trial in the accused's preferred language—mostly operate in English and Afrikaans (Ralarala, 2012), or schools. The majority of schools where the medium-of-instruction (MoI) in the initial years is a language other than English introduce English as the MoI in grade 4 (Taylor & Fintel, 2016), hence forcing learners to navigate at least two languages in the course of their primary and secondary education alone. The change of MoI to English often impedes the development of students' home language, while at the same time their status as additional-language English-speakers still sees them more likely to trail behind native English-speakers, particularly with regard to their vocabulary development (Cockcroft et al., 2015; Oller et al., 2007).[2]

Very often, especially in rural and non-fee-paying schools, the change in MoI remains largely an official one and, practically, sees teachers either switch back and forth between the former MoI and English, or simply continue to teach in the former MoI (Spaull, 2013a). This situation is often the result of poorly qualified teachers and a general shortage of teaching staff at many public schools, yet much less so at fee-paying and private schools (see, e.g., Msila, 2014). Thus, in addition to changes in MoI, one needs to take into consideration the quality and length of education—both of which are very unevenly distributed in South Africa (Spaull, 2013b)—and

[2] Here it is important not to equate apparent or actual conversational fluency in a given single language to test-readiness in that same language. Multilinguals frequently attain conversational fluency, but especially those from minority backgrounds are often found to be disadvantaged in their performance on standardized verbal cognitive tests. This disadvantage is likely due to their reduced exposure to all their languages, compared to monolingual speakers of any of these languages (see, e.g., Hebben & Milberg, 2009).

both of which influence cognitive performance (see, e.g., Manly, Jacobs, Touradji, Small, & Stern, 2002; Rosselli & Ardila, 2003; Walker, Batchelor, & Shores, 2009). Linking this back to the historical inequalities between population groups (which, as shown above, closely resemble linguistic groups) and bearing in mind the hegemonic status of the English language, one can observe that those having attended non-fee-paying schools with MoIs other than English in the first 3 years are at a disproportionate disadvantage (Branson, Hofmeyr, & Lam, 2014; Spaull & Kotze, 2015). Not only do formerly White-only English-MoI schools not experience the change in MoI, they also tend to be, on average, better resourced. These resource difference result in statistically observable differences in terms of quality of education between the different racial and linguistic groups, with (White) E1-speakers much more likely to have received a higher-quality education (Salisbury, 2016; Shuttleworth-Edwards & Kemp, 2004).

Moreover, the frequent mixing of the three languages of the Western Cape province (within and outside the school environment), and the numerous borrowings from one another (both in colloquial varieties and in the standard language), remind us of Grosjean's (1989) hypothesis that all multilinguals likely have their own and unique fluid pattern of multilingualism, which renders the task of choosing the one 'right' language for an assessment even more difficult. Consequently, and given that the majority of (particularly Coloured and Black) South Africans speak more than one language, one might argue that the fairest possible way of assessing them is with an inherently multilingual assessment tool (Barac et al., 2014).

Neuropsychological Assessment in South Africa

Beside linguistic diversity, a major issue South African neuropsychologists face is the lack of relevant post-Apartheid research that fully considers the current South African population's needs and sociodemographic profile. This situation has resulted in the non-availability of linguistically fair tests (Cockcroft et al., 2015; Knoetze, Bass, & Steele, 2005).

Most measures currently in use are only available in English or only normed using English-speaking standardization samples. Only recently have scholars begun to develop appropriate local norms for international measures (see, e.g., Ferrett, 2011; van Wijk & Meintjes, 2015), or to adapt those measures to the South African context (see, e.g., Cawthra, 2016; van Wyhe, 2012; Wechsler, 2014). This development is long overdue, and it helps to finally leave behind the remnants of racially motivated psychometrics.

Currently, two of the most progressive South African-adapted measures are the Wechsler Adult Intelligence Scale-Fourth South African Edition (WAIS-IV SA; Wechsler, 2014) and the South African-adapted Wechsler Abbreviated Scale of Intelligence (SA-WASI; Ferrett, 2011). Although the former has locally relevant normative data, it is only available in English. The latter has been translated into Afrikaans and isiXhosa, but preliminary norms are only available for Afrikaans (see Ferrett, 2011; van Wyhe, 2012). It is worth pointing out here that translating measures from English into other languages gives rise to differential item functioning, through the introduction of a bias toward one particular (in this case) linguistic group (Van De Vijver & Rothmann, 2004).

Consequently, the currently available range of monolingual assessments, even if normed on a South African population, comprises predominantly English and Afrikaans measures. No measure caters appropriately for isiXhosa-speakers (or, for that matter, speakers of the other eight official languages), and no measure takes into account the multilingual reality experienced by most South Africans (Shuttleworth-Edwards, 2016; van Wyhe, 2012). Work on the development and validation of isiXhosa translations of the SA-WASI Vocabulary subtest (see, e.g., Ferrett, 2011; Zieff, 2017) heralds a late, yet welcome addition to the pool of South African-adapted neuropsychological assessment tools. However, it still assumes a monolingual approach

to assessment and does not address the aforementioned problems inherent to neuropsychological assessment in multilingual populations.

Due to the slow development of local or appropriately adapted measures, South African neuropsychologists are left with lack of appropriate tests in languages other than English and Afrikaans, and without verbal IQ measures suitable for use with the country's multilingual population (Foxcroft & Aston, 2006; Shuttleworth-Edwards, 2016; van Dulm & Southwood, 2013). In addition to the lack of development in the field, South Africa also faces infrastructural and financial challenges in the health and education sectors (Branson, Garlick, Lam, & Leibbrandt, 2012; Das-Munshi et al., 2016), both of which mean they require quick and affordable, yet reliable and valid, cognitive assessment tools. All of this occurs in a context that features a high burden of disease and a huge treatment gap (Breuer et al., 2015; Burns, 2015; Watts & Shuttleworth-Edwards, 2016). Hence, South African neuropsychologists require a sensitive screening tool that can quickly indicate the presence or absence of cognitive impairment at the beginning of a process-based testing chain that can ultimately culminate in a comprehensive neuropsychological assessment. The digital version of the Multilingual Vocabulary Test (MVT) addresses precisely this requirement. In doing so, it rests on findings showing that the Wechsler Vocabulary subtest is solidly predictive of full-scale intelligence quotient (FSIQ) scores, with criterion correlations of .7 and above (see, e.g., van Wyhe, 2012).

The construction of an inherently multilingual test, however, is fraught with methodological complexities in terms of test construction, item translations, scoring decisions, and subsequent psychometric analyses. The nature of some of these complexities is described in the next section and is elaborated upon in the empirical studies presented in later chapters.

A Brief Overview of Approaches to Psychometric Test Evaluation

In light of the potentially severe repercussions of using psychological tests that are unreliable, inappropriate for their purpose, or unsuited for administration to the population at hand, psychometric analyses are of utmost importance during test development and evaluation. Two central paradigms in psychometrics are Classic Test Theory (CTT) and the more contemporary Item Response Theory (IRT). The brief review of each that I provide below are not meant to summarize entire textbook-magnitude amounts of material; rather, they serve to familiarize readers without any prior knowledge of psychometrics with the basic principles of and approaches to psychometric test analysis.

Classical Test Theory

CTT provides a scale-level analysis of measurement instruments. Fundamentally, it rests on the notion that, for every individual, there exists a true score (T) denoting, for instance, ability within a cognitive domain or strength of a personal trait. This true score cannot be measured, however; one can only observes someone's test score (X), which is the sum total of the true scores and any errors (e). Those errors can be systematic or random. Systematic errors are specifiable individual or situational effects introduced by invalid measures, those measure that do not tap into the construct of interest, or those that consistently produce inaccurate results. Random errors are trial-specific errors, due to unknown factors, such as the consequence of poorly designed or phrased items or other procedural errors and inaccuracies in the test administration. Random errors are quantified using the Standard Error of Measurement (SEM), which describes the standard deviation of the random errors—assumed to be equal for all individuals. Hence, random errors can be removed from testing by means of thorough reliability analyses. Generally, incorporating e into the theory, the overall notion of CTT is encapsulated in the formula $X = T + e$ (Kaplan & Saccuzzo, 2005; Lord & Novick, 1968).

Validity. The most crucial consideration in test development is validity, or the degree to which a test measures what it purports to measure. The most important type of validity is *construct validity*, or the degree of agreement between the outcome of measure under scrutiny and the underlying theoretical concept. In order to achieve construct validity, the first step in the test development process should be defining a clear target construct—in the case of the MVT, general cognitive functioning.

The most feasible and practical way of establishing whether a new instrument taps into the desired construct is assessing *criterion validity*. This form of validity approximates construct validity, by deeming an established measure of the construct of interest, the criterion measure, as representative of said construct. Criterion validity is then assessed by computing correlations between the sample's performance on the new measure and their performance on the criterion measure (Hall et al., 2014; Slocum-Gori & Zumbo, 2011).

Moreover, recent research suggests that validity is more likely to be achieved if the focus during the test development lies on creating a unidimensional measure (i.e., one that only loads onto the factor of interest). Hence, unidimensionality, rather than internal consistency should be the focus of test development, especially in the early stages. From a CTT perspective, however, a measure cannot be valid unless it possesses high reliability (Clark & Watson, 1995).

Reliability. The most widely reported statistical output in CTT is that of reliability, the ratio between the true and observed score variances (Spearman, 1904a). There are various ways of calculating reliability, such as (a) test-retest reliability (a measure's ability to produce consistent results across multiple administrations), (b) split-half reliability and its Spearman-Brown correction for loss of scale-length (the correlation between equal halves of the same measure), and (c) internal consistency, which is defined as the mean of all split-half correlations, the lower bound reliability, reliability at tau-equivalence, a measure of first-factor saturation, or a

general form of the Kuder-Richardson coefficient of equivalence (Cortina, 1993; Cronbach, 1951). Nonetheless, all reliability coefficients, in some way or another, describe the degree of precision in the measurement, as well as the likelihood the instrument will produce similar results at different times of administration or in different forms.

Beyond reliability, the foci of many statistical analyses within CTT are: (a) item difficulty (i.e., the relative frequency of correct responses to each item on the scale); (b) response frequencies (i.e., the number of times a given response options was chosen); and (c) item-total correlations (i.e, the bivariate correlation between test-takers' score on the given item and their test score).

In this thesis, however, the SEM is of primary interest. The SEM behaves inversely to reliability. Hence, a high SEM is associated with low reliability values, and a measure with a low SEM will be more likely to provide stable results over time and over different forms of administration (Finchilescu, 2013).

Shortcomings of CTT. Despite its popularity as a scale evaluation method and the relatively simple computational methods used to derive its primary outcome statistics, sole reliance on CTT and its reliability-driven assessments is problematic for a number of reasons. Primary among these reasons is that all resultant statistics (including the approximated true score) are specific to the particular sample and the items used for the analysis and, hence, cannot be interpreted in isolation from one another. Moreover, in CTT, *T* is inevitably dependent on the content of the test, and not considered inherent to the test-taker. IRT addresses some of these issues by shifting its analytical focus from the test level to the item level and by attempting to achieve sample-independence (Hambleton, 2000; Hambleton, Swaminathan, & Rogers, 1991).

Item Response Theory

IRT is often also referred to as latent trait theory, which highlights one of its key postulates: the concept of an individual's test-independent unidimensional *latent trait* (or *latent ability*), which the test purports to measure. The latent ability of test-taker j is, then, symbolised by θ_j and it is inferred from their test performance. This consideration of θ allows for an item response function (IRF), or $P_i(\theta)$, which expresses the probability of selecting a certain response of item i at each level of θ. A second postulate is that the relationship between examinees' performance on a given test item and their latent ability level can be modelled mathematically. This foundation on mathematical models renders any IRT model falsifiable through an adequacy assessment using goodness-of-fit statistics (Hambleton, 2000; Magno, 2009). This mathematical modelling becomes possible given a set of assumptions, as outlined below.

IRT assumptions. IRT is subject to stricter assumptions than CTT. The set of mathematical functions underlying all IRT models posits that the probability of a test-taker's responding correctly to a given item is a function of the test-taker's ability and of the item characteristics. As a consequence, IRT models are built on the following assumptions:

- The existence of a test-independent ability, or latent trait, denoted as θ.
- Unidimensionality: All items on a scale measure the same underlying construct or factor. Acknowledging that many psychometric tests are influenced by multiple test-taking, cognitive, and personality factors, the presence of one dominant factor or trait (i.e., θ, the ability of interest) suffices. The assumption of unidimensionality is evaluated by means of a factor analysis. Should this condition not be met (or at least approximated), one can resort to multidimensional IRT models (Lord & Novick, 1968).

- Local independence: Test-takers' responses are solely influenced by θ, which is independent of the instrument used to approximate it. Thus, responses to different items are statistically independent. And while this, at first, seems to be counterintuitive, it is a logical consequence if the assumption of unidimensionality is upheld, and thus allows those two concepts to be treated as equivalent. Local independence can, however, also be achieved in a multidimensional dataset (Hambleton et al., 1991).

Item characteristic and information curves. Item characteristic curves (ICCs) are monotonically increasing functions, which describe the probability of a correct response given the test-taker's ability level. Where, in CTT, one computes item difficulty and discriminability scores separately, in IRT the ICCs provide this information in one instance, through the curve's slope (b_i) and location (a_i), respectively. For polytomous items, item response category characteristic curves (IRCCCs) show the likelihood of selecting each of the available response options in relation to θ (Hambleton et al., 1991).

Another major difference between CCT and IRT is that, in contrast to the concept of CTT reliability, in IRT measurement precision is understood as the amount of information an item and/or a test provides about the latent trait of interest. This measure of precision, termed *item information*, is specific to a certain level of said trait. IRT models are built on the premise that item performance is dependent on the test-taker's ability level (the estimated degree of the latent trait). IRT models also account for the fact that not all items are of equal difficulty for everyone. This then means that, unlike in the case of CTT, the probability of a correct response is a mathematical function of both person and item parameters (Anastasi & Urbina, 2002; Hambleton & Swaminathan, 1985).

The fact that each item's measurement efficiency (and hence the amount of information it provides) differs across ability levels is captured by the item information function (IIF), or $I_i(\theta)$, where that term denotes the amount of information provided for a given item i in relation to θ. IIFs are graphically represented in the form of item information curves (IICs), which plot the amount of information an item conveys on the y-axis over the ability at which it conveys that information (on the x-axis).

Test characteristic and information curves. Both levels of analysis of the item (i.e., the probability of attaining a certain score, $P_i(\theta)$, and the information obtained given a certain level of θ, $I_i(\theta)$), can be transferred to the test level (Lord & Novick, 1968; Ostini & Nering, 2006). The sum of all *item information* constitutes the *test information*. In other words, the sum of all IIFs results in the test information function (TIF), which is graphically represented by the test information curve. Just as IIFs describe the amount of information an item conveys at a given ability level, the TIF maps out the amount of information the set of items (i.e. the test as a whole) conveys at a given ability level. Analogous to the relationship between reliability and the SEM in CTT, test information and SEM are inversely related.

The last IRT component of interest here is the test characteristic curve (TCC), which plots individuals' expected true score given their ability level. This function allows the creation of an output score in circumstances where an ability level cannot be meaningfully interpreted (e.g., a situation where test-takers are compared against an absolute cut-off score, rather than in relation to one another). Prior to computing any of the above, however, one has to settle on one of various IRT models. The process of selecting the appropriate model is addressed below.

IRT model selection. At a primary level, IRT models differ based on the kinds and number of item characteristics that are hypothesized to influence an individual's test

performance. Here, I will only consider two-parameter models, given their greater accuracy in comparison to one-parameter models such as Rasch models (Hambleton et al., 1991).

Next, there is a distinction between models for dichotomous data and those for polytomous data. Given the item format of the MVT, I only considered polytomous IRT models. When dealing with polytomous items, the choice of two-parameter IRT models is restricted to one of Nominal Response Models (Bock, 1972); General Partial Credit Models, a variant of the Partial Credit Model and the Rating Scale Model (GPCM; Muraki, 1992); or Graded Response Models (GRM; Samejima, 1969). Given the MVT's interval response data, Nominal Response Models are ruled out, which leaves GPCMs, as well as GRMs for consideration. Although GCPMs and GRMs often produce similar results (Maydeu-Olivares, Drasgow, & Mead, 1994), the decision of which model to choose is usually made after consultation of a goodness-of-fit test comparing the precision of the competing models (Ostini & Nering, 2006).

Advantages of IRT. The mathematical and probabilistic underpinnings of IRT approaches offer several advantages over CTT approaches. Of primary importance here are these two advantages: (1) item parameter invariance – ability estimates are independent from item choice, and (2) ability parameter invariance – item statistics are obtained independently from sample ability. From a practical perspective, this means that (a) although item and ability parameters influence each other, they can be treated as independent, (b) one can use different items in the creation of distinct, yet equivalent, scales, and (c) the ability and true score results can be considered universal, as opposed to being test-specific (Hambleton, 2000). Moreover, relating item performance to ability level allows one to identify the ability level at which certain items provide useful differentiation and that at which they do not. CTT provides no such statistics, and so IRT is better at providing more nuanced item and scale evaluation (Magno, 2009).

One caveat is that, despite the novel insights IRT provides, the complexities of its calculations as well as the great variety of available models provide plenty of possible sources of error. Hence, rather than blindly subscribing to IRT models in lieu of CTT methods, the results of both are best reported and interpreted alongside one another. I adopted this approach in the empirical studies described in this thesis.

Study Aims and Rationale

The prime rationale underlying the research project described in this Master's thesis was to overcome many of the challenges facing neuropsychological assessment in South Africa (e.g., its linguistically heterogeneous population, the high occurrence of multilingualism, and the lack of infrastructure and resources to routinely conduct comprehensive assessments). The central aim was therefore to develop an inherently multilingual and, hence, linguistically fair measure of general intellectual functioning—one that would be appropriate for South Africa's multilingual population, and that would be able to detect cognitive impairment, regardless of etiology, quickly and accurately. Such a tool needs to be particularly sensitive in the lower ability spectrum, around the cut-off for cognitive impairment.

Importantly, the focus of this study is neither the addition of another language into an existing test, nor the accommodation of speakers of a particular single language by developing a test in that language. Instead, the project's central tenet is the accommodation of multilingualism in cognitive testing. At the same, I seek to detect the presence of a multilingual advantage (see, e.g., Friesen, Luo, Luk, & Bialystok, 2013; Weber, Johnson, Riccio, & Liew, 2015), using a sample comprising both monolingual and multilingual individuals and to identify linguistic variables that influence test performance on the criterion measures and on the MVT.

Furthermore, although the ultimate aim was the development of the said measure (viz., the MVT), an important by-product is that, by incorporating multiple languages in a single

measure, I introduce a new way of thinking about linguistically fair assessment. Hopefully, this induces a broader change of mindset among those developing and working with any type of (verbal) assessment tools. This new consideration applies to assessment regardless of domain or application; it holds true in neuropsychological assessment situations (as is the subject matter of this thesis), but also in the realms of, for instance, educational or other performance assessments.

In summary, this project seeks to establish the MVT as a linguistically fair IQ screening tool, one that elegantly circumvents the constraints of monolingual testing and that can be used confidently in multilingual populations. In so doing, I seek to bring the discipline one step closer to finding a solution for what has been termed “one of the most serious challenges facing the field of neuropsychology” (Razani et al., 2007, p. 107): the fair assessment of multilingual individuals.

CHAPTER 3:

STUDY 1—DEVELOPMENT AND PILOT PSYCHOMETRIC ANALYSIS OF THE MULTILINGUAL VOCABULARY TEST

This study constitutes the pilot investigation for the MVT Research Project with the aim of developing and evaluating a pilot version of a linguistically fair and inherently multilingual screening tool for overall cognitive functioning. The study comprises two logically connected stages: In the first stage, I piloted both a pen-and-paper (p-MVT) and the final digital version of the MVT, before obtaining a validity and preliminary reliability analysis (Study 1a). Subsequently, I assessed the reliability of both versions of the MVT in a more heterogenous sample in Study 1b. Moreover, the latter study helped to pilot the large-scale combined online administration of the MVT and two self-report questionnaires. As a whole, Study 1 laid the foundation for this research project and informed modifications made to the MVT and to the measures and procedures used in Studies 2 and 3.

Methods

Design and Setting

Study 1a study used an intra-individual correlational design, whereas Study 1b was descriptive in nature. Specifically, in Study 1a, I correlated participants' performance on the MVT with their performance on two criterion measures: the 12-Item SA-WASI Vocabulary Subtest (Cawthra, 2016), which served as the basis of the MVT, and Raven's Advanced Progressive Matrices (APM; Court & Raven, 1993; Raven, 2000), a nonverbal measure of intelligence. In Study 1a all psychometric testing was conducted at the University of Cape Town's (UCT) Department of Psychology's ACSENT Laboratory, whereas the surveys were administered online, using the SurveyMonkey platform (www.surveymonkeycom). In Study 1b

the self-report questionnaires and the MVT were administered in a combined online survey, using the SurveyMonkey platform.

Participants

In Study 1a, I used convenience sampling to recruit self-reported multilingual (Afrikaans/English or isiXhosa/English) individuals via the UCT Department of Psychology's Student Research Participation Programme (SRPP). Participants were invited to the study via email. After the exclusion of 2 individuals who had not completed the online survey component, the final sample for Study 1a was $N = 65$ (46 women, 19 men), aged 18-29 years ($M = 20.46$, 95% CI [19.84, 21.08], $SD = 2.49$). They had completed between 11 and 19 years of education ($M = 13.60$, 95% CI [13.22, 13.98], $SD = 1.52$), and all were currently registered within the UCT Faculty of Humanities. A post-hoc power analysis using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009), with $\alpha = .05$, number of predictors = 4, $n = 30$, and an estimated effect size of Cohen's $f = .66$ (corresponding to a partial R^2 value of .40), suggested a linear regression design would yield statistical power of .97. Smaller effect size estimates of $R^2 = .30$ and $f = .25$, and of $R^2 = .20$ and $f = .43$, reduced the achieved power to .86 and .63, respectively.

In Study 1b, I used convenience sampling to recruit participants via UCT's Department of Student Affairs, which circulated an email invitation to the university's general student population. Of the 281 people who responded to the invitation by opening the link to the survey contained therein, 221 completed the MVT only, and 106 completed both the MVT and the remainder of the survey. Consequently, the psychometric analysis of the MVT draws on a sample of $N = 221$, whereas the regression analysis that uses linguistic and demographic variables to predict MVT performance draws on a sample of $n = 106$. The latter comprised 84 women and 22 men aged between 18 and 34 years ($M = 22.78$, 95% CI [22.13, 23.55], $SD = 3.71$), with between 12 and 21 years of education ($M = 15.19$, 95% CI [13.22, 13.98], $SD = 2.90$). A post-hoc power

analysis with $\alpha = .05$, number of predictors = 4, $f = .43$, $R^2 = .30$, $n = 106$ suggested a linear regression design would yield statistical power of .99. Only with effect size estimates smaller than $f = .12$ and $R^2 = .11$ did the achieved power drop below .90.

For both Study 1a and Study 1b, I only enrolled individuals who (a) were aged 18-34 years, an age range consistent with that of the Wechsler IQ scales' reference group (Wechsler & Zhou, 2011), and (b) self-reported not experiencing any psychological, psychiatric, or neurological disorders, and not being prescribed any kind of chronic medication.

Measures

In Study 1a, the sociodemographic questionnaire and the adapted Language Experience and Profile Questionnaire (LEAP-Q) were administered as one combined online survey; all other measures formed part of the study proceedings in the laboratory stage. Study 1b was entirely conducted online, with the revised MVT preceding the sociodemographic questionnaire and adapted LEAP-Q in a single combined online survey.

Sociodemographic questionnaire. This online instrument (Appendix A) collected self-reported information regarding basic biographical variables (e.g., age, sex, current level of education, quality of previous education), socioeconomic status, and brief medical history. All of these factors influence cognitive performance (Hebben & Milberg, 2009; Lezak, Howieson, Bigler, & Tranel, 2012), and hence gathering data regarding them allowed their inclusion in subsequent statistical modelling. Given that previous studies used race to approximate the above socioeconomic factors (see, e.g., Cawthra, 2016; van Wyhe, 2012), the measure also recorded the race participants identified with.

Adapted Language Experience and Profile Questionnaire. This instrument (Appendix B) gathered information about participants' linguistic profile (e.g., order in which languages were acquired, years spent in different language environments, and subjective language dominance

ratings). I adapted (i.e., translated into Afrikaans and isiXhosa and modified for online administration) it from the LEAP-Q (Blumenfeld et al., 2016). The instrument has been successfully translated into 16 languages, which increases confidence in the translations used here (Bilingualism and Psycholinguistics Research Group, 2017). The data gathered from the Adapted LEAP-Q helped to model the influence of language-specific factors on cognitive performance.

Raven's Advanced Progressive Matrices. The APM (Court & Raven, 1993) is a brief nonverbal measure of the fluid intelligence component of general intelligence (g; Spearman, 1904b). It measures abstract reasoning, asking test-takers to complete black-and-white geometric design patterns by choosing one of eight possible options. The APM is widely regarded as one of the closest approximations (and one of the most culture-free measures) of fluid intelligence and of *g* (Mackintosh, 1998; Shuttleworth-Edwards & Kemp, 2004; Strauss, Sherman, & Spreen, 2006). South African studies found internal consistency and test-retest reliability values ($\alpha = .87$ and $\alpha > .90$, respectively) similar to those published in the test manual (Raven, Raven, & Court, 1998).

In an attempt to avoid fatigue effects, I opted to use the 20-minute timed version (see Hamel & Schmittmann, 2006), but otherwise adhered strictly to the procedure outlined in the administration manual (Raven et al., 1998). First, I explained to participants how to correctly respond on the answer sheet provided, which they used for both the practice set and the test set. Then, using the practice set, I explained the task to the participants, illustrating it by pointing out the pattern in item 1, the cut-out patch, and the eight answer options. Next, I ran a finger along the horizontal and vertical lines in the pattern and elicited a response. I indicated incorrect responses and encouraged repeated trials. The process was repeated for the second item. If answered correctly, I instructed participants to complete the practice set in their own time, and

ensured they understood the measure, before proceeding to the 20-minute timed task, using the 36-item set. After 10 minutes, I alerted participants to the fact that half of their allotted time had elapsed.

Despite sometimes being criticized in the debate surrounding culture-free testing, the APM is commonly considered one of the best approximations of culture-fair testing (Shuttleworth-Edwards & Kemp, 2004; Strauss et al., 2006). Some evidence for this claim comes from studies that assessed the APM's cross-cultural validity amongst a heterogeneous group of South African students. Its scores were found to be equally valid for black, Indian, and white individuals (Rushton & Skuy, 2000; Rushton, Skuy, & Bons, 2004). Although there are conflicting findings for other ethnic subgroups within South Africa (see, e.g., Grieve & Viljoen, 2000), the preponderance of the evidence at the time of this study suggested it was the best available measure for the purposes of studies such as this one.

12-Item SA-WASI Vocabulary subtest. The SA-WASI Vocabulary subtest (Appendix C) measures expressive vocabulary and verbal knowledge. In an attempt to improve the instrument's monolingual (English) Vocabulary subtest and to shorten its administration time, Cawthra (2016) developed a condensed form of the instrument that contained 12, instead of the original 34, items. Like its parent, the 12-Item SA-WASI Vocabulary subtest is administered orally and individually, and it requires test-takers to orally explain the meaning of increasingly difficult (in graded order, based on relative item difficulty) words presented to them by the test administrator one at a time. Answers are scored on a scale from 0 to 2. The highest scores are awarded to comprehensive and abstract responses and definitions; a directed, yet incomplete response receives a score of 1; and a vague or irrelevant response receives a score of 0. The instrument is highly reliable (Cronbach's $\alpha = .82$) and has good construct validity, with

correlations of .76 and .70 with the SA-WASI Verbal IQ and FSIQ scores, respectively (Cawthra, 2016).

Multilingual Vocabulary Test. One of the major aims of this research project—and of Study 1, in particular—was to develop this multilingual (Afrikaans/English/isiXhosa) instrument. The MVT is modelled on the 12-Item SA-WASI Vocabulary subtest, which is described above. To meet clinicians' need for a quick IQ screening tool, and given Cawthra's (2016) successful abbreviation of the SA-WASI Vocabulary subtest, the MVT also features 12 items.

Development. The multilingual nature of the measure required a carefully planned word-selection process. To address South Africa's multilingual reality, and to maintain fairness, items were translations of the same concept into Afrikaans, English, and isiXhosa. Items were chosen based on similar frequency of occurrence and similar syllable length across the three languages. Native Afrikaans and isiXhosa speakers, and university lecturers in the relevant language departments, suggested items, translated and back-translated words meeting the above criteria, and provided culturally appropriate definitions to be used in the scoring rubrics.

Format and administration. The MVT was developed in both a paper-and-pencil and digital format (Appendices D and E, respectively), to tackle the need for a quick, easy-to-administer, and self-scored IQ screening tool in the clinical setting. The p-MVT requires test-takers to provide brief oral definitions of 12 words presented to them orally (and, if needed, visually), one at a time. The digital version (hosted on SurveyMonkey) differs insofar as the stimuli are only presented visually, and test-takers are required to select the most correct meaning from five response options. In both versions, items are presented in graded order, from easiest to most difficult, where difficulty was approximated by frequency of occurrence in the 5.3-billion-entry News on the Web Corpus (Davies, 2013).

It is, however, important to note that the version of the MVT used in Study 1b was modified based on the results obtained in Study 1a. First, some minor changes were made to the answer options for items in positions 4, 6, 7, 10, 11, and 12 during Study 1a.[3] These changes were made in response to answers given by Study 1a participants on both the p-MVT and MVT, as well as ongoing discussion with language experts during and after the conclusion of Study 1a. Then, in Study 1b, MVT items were presented in new graded order, informed by the Study 1a item difficulty analyses. The new order of items in Study 1b was: 6, 5, 1, 8, 7, 11, 10, 9, 2, 3, 4, 12 (the numbers here reflecting the Study 1a positions).

An important aspect of the MVT's administration is that test-takers are allowed to respond to each item in whichever language they prefer, as all languages are presented simultaneously. Hence, test-takers can draw on their linguistic knowledge across all three languages, as opposed to only one language, as is the case in the original SA-WASI Vocabulary subtest and the vast majority of other standardized cognitive tests. Such administration likely produces results that provide a more accurate representation of test-takers' overall cognitive abilities (Bialystok, 2009; Bialystok et al., 2012; Nell, 1994).

Scoring. For both the p-MVT and MVT, responses are scored on a 0-1-2 scale. On the p-MVT, test-takers receive a score of 2 for providing a comprehensive definition, a score of 1 for an incomplete, yet directed definition, and a score of 0 for an irrelevant or vague response. On the MVT, test-takers receive a score of 2 for choosing the most correct option, a score of 1 for choosing one of two partly correct options, and a score of 0 for choosing one of two distractors. For instance, for the item (deliberation | deliberasie | ukucamngca), the response (consideration | oorweging | ukucingisisa nzulu), awards the test-takers 2 marks; (carefulness | versigtigheid |

[3]The numbering of items changes across studies. Where references to item number are made, they refer to item position in the current study's order of administration. Where possible, references to items include the current position *and* the actual item wording in English, Afrikaans, and isiXhosa. An overview of item positions across studies can be found in Appendix F.

ukucinga kakhulu) or (thinking | dink | ukucinga ngento), and (freedom | vryheid | ukuqwalasela) or (communication | kommunikasie | ukuphonononga) will results in 1 and 0 marks, respectively (also see scoring rubric for the p-MVT in Appendix G).

Procedure

Individuals willing to participate in Study 1a signed up for a time slot of their choice on the SRPP site, hosted on Vula, UCT's intranet platform. They then received confirmation and subsequent reminder emails containing (a) instructions on how to find the research laboratory, (b) the date and time of their slot, and (c) a link to an online survey containing a consent form, the sociodemographic questionnaire, and the adapted LEAP-Q. Participants were instructed to complete the online survey prior to arriving for their laboratory appointment. The link in the recruitment email took participants to an informed consent document. After giving consent, they were asked to complete the sociodemographic questionnaire and the adapted LEAP-Q. Both questionnaires were available in English, Afrikaans, and isiXhosa, and participants were given the choice to complete them in any one of those languages. Upon completion, they saw a message reminding them to attend the laboratory session they had signed up for.

At the appointed time, I welcomed the participant to the laboratory, provided a detailed explanation of the study purposes and procedures, and explained participants' rights as outlined in the informed consent document. After consenting to participation, myself or one of my research assistants (RAs; four female students recruited from a third-year psychology research class) administered the cognitive measures individually, in separate and quiet rooms, using the exact procedures outlined above. All participants completed the 12-Item SA-WASI Vocabulary subtest and the APM, and the first 37 participants completed the p-MVT. After a preliminary face-value psychometric analysis, I finalized the MVT and changed the administration format to the digital version for the next 30 participants. The three measures (12-Item SA-WASI

Vocabulary subtest, APM, and (p-)MVT) were counterbalanced throughout to avoid practice and fatigue effects.

Upon completion of the test procedures, I used a set of open-ended questions (Appendix H) to encourage participants to comment on their testing experience. I answered any questions participants had, debriefed them, and thanked them for their time. Psychology students received 3 SRPP points and an SRPP participation slip, while all other students received an entry form into a draw, where they stood a chance to win a R1 000 shopping voucher.

In Study 1b, upon clicking on the link in the recruitment email, participants saw an informed consent document. After having read that document and given consent to participate, they saw the MVT instructions and then completed that measure. Subsequently, they were asked to complete the adapted LEAP-Q and the sociodemographic questionnaire (as above). The survey concluded with a page showing a thank-you message, as well as my contact details in case participants were left with any questions. The entire procedure (both Studies 1a and 1b) received ethical clearance for these procedures from the Ethics Review Committee of the UCT Faculty of Humanities (Appendix I).

Statistical Analyses

I used SPSS (version 24.0) to complete all statistical analyses. Unless stated otherwise, assumptions underlying the various types of inferential analyses were met, and α was set at .05 for all decisions regarding statistical significance. Correlations were considered low when less than .40, moderate when between .40 and .70, and high when above .70 (Lachenicht, 2013).

Preliminary analyses. Initial reports of descriptive statistics outlined the sociodemographic and linguistic characteristics of both the Study 1a and Study 1b samples. Independent-sample t-tests assessed between-sex differences for the continuous variables of age, years of education completed, current year of education, and number of languages spoken. Chi-

square analyses or Fisher's exact tests assessed for the presence of between-sex differences for the categorical variables of race, primary and high school types, dominant language, and language acquired first.

Psychometric analyses. In both Study 1a and Study 1b, I calculated the (p-)MVT's internal consistency and conducted an item difficulty analysis. In Study 1a, bivariate correlational analyses (using Pearson's *r*) described the magnitude of association between participants' performance on the MVT (both paper-and-pencil and digital versions) and the criterion measures. In Study 1a, correlating (p-)MVT scores with scores on the 12-Item SA-WASI Vocabulary Subtest and APM helped determine the measure's construct validity as an IQ screening tool.

Regression modelling. In Study 1b, three simple linear regression models sought to identify linguistic factors predicting performance on the MVT. The findings of significant predictors were corroborated by means of independent-samples *t*-tests comparing mean test performance between the relevant groups under investigation.

Results

Sample Characteristics

In Study 1a, all participants ($N = 65$) completed the 12-Item SA-WASI Vocabulary Subtest and the APM, 35 completed the p-MVT, and 30 the MVT. In Study 1b, 221 participants completed the MVT, and 106 of those completed the entire set of questionnaires (viz., the LEAP-Q and sociodemographic questionnaire) as well. Tables 2 and 3 summarize the key sociodemographic characteristics of the Study 1a sample and the Study 1b sub-sample of, respectively. (A similar table could not be generated for the entire Study 1b sample ($N = 221$) because 115 (i.e., 221 – 106) of those participants did not complete the sociodemographic questionnaire.)

Table 2
Study 1a: Sample's Sociodemographic Characteristics (N = 65)

Variable	Entire Sample (N = 65)	Women (n = 46)	Men (n = 19)	t / χ^2	p	ESE
Age (years)	20.46 (2.49)	19.74 (1.24)	22.21 (3.71)	4.05	< .001***	1.10
Education (years completed)	13.60 (1.52)	13.43 (1.31)	14.00 (1.92)	1.37	.174	0.37
Number of Languages	2.62 (0.90)	2.63 (0.90)	2.58 (0.90)	0.21	.835	0.06
Race				3.47	.304	.40
Black	26 (40.00)	17 (36.96)	9 (47.37)			
Coloured	24 (36.92)	20 (43.48)	4 (21.05)			
White	12 (18.46)	7 (15.22)	5 (26.32)			
Other/Not declared	3 (4.55)	2 (4.35)	1 (5.26)			
Dominant Language				3.23	.369	.32
Afrikaans	6 (9.23)	5 (10.87)	1 (5.26)			
English	43 (66.15)	29 (63.04)	14 (73.68)			
isiXhosa	15 (23.01)	12 (26.09)	3 (15.79)			
Other	1 (1.54)	---	1 (5.26)			
Language Acquired First				2.63	.426	.47
Afrikaans	9 (13.85)	8 (17.39)	1 (5.26)			
English	29 (44.62)	21 (45.65)	8 (42.11)			
isiXhosa	25 (38.46)	16 (34.78)	9 (47.37)			
Other	2 (3.08)	1 (2.17)	1 (5.26)			

Notes. For the continues variables (*Age*, *Education*, *Number of Languages*), means are presented with standard deviations in parentheses. For the remaining (categorical) variables, frequencies are given with percentages in parentheses. Between-group differences were assessed using independent-samples *t*-tests for the continuous variables and Fisher's exact tests for the categorical variables (as some of the expected cell frequencies were smaller than 5). ESE = effect size estimate (Cohen's *d* for continuous variables and Cramer's *V* for categorical variables). If percentages do not add up to 100%, it is due to rounding.
***$p < .001$, two-tailed.

Table 3
Study 1b: Sociodemographic Characteristics of Regression Subsample (n = 106)

Variable	Entire Sample (n = 106)	Women (n = 84)	Men (n = 22)	t / χ^2	p	ESE
Age (years)	22.78 (3.71)	22.60 (3.44)	23.50 (4.63)	1.02	.311	0.24
Education (years completed)	15.19 (2.90)	15.14 (2.71)	15.36 (3.59)	0.32	.752	0.08
Number of Languages	2.80 (1.12)	2.74 (1.10)	3.05 (1.21)	1.14	.256	0.27
Race				5.02	.161	.20
Black	14 (13.21)	9 (10.71)	5 (22.73)			
Coloured	15 (14.15)	12 (14.29)	3 (13.64)			
White	62 (58.49)	53 (63.10)	9 (40.91)			
Other/Not declared	15 (14.15)	10 (9.43)	5 (22.73)			
Dominant Language				6.74	.999	.06
Afrikaans	5 (4.72)	4 (4.76)	1 (4.55)			
English	94 (88.68)	74 (88.10)	20 (90.91)			
isiXhosa	1 (0.94)	1 (1.19)	---			
Other	6 (5.6)	5 (5.95)	1 (4.55)			
Language Acquired First				2.45	.475	.15
Afrikaans	17 (16.04)	15 (17.86)	2 (9.10)			
English	50 (47.17)	54 (64.29)	16 (72.72)			
isiXhosa	5 (4.72)	3 (3.57)	2 (9.10)			
Other	14 (13.21)	12 (14.29)	2 (9.10)			

Notes. For the continues variables (*Age*, *Education*, *Number of Languages*), means are presented with standard deviations in parentheses. For the remaining (categorical) variables, frequencies are given with percentages in parentheses. Between-group differences were assessed using independent-samples *t*-tests for the continuous variables and Fisher's exact tests for the categorical variables (as some of the expected cell frequencies were smaller than 5). ESE = effect size estimate (in this case, Cohen's *d* for continuous variables and Cramer's *V* for categorical variables). If percentages do not add up to 100%, it is due to rounding.
[a]Data from all those currently studying (n = 99, 78 women, 21 men)

All Study 1a participants had at least matriculated from high school (i.e., completed at least 12 years of education). The modal participant in that study was black[4], female, primarily English-speaking, and studying at second-year level. Analyses detected no significant between-sex differences with regard to years of education completed, year of study, number of languages spoken, race, dominant language, language acquired first, and number of languages spoken (all assumptions other than that of normality were upheld). With regard to age, however, the analyses detected a significant between-sex difference, but simple linear regression models showed that age was not a significant predictor of performance on any of the outcome measures. This latter result, and the relative homogeneity of the variable with regard to age (ranges were 18-23 and 18-29 for women and men, respectively), allowed me to disregard that variable as a predictor of cognitive ability in this study.

The modal participant in Study 1b was a white, female, first-language English-speaker, studying at the postgraduate level. Analyses detected no significant between-sex differences with regard to age, years of education completed, current year of study, number of languages spoken, race, dominant language, and language acquired first. The assumption of normality was not met for age and current year of education, which, again, demands a cautious analysis.

[4]Here, the term ‘black’ is used in Biko’s sense, encompassing all historically disadvantaged population groups in South Africa. Where capitalised (Black), it serves as a finer distinction between the different historically disadvantaged race groups. Throughout this work, I avail myself to the population group labels commonly utilized by Statistics South Africa: Black (African), Coloured, Indian or Asian, and White(e.g., Statistics South Africa, 2012b).

Performance on Cognitive Measures

Table 4 presents a summary of the Study 1a participants' performance on the various outcome measures. Because normative data was unavailable for the measures, no standardized scores could be computed; hence all scores are raw scores. For the APM, however, performance could be evaluated by comparing mean scores obtained here to those Hamel and Schmittmann (2006) obtained in their sample (N = 397, M = 21.19, SD = 4.29). A one-sample t-test suggested that, on average, participants in the current sample obtained significantly lower scores than those in the Hamel and Schmittmann sample, $t(64) = 7.64$, $p < .001$. On a descriptive level, this amounts to a mean difference of almost 1 SD.

Further, given the uneven sex distribution of those who completed the p-MVT (n = 35; 30 women, 5 men), I investigated the effect of sex on cognitive performance (see Appendix J for performance by sex). As the Table shows, the analysis detected no significant between-sex differences on any of the measures. Further, none of the regression coefficients obtained in a set of single linear regression models assessing the predictive power of sex on those four outcome measures were statistically significant. Hence, it is safe to conclude that, in this sample, test performance on the p-MVT, MVT, 12-Item SA-WASI Vocabulary subtest, and APM was not significantly influenced by sex.

Despite the changes made to the MVT before its administration in Study 1b, and despite the bigger sample size in that study (N = 221 for the psychometric analysis), performance on the test was similar: For Study 1a, M = 17.60, 95% CI [16.71, 18.49], SD = 2.39, and for Study 1b, M = 17.88, 95% CI [17.58, 18.19], and SD = 2.31. Statistical analyses confirmed this impression of similarity, $t(249) = 0.63$, $p = .531$, Cohen's d = 0.08. Figure 1 shows the distribution of Study 1b MVT scores. As is clear, the scores approximate the desired normal distribution.

Table 4
Study 1a: Performance on the Outcome Measures (N = 65)

				95% CI (Means)		Sex differences		
Measure	*M*	*SD*	*SE*	*LL*	*UL*	*t*	*p*	ESE
p-MVT (*n* = 35)	15.20	2.71	0.46	14.27	16.13	0.35	.727	0.17
MVT (*n* = 30)	17.60	2.39	0.44	16.71	18.49	0.09	.929	0.03
SA-WASI	12.08	3.97	0.49	11.09	13.06	1.50	.410	0.41
APM	17.29	4.11	0.51	16.27	18.31	0.69	.493	0.19

Notes. Mean raw scores are presented, with standard deviations in parentheses. Between-group differences were assessed using independent-samples *t*-tests. p-MVT = pen-and-paper Multilingual Vocabulary Test; MVT = digital Multilingual Vocabulary Test; SA-WASI Vocabulary = 12-Item South African-adapted Wechsler Abbreviated Scale of Intelligence Vocabulary subtest; APM = Raven's Advanced Progressive Matrices; ESE = effect size estimate (in this case, Cohen's *d*).

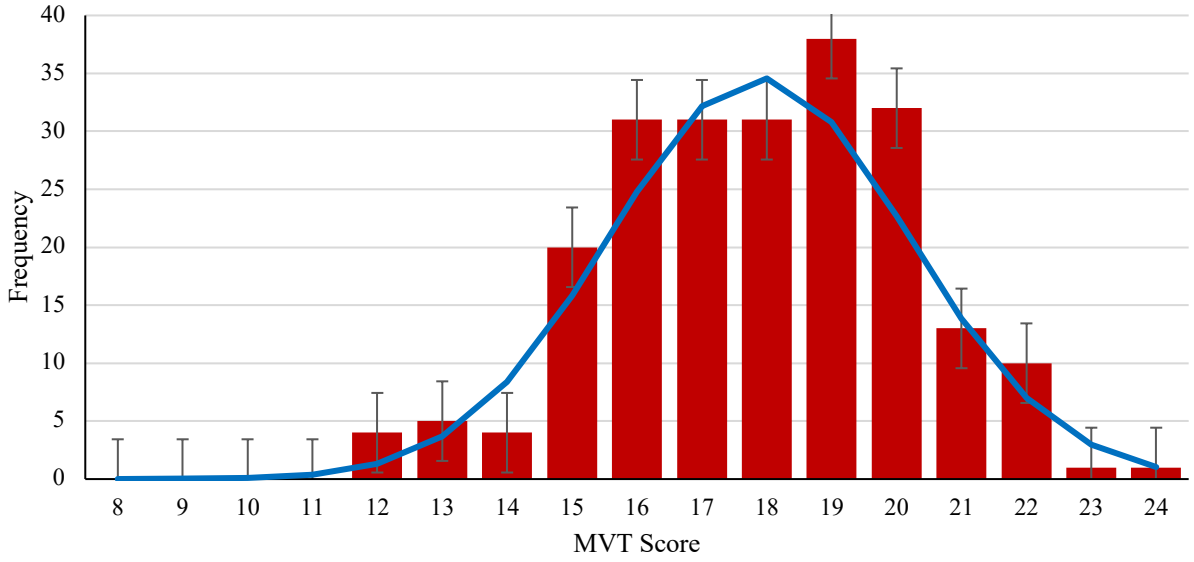


Figure 1. Distribution of MVT scores in Study 1b (N = 221), with normal curve. The entire range of scores is depicted.

MVT Psychometric Analysis

This section presents a brief psychometric analysis of the pilot version of the (p-)MVT. Given its preliminary nature, it focuses on a CTT approach and reports the instrument's reliability and validity, as well as an item analysis.

MVT reliability. In Study 1a, Cronbach's α was .37 for the p-MVT and .24 for the MVT (12 items each). Both of these values are too low for the measure to be considered reliable

(Finchilescu, 2013). When deleting individual items, alpha increased marginally to maximally α = .43 (deletion of item 8) and α = .33 (deletion of item 7) for the p-MVT and MVT, respectively. Due to the different modes of administration, however, the weak items differ between the versions. Split-half reliability estimates, using the Spearman-Brown correction to account for loss of scale length, produced a marginally higher reliability coefficient for the p-MVT, r =.44, but a lower one for the MVT, r = .24. These low reliability values mean that the instrument requires revision as it is unlikely to produce consistent results across multiple administrations. However, for illustrative purposes, I decided to continue with the analysis.

Analyses of the Study 1b data (N = 221) produced a Cronbach's coefficient alpha value of .73, indicating strong internal consistency. This increase of α by a magnitude of .36 over the Study 1a value allows the conclusion that the revisions made between Study 1a and Study 1b were effective, and it provides compelling evidence for the reliability of the modified MVT.

MVT item analysis. Figure 2 displays the item difficulty levels and response patterns for the p-MVT (Study 1a). Apart from the small drop for item 2 (*picture* | *prent* | *umfanekiso*) and the spike of item 11 (*effort* | *poging* | *umzamo*), item difficulty is relatively low and constant up to item 6, and then gradually and smoothly increases up to the last item. Even though the frequency of 1- and 2-mark responses is erratic for the first eight items, the curves cross at item 9, indicating that, from this item onward, more people scored 1 mark than 2 marks—another indicator of increased difficulty.

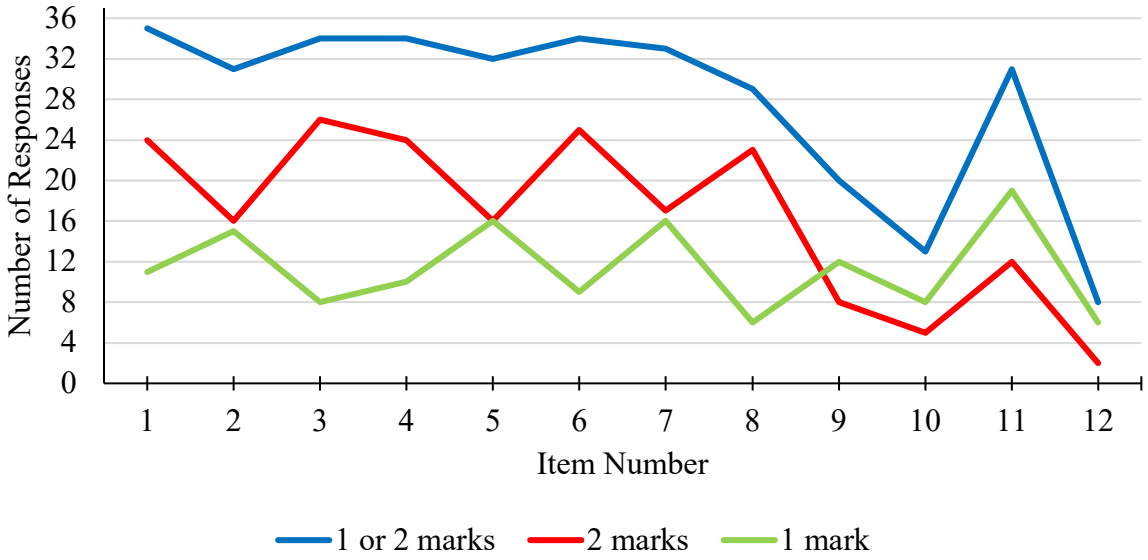


Figure 2. Relative item difficulty curves for the p-MVT (n = 35) in Study 1a.

Figure 3 displays the item difficulty levels and response patterns for the original version of MVT (Study 1a). Here, the pattern is less straightforward: Overall item difficulty remains relatively constant, and it is difficult to discern a clear response pattern across the first four items. From item 5 onward, however, more people score 2 marks than 1 mark, a contraindication of item difficulty. This drastic change in response pattern, despite featuring the same stimuli as the p-MVT, is most likely due to the multiple-choice administration format.

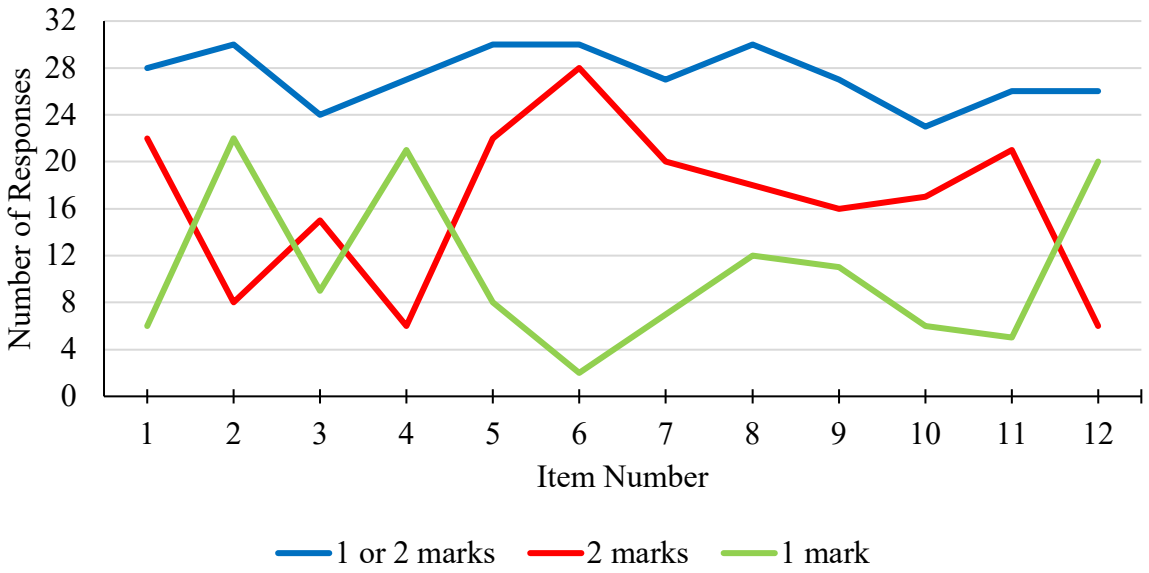


Figure 3. Relative item difficulty curves for the MVT (n = 30) in Study 1a.

Hence, before beginning Study 1b, I modified the MVT item order and the response options associated with items 4, 6, 7, 10, 11, and 12. As Figure 4 illustrates, these modifications had positive effects. For the modified MVT, as administered in Study 1b, the overall item difficulty curve showed a fairly smooth, yet slow downward trend, with the exception of items 11 (formerly item 4, *announce* | *aankondig* | *ukwazisa*) and 12 (*tumult* | *rumoer* | *isidubedube*). Even though the pattern is erratic from item 7 onward, from items 1 to 6, more test-takers score 2 marks than 1 mark, with a downward trend, indicating an appropriate difficulty grading.

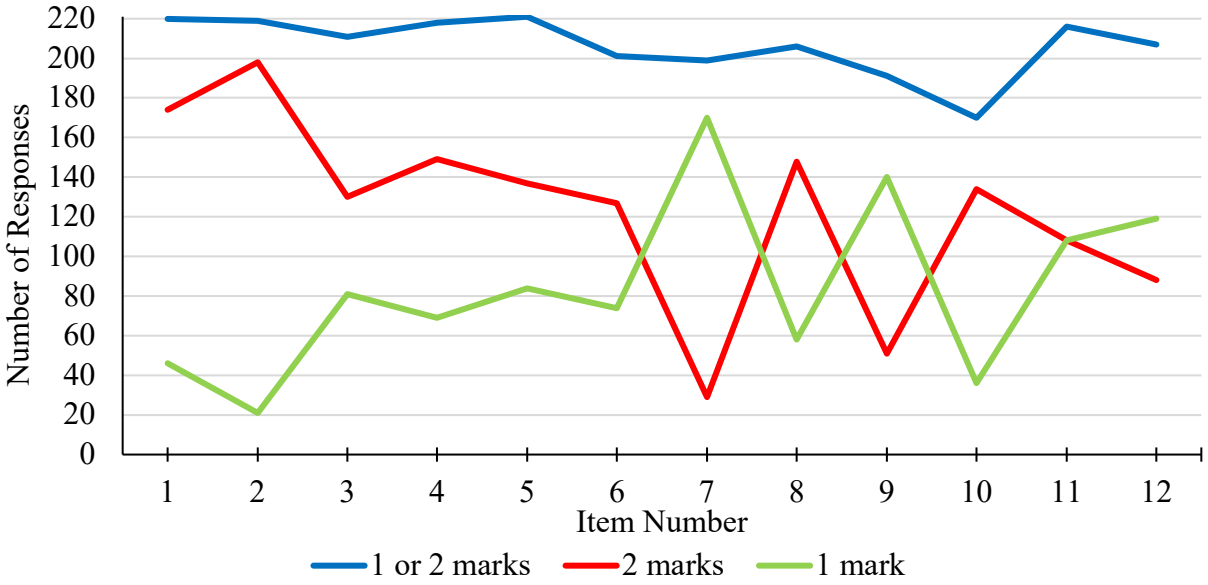


Figure 4. Relative item difficulty curves for the MVT ($N = 221$) in Study 2.

MVT criterion validity. In Study 1a, the analysis detected a significant, moderate, positive correlation between scores on the p-MVT and those on the 12-Item SA-WASI Vocabulary subtest, $r(33) = .52$, $p = .001$, as well as a smaller, non-significant positive correlation between scores on the p-MVT and those on the APM, $r(33) = .20$, $p = .246$.

The strength of the correlation between the two measures differed, however, depending on the test-taker's dominant language. When examining data only from those who reported English as their dominant language ($n = 24$), there was a significant, strong, positive correlation between scores on the p-MVT and those on the 12-Item SA-WASI Vocabulary Subtest, $r(22) = .77$, $p = .006$. In contrast, when examining data only from those who reported Afrikaans or isiXhosa as their dominant language ($n = 11$), the statistics were $r(9) = .35$, $p = .090$. This pattern of data suggests that either multilinguals are disadvantaged when tested using the 12-Item SA-WASI Vocabulary Subtest, or that two instruments measure different constructs.

Regarding the original version of the MVT, as administered in Study 1a, the analysis detected a significant, positive but small correlation between scores on that instrument and those on the 12-Item SA-WASI Vocabulary Subtest, $r(28) = .38$, $p = .038$. However, there was no significant correlation between scores on the MVT and those on the APM, $r(28) = .003$, $p = .988$. Even when restricting this sample to those reporting English as their dominant language, the analyses detected no significant correlations. This set of results highlighted the need for a further cycle of revisions to the MVT and a re-evaulation of the use of the APM as a criterion measure.

Linguistic Factors Predicting Test Performance

Table 5 summarizes the results of a set of simple linear regression analyses of performance on the outcome measures in Study 1a. It shows that, whereas performance on the 12-Item SA-WASI Vocabulary Subtest was significantly influenced by participants' linguistic profile, both versions of the MVT are unaffected by the language test-takers acquired first, by their dominant language, or by the number of languages they spoke. To further illustrate these effects of linguistic profile on the 12-Item-SA-WASI Vocabulary subtest but not on the MVT, mean comparisons showed that on the WASI test, (a) those who reported having acquired English as a first language (E1 status) outperformed their peers who reported having acquired any other language first , $t(63) = 3.81$, $p < .001$, Cohen's $d = 0.96$, and (b) those who reported English as their dominant language outperformed those who reported any other language as their most dominant language, $t(63) = 4.21$, $p < .001$, Cohen's $d = 1.10$, but that (c) for both forms of the MVT in Study 1a, analyses detected no such significant between-groups differences, with all ps > .075. Moreover, when employing the same regression models with the bigger Study 1b dataset, none of the three linguistic profile variables significantly predicted MVT performance, R^2s < .04, ps > .103.

Table 5

Study 1a: Summary of Simple Linear Regression Models Predicting Effects of Select Linguistic Factors on Verbal Test Performance (N = 65)

	SA-WASI Vocabulary (N = 65)				p-MVT (n = 35)				MVT (n = 30)			
Variable	R^2	F	β	p	R^2	F	β	p	R^2	F	β	p
E1 status	.19	14.54	.43	< .001***	.18	0.60	.13	.446	.11	3.39	.33	.076
English dominance	.13	17.71	.47	.033*	.01	0.31	.10	.580	.05	1.48	.22	.234
Number of languages	.11	7.48	-.33	.008**	< .01	0.01	-.02	.908	< .01	.00	.00	.987

Notes. *E1 status* and *English dominance* are dummy variables created for the purposes of the regression analyses. SA-WASI Vocabulary = 12-Item South African-adapted Wechsler Abbreviated Scale of Intelligence Vocabulary subtest. (p-)MVT = (pen-and-paper) Multilingual Vocabulary Test.

***p < .001. **p < .01. *p < .05. All p-values are two-tailed.

Test-takers' MVT Experience

The decision to continue with the Study 1a analyses after the initially low reliability values was bolstered by the fact that, beside the psychometric properties, an important factor in interpreting the results of cognitive tests is the test-takers' experience when taking the test (Leong, Park, & Leach, 2013). When asked about their testing, all Study 1a participants, apart from one, indicated that they preferred the MVT (regardless of administration format) over the 12-Item SA-WASI Vocabulary Subtest, for the same reasons that motivated the development of the instrument: They enjoyed having the option to respond in whatever language they felt they knew a given word best, as they thought it better represented their actual knowledge. Moreover, participants stated that being able to refer to a language other than the one in which they had responded boosted their confidence in their responses.

Discussion

Study 1, the project's pilot study, aimed to provide a preliminary psychometric analysis of the MVT and to identify factors influencing performance on the instrument. I assessed the instrument's criterion-related validity as an IQ screening tool, using the 12-Item-SA WASI Vocabulary subtest and the APM as criterion measures, and reported the internal consistency of both paper-and-pencil and digital versions of the MVT. The construct validity and internal consistency values observed in Study 1a were too low to recommend the use of the measure without changes. Data from the administration of a modified MVT in Study 1b, however, showed an increase in Cronbach's α by .36. This change can be attributed, at least partially, to the changes made to the MVT based on the results and observations from Study 1a.

Given that Study 1 was a pilot study not only for the MVT, but for this form of inherently multilingual assessment, in general, and given that further examination of the instrument was encouraged by the positive feedback I received from participants in short, open-ended interviews

after the test sessions, Study 2 continued the iterative cycle of analysis and revision. Moreover, even though these initial results do not allow for the MVT to be considered equivalent to the 12-Item SA-WASI Vocabulary subtest, the negligible influence of linguistic profile (comprising factors with huge variation in South Africa) on MVT performance, especially when compared to the substantial influence of that profile on the WASI test, encouraged a continuation of the project. Therefore, the data obtained in Study 1 was used to improve the MVT for further evaluation and analysis in Study 2.

Finally, in preparation for Study 2, I evaluated the criterion measures used here. The 12-Item SA-WASI Vocabulary subtests proved useful (given its verbal nature and given the fact that it served as a model for the MVT). The APM, on the other hand, despite its supposed culture-free nature and its successful use among African students (Rushton et al., 2004), produced significantly lower results for the current sample. Additionally, using an entirely non-verbal measure of *g* as a criterion measure was perhaps too ambitious given that the project sought to develop a verbal-based screening tool. Consequently, Study 2 addressed this shortcoming by including more appropriate criterion measures.

CHAPTER 4:

STUDY 2—PRELIMINARY RELIABILITY AND VALIDITY ANALYSIS OF A REVISED VERSION OF THE MVT

Study 2 constitutes a logical continuation of Study 1; it aimed to revise the MVT and to continue its development process toward becoming a linguistically fair screening tool for overall cognitive functioning. Given South Africa's great need for a quick and easy-to-administer IQ screening tool, I opted to exclusively focus on the digital administration format, presented on an iPad. This is advantageous, given its self-scored nature, as well as its cost- and time-efficient administration.

The study used an improved design and more appropriate criterion measures compared to Study 1. The major aims of the study were to (a) describe the rationale for, and process of, revising the MVT based on the results obtained in Study 1, and to (b) offer a psychometric analysis of the revised instrument. In doing so, this study continued to shape the MVT toward becoming a reliable and valid multilingual IQ screening tool.

Methods

Design and Setting

This study featured an intra-individual correlational design that sought to establish the revised MVT's construct validity. Analyses correlated participants' scores on the MVT with their scores on three criterion measures. All face-to-face testing took place in the UCT Department of Psychology's ACSENT laboratory. All online components were hosted on the SurveyMonkey platform (www.surveymonkey.com).

Participants

Recruitment. Using convenience sampling, I invited potential participants to sign up for this study in an email sent via the UCT Department of Psychology's Student Research Participation Programme (SRPP). The email contained a link to the SRPP sign-up page, which featured an overview of available time slots from which potential participants could choose. A sign-up confirmation email provided further details (see Procedure section below).

Eligibility criteria. Participants were required to be either monolingual English-speakers or multilingual individuals reporting Afrikaans/English or isiXhosa/English as their home languages. Given the age range of the Wechsler IQ scales' reference group, participants had to be aged between 18 and 34 years (Wechsler, 2008). Individuals with current or past neurological, psychiatric, or psychological disorders, as well as those taking medication for chronic illness, were not allowed to participate, as these factors might affect cognitive performance (Hebben & Milberg, 2009; Lezak et al., 2012).

Final sample. A total of 131 individuals signed up to participate. The first 18 comprised the sample for a pilot study that sought to confirm that the order of items on the revised MVT was acceptable and suitable for subsequent purpose. Hence, 113 individuals comprised the initial sample for the main study. Figure 5 shows the process of participant attrition from that number, explaining the composition of the final sample ($N = 101$) whose data were analyzed. This final sample comprised 77 women and 24 men aged between 18 and 31 years ($M = 19.53$, 95% CI [19.15, 19.92], $SD = 1.97$). They had completed between 10 and 18 years of education ($M = 13.20$, 95% CI [12.92, 13.47], $SD = 1.39$) and were studying toward a Humanities undergraduate degree with a major in Psychology at UCT.

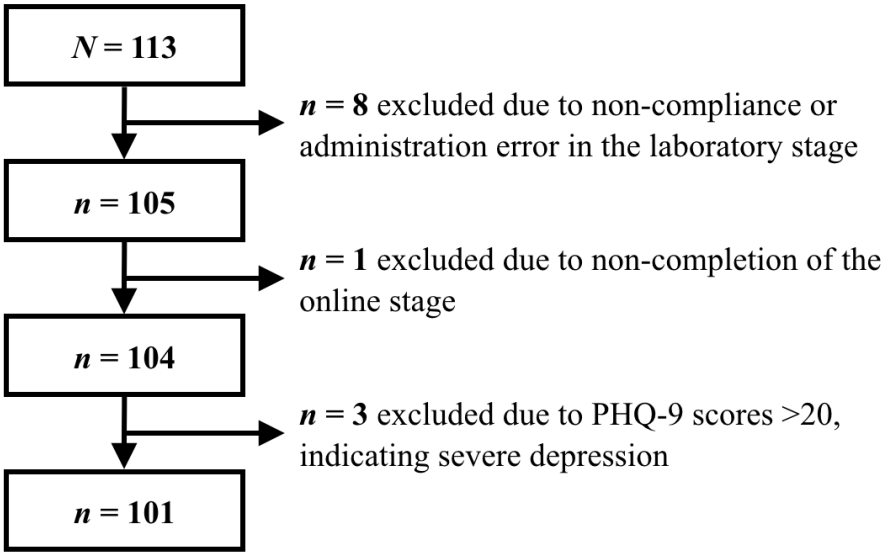


Figure 5. Participant attrition chart for Study 2.

A post-hoc power analysis using G*Power (Faul et al., 2009) with $N = 101$, $\alpha = .05$, and an estimated effect size of $r = .45$ (corresponding to a coefficient of determination of $r^2 = .20$) computed an achieved power of .99 for bivariate correlations and means comparisons using *t*-tests. The computed power only dropped below .90 for effect sizes lower than $r = .32$ ($r^2 = .10$). The statistical power for regression analyses with the above criteria, number of predictors = 4, and an estimated effect size of Cohen's $f = 0.25$ (corresponding to $R^2 = .20$) was .99 and only dropped below .90 for effect sizes of $f < 0.16$ ($R^2 < .14$).

Measures

In addition to the measures described below, this study used the sociodemographic questionnaire, Adapted LEAP-Q, and 12-Item SA-WASI Vocabulary subtest described in Study 1 (Chapter 3). Further, I decided to replace the APM as a criterion measure for two main reasons: the APM seemed to have been too challenging for my sampling frame and it lacked a verbal component. As noted above, I used new criterion measures to replace the APM. These measures (a) are quicker to administer, (b) contain both a verbal and a non-verbal component, and (c)

promised to be more appropriate to the population under investigation. Unless stated otherwise, the language of instruction and administration for the measures was English.

9-Item Patient Health Questionnaire (PHQ-9). The Patient Health Questionnaire is a self-administered mental health screening tool based on the Primary Care Evaluation of Mental Disorders, a diagnostic measure used to assess for the presence of common mental health disorders (Spitzer, Kroenke, Williams, Patient Health Questionnaire Primary Care Study Group, 1999; Spitzer et al., 1994). Its 9-item depression module can be administered as a standalone depression screening (hence, PHQ-9). This instrument scores test-takers' responses to each of the DSM-V's nine depression criteria on a scale from 0 (*not at all*) to 3 (*nearly every day*), with the intermediate options of 1 (*several days*) and 2 (*more than half the days*). Scoring 20 or higher indicates a likelihood ratio of being diagnosed with major depression in a subsequent clinical interview at 36.8 (Kroenke, Spitzer, & Williams, 2001).

The original English, as well as translated versions of the PHQ-9, have been validated and successfully used in different racially and ethnically diverse settings (see, e.g., Huang, Chung, Kroenke, Delucchi, & Spitzer, 2006), including African contexts (Adewuya, Ola, & Afolabi, 2006; Bhana, Rathod, Selohilwe, Kathree, & Petersen, 2015). Cholera et al.'s (2014) recent South African validation study found a 75% post-test probability of being diagnosed with severe depression when scoring above 20. Hence, the current study employed a cut-off score of 20 when screening participants for severe depression.

Controlled Oral Word Association Test (COWAT). This task assesses both phonemic and semantic verbal fluency by measuring spontaneous oral word production, in response to either a letter or a category cue, within 1 minute. For the phonemic component, I used the letters SBL/IBL (S for Afrikaans and English, I for isiXhosa). This letter set is the result of a careful study drawing on dictionary research and language expert consultation with the aim of creating

an equivalent letter set for Afrikaans, English, and isiXhosa that allows the comparison of results across these three languages (see Ferrett et al., 2014 for a description of the letter selection process). This was complemented by the semantic component, which requires test-takers to produce as many unique items belonging to a certain semantic category within the one-minute time limit. Given its superior psychometric properties, I chose to use the category of animals (see, e.g., Portocarrero et al., 2007; Strauss et al., 2006).

Kaufman Brief Intelligence Test-Second Edition (KBIT-2). This is the first new criterion measure. The KBIT-2 (A. S. Kaufman & Kaufman, 2004) is a brief (approximately 25 minutes administration time) intelligence screening measure suitable for use with individuals aged 4 to 90 years. The instrument measures both fluid and crystallised intelligence. Its major outcome variable, the KBIT-2 IQ Composite score, is an overall IQ score that results from a combination of scores on the Verbal Knowledge and Riddles subtests (both measures of verbal IQ (VIQ)/crystallized intelligence), and the Matrices subtest (a measure of non-verbal IQ (NVIQ)/fluid intelligence). These subtests are closely aligned to their counterparts in the Wechsler intelligence scales. Subscale scores (VIQ and NVIQ) are not normally interpreted separately, but they are useful in the kinds of correlation analyses reported here. Psychometric studies suggest the KBIT-2 is highly reliable (internal consistency = .93), and that it has excellent construct validity, correlating at .90 and .89 with WASI FSIQ and WAIS-III FSIQ, respectively (Bain & Jaspers, 2010; A. S. Kaufman & Kaufman, 2004).

Shipley-2. This second new criterion measure is a revised and re-standardized version of the Shipley Institute of Living Scale (see Shipley et al., 2009). The measure gives users two options of subtest combinations: Test administrators can administer the Vocabulary subtest, a measure of crystallized intelligence, alongside either the Abstraction or the Block Patterns subtests, both of which tap into fluid intelligence. I chose to administer the Block Patterns

subtest because of its multiple-choice format, which meant that the Shipley-2 was not only an established criterion measure of IQ, but also controlled for the effects of test administration format (both Shipley-2 Vocabulary and Block Patterns subtests are multiple-choice measures, as is the MVT). The resultant Shipley-2 Composite B score provides a standardized IQ score with average composite reliabilities of .92 and test-retest reliabilities between .87 and .94 for adults. Criterion-related validity, measured by correlations between the Shipley-2 Composite B score and the WAIS-III and WASI FSIQ scores is high, at .85 and .72, respectively (Shipley et al., 2009). Including instructions, administration of the Shipley-2 takes a maximum of 25 minutes, as each subtest has a 10-min time limit.

Revised MVT. For Study 2, I revised the MVT based on Study 1b results and changed its medium of administration to an iPad. In response to the low reliability values obtained in Study 1b and following the basic psychometric rule of thumb that reliability increases with scale length (Finchilescu, 2013), I doubled the number of items from 12 to 24. Analogous to the development process outlined in Study 1, and with the help of language experts and teachers, I devised additional items in English, had them translated and back-translated by Afrikaans and isiXhosa language experts. Eventually, I only included 12 additional items that (a) allowed for meaningful and logical translations, (b) were, at face value, more or less equivalent in terms of difficulty, (c) had a similar frequency (see below for frequency assessment procedure), and (d) had a similar number of syllables and were of similar length across the three languages. The administration format and scoring mechanism remained the same.

Given the unavailability of empiric difficulty data for the 12 new items, the order of administration for the 24 items was initially based on the English translation's frequency of occurrence in the 155-billion-entry American English Google Books N-Gram (Davies, 2000). Items were arranged from least to most difficult (i.e., most to least frequent occurrence in the

database). Then, based on item difficulty data (sum of scores awarded per item, from highest to lowest) obtained in the pilot study ($N = 18$), I changed the order of administration for the main study. A complete list of the 24 items and response options in the order of administration across all studies presented in the thesis can be found in Appendix F.

Procedure

The main study proceedings took place in two distinct stages. Participants completed an online survey at any convenient time prior to the in-person test session, which took place at the UCT Department of Psychology's ACSENT laboratory during the time slot chosen by the participant in the course of the sign-up process. The procedures outlined below were approved by the Ethics Review Committee of the Humanities Faculty at UCT (Appendix K)

Online survey stage. Once participants had signed up for a time slot using the link in the recruitment email, they were prompted to complete an online survey, accessible via a link included in the sign-up confirmation email. The survey started with an informed consent document. Participants were only able to proceed to the rest of the survey after they had confirmed that they had read the consent form, and that they agreed to voluntary participation. After the participants had completed the PHQ-9, the sociodemographic questionnaire, and the Adapted LEAP-Q (in that order), the final page of the survey reminded them to attend the laboratory session during their chosen time slot. Completion of the online survey took 10-15 minutes.

Laboratory stage. After welcoming participants to the laboratory, I explained to them the purpose and procedure of the study and highlighted their rights as participants, as stated in the informed consent document. Once participants had signed the informed consent document, I ascertained that they had completed the online survey. If not, they were given the opportunity to do so in the laboratory. Subsequently, research assistants (final-year undergraduate psychology

students) fluent in the participants' home language(s) conducted the verbal fluency tests in English, as well as in either Afrikaans or isiXhosa. Participants then completed the MVT and the three criterion measures in a block-randomized order (see A. S. Kaufman & Kaufman, 2004; Shipley et al., 2009; Wechsler & Zhou, 2011 respectively, for a detailed description of the administration procedures). After the research assistants or I had administered all psychometric measures, we encouraged participants to comment on their testing experience, to compare their MVT testing experience to that of the 12-Item SA-WASI Vocabulary subtest, as well as to provide general feedback on the MVT. This feedback session took the form of a structured interview based on the interview schedule provided in Appendix L. Finally, I gave participants a chance to ask questions, debriefed them, thanked them, and provided them with a debriefing form and an SRPP confirmation slip. This stage of the study lasted approximately 70-90 minutes.

Statistical Analyses

For most statistical analyses, I used SPSS (version 25). For the IRT analyses, I used the *ltm* package (Rizopoulos, 2006) for *R* (R Core Team, 2018), and for the congeneric reliability estimates, I used the *jamovi* software (version 0.9; jamovi project, 2018).

As in Study 1, all assumptions underlying the various types of inferential analyses were met, unless stated otherwise. I set α at .05 for all decisions regarding statistical significance; correlations below .40, between .40 and .70, and above .70 were considered low, moderate, and high, respectively (Lachenicht, 2013). Effect sizes below .10 were considered very small, while those around .20, .50, and .80 were considered small, medium, and large, respectively (Cohen, 1988).

Preliminary analysis. The steps taken in the analyses of Study 2 data were similar to those taken in Study 1. Hence, a preliminary descriptive analysis provided an overview of the sample's key sociodemographic and linguistic characteristics and basic performance data on the

outcome measures. Between-sex differences were assessed using independent-samples t-tests and chi-square tests of contingency where appropriate. Unlike in Study 1, however, I also analyzed the sample characteristics along linguistic differences, by comparing E1- to E2-speakers and English-dominant to non-English-dominant participants.

Psychometric analysis. Drawing on the principles of both Classical Test Theory and Item Response Theory, I went beyond the scope of the pilot psychometric analyses presented in Study 1. Specifically, beyond calculating item difficulty scores and item-total correlations, I computed IRFs for each item, and then analysed them on a descriptive level using IRCCCs. Subsequently, for test-level analyses, I calculated the split-half reliability, using the Spearman-Brown correction, and the internal consistency of the revised MVT. Here, in addition to Cronbach's α, I calculated McDonald's ω as an estimate of internal consistency, given its less stringent assumptions regarding tau-equivalence, unidimensionality, and normality of the measured construct—some of the most well-known shortfalls of coefficient α (Cortina, 1993; Lord & Novick, 1968; Yang & Green, 2011). Moreover, to complete the picture, I took into consideration the instrument's TCC and TIF.

To assess the MVT's criterion-related validity, I conducted bivariate correlation analyses using Pearson's *r*. Criterion-related validity, here, constituted associations between MVT scores and performance on each of the criterion measures (the 12-Item SA-WASI Vocabulary subtest, KBIT-2, and Shipley-2) separately. For the latter two criterion measures, I assessed correlations of MVT scores with their VIQ, NVIQ, and FSIQ/Composite scores.

Regression modeling. Guided by Study 1 results, I assessed linguistic factors predicting performance on the verbal outcome measures in order to (a) confirm factors suggested in the simple linear regression analyses presented in Studies 1a and 1b, as well as (b) probe additional factors that exert an influence onto MVT performance.

Results

Sample Characteristics

All participants were UCT students enrolled for at least one undergraduate psychology course. The modal participant was an 18-year-old woman who self-identified as Coloured, reported English as her dominant language, and was registered as a first-year-student. Table 6 summarizes the sample's key sociodemographic and linguistic characteristics; it shows that analyses detected no significant between-sex differences with regard to any of those characteristics. However, even though the assumption of homogeneity of variances was upheld for the continuous variables, that of a normal distribution was violated for number of languages (most likely due to the relatively small sample size). Hence, this set of analyses should be treated with caution.

Tables M1 and M2 (see Appendix M) show the characteristics of subsamples, based on linguistic criteria. Specifically, I compared E1- and E2-speakers as well as English-dominant and non-English-dominant individuals. Not surprisingly, those who had reported English as their first or dominant language (98.46% of E1-speakers reported English as their dominant language) spoke a significantly lower number of languages compared to their respective non-English counterparts, with $t(100) = 5.11$, $p < .001$, $d = 1.06$, and with $t(100) = 3.35$, $p = .001$, $d = 0.85$, respectively.

Table 6
Study 2: Sample's Sociodemographic Characteristics (N = 101)

Variable	Total ($N = 101$)	Women ($n = 77$)	Men ($n = 24$)	t / χ^2	p	ESE
Age (years)	19.53 (1.97)	19.38 (1.76)	20.04 (2.49)	1.45	.149	0.34
Years of Education Completed	12.73 (2.64)	12.47 (2.83)	13.58 (1.67)	1.83	.070	0.43
Number of Languages Spoken	2.53 (0.94)	2.52 (0.95)	2.58 (0.93)	0.29	.774	0.07
Race				0.94	.815	.10
Black	34 (34.65)	25 (32.47)	9 (37.50)			
Coloured	46 (45.54)	35 (45.45)	11 (45.83)			

White	12 (11.88)	9 (11.69)	3 (12.50)			
Other/Not declared	9 (8.91)	8 (10.39)	1 (4.17)			
Dominant Language				1.87	.867	.14
Afrikaans	1 (0.01)	1 (1.30)	---			
English	82 (81.12)	62 (80.52)	20 (83.33)			
isiXhosa	11 (10.89)	9 (11.69)	2 (8.33)			
Other	7 (6.93)	5 (6.49)	2 (8.33)			
Language Acquired First				8.41	.394	.29
Afrikaans	6 (5.94)	5 (6.49)	1 (4.17)			
English	65 (64.36)	53 (68.83)	13 (54.17)			
isiXhosa	18 (17.82)	12 (15.58)	6 (25.00)			
Other	11 (10.89)	7 (9.09)	4 (16.67)			

Notes. For the continues variables (*Age, Years of Education Completed*, *Number of Languages Spoken*), means are presented with standard deviations in parentheses. For the remaining (categorical) variables, frequencies are given with percentages in parentheses. Between-group differences were assessed using independent-samples *t*-tests for the continuous variables and Fisher's exact tests for the categorical variables (as some of the expected cell frequencies were smaller than 5). ESE = effect size estimate (Cohen's *d* for continuous variables and Cramer's *V* for categorical variables). If percentages do not add up to 100%, it is due to rounding.

Performance on Cognitive Measures

Table 7 summarizes performance on the various outcome measures and presents results of between-sex comparisons (see Appendix N for performance on the outcome measures by sex). Analyses detected no significant differences between average male and female performance, except on the 12-Item SA-WASI Vocabulary subtest where men scored significantly higher than women, with $t(100) = 2.44$, $p = .016$, $d = 0.57$. Of further note here is that, for all of the individual outcomes on both of the standardized IQ measures (i.e., KBIT-2 and Shipley-2 VIQ, NVIQ, and Composite IQ), the sample's mean scores fell within the range conventionally labelled as average (A. S. Kaufman & Kaufman, 2004; Shipley et al., 2009).

Table 7
Study 2: Performance on the Outcome Measures (N = 101)

				95% CI (Means)		Sex differences		
Measure	*M*	*SD*	*SE*	*LL*	*UL*	*t*	*p*	ESE
12-Item-SA-WASI Vocabulary Subtest	10.73	2.59	0.26	10.22	11.24	2.44	.016*	0.57
MVT	34.22	3.62	0.36	33.50	34.93	0.37	.711	0.09
KBIT-2								
Verbal IQ	94.67	9.84	0.98	92.73	96.61	0.99	.322	0.23
Non-Verbal IQ	97.79	11.92	1.19	95.44	110.14	0.16	.876	0.04
Composite Score	95.90	10.14	1.01	93.90	97.90	0.42	.674	0.10
Shipley-2								
Verbal IQ	101.53	9.70	0.97	99.62	103.45	1.09	.278	0.25
Non-Verbal IQ	100.28	12.19	1.21	97.87	102.68	1.34	.185	0.31
Composite B	101.18	10.86	1.08	99.03	103.32	1.58	.118	0.37

Notes. Mean scores are presented, with standard deviations in parentheses. SA-WASI = 12-Item-SA-WASI Vocabulary subtest; MVT = digital Multilingual Vocabulary Test; KBIT-2 = Kaufman Brief Intelligence Scale-Second Edition; ESE = effect size estimate (in this case, Cohen's *d*). 12-Item SA-WASI Vocabulary subtest and MVT scores are raw scores, KBIT-2 and Shipley-2 are standard scores. Sex differences were assessed using independent-samples *t*-tests. *$p < .05$, two-tailed.

Guided by the results of the regression models presented Study 1, which showed a significant influence of E1 status and English dominance on criterion measure performance, and encouraged by the aforementioned difference in sample characteristics between linguistic subsamples, I analyzed test performance for each linguistic subsample separately (see Tables 8 and 9). As might be expected, participants who had reported a language other than English as their first or as their dominant language scored significantly more poorly on most criterion measures than those who had reported English as their first or as their most dominant language. Specifically, they scored lower on the 12-Item SA-WASI Vocabulary subtest, on the Shipley-2 VIQ and NVIQ indices, as well as on the KBIT-2 VIQ index. As a consequence, their Shipley-2 and KBIT-2 Composite scores were significantly lower than those of their English-speaking/-dominant counterparts. The only outcome variable for which analyses no significant between-group difference for first and dominant language were the KBIT-2 NVIQ index and the MVT score.

Table 8
Study 2: Performance on the Outcome Measures by L1 (N = 101)

Measure	Total (*N* = 101)					E1-speakers (*n* = 65)					E2-speakers (*n* = 36)					Means comparison		
				95% CI					95% CI					95% CI				
	M	*SD*	*SE*	*LL*	*UL*	*M*	*SD*	*SE*	*LL*	*UL*	*M*	SD	SE	LL	UL	*t*	*p*	*ESE*
SA-WASI	10.73	2.59	0.26	10.20	11.24	11.31	2.33	0.30	10.73	11.89	9.69	2.75	0.46	8.77	10.62	3.12	.002**	0.57
MVT	34.22	3.62	0.36	33.50	34.93	34.62	3.80	0.47	33.67	35.56	33.50	3.20	0.53	32.42	34.58	1.49	.139	0.09
KBIT-2																		
Verbal IQ	94.67	9.84	0.98	92.73	96.61	98.20	8.14	1.01	96.18	100.22	88.31	9.51	1.59	85.09	91.52	5.51	<.001***	0.23
Non-Verbal IQ	97.79	11.92	1.19	95.44	100.14	98.69	11.75	1.46	95.78	101.60	96.17	12.20	2.03	92.04	100.30	1.02	.310	0.04
Composite	95.90	10.14	1.01	93.90	97.90	98.48	9.63	1.19	96.09	100.86	91.25	9.47	1.58	88.04	94.46	3.63	<.001***	0.10
Shipley-2																		
Verbal IQ	101.53	9.70	0.97	99.62	103.45	104.22	8.77	1.09	102.04	106.39	96.69	9.52	1.59	93.47	99.92	4.00	<.001***	0.25
Non-Verbal IQ	100.28	12.19	1.21	97.87	102.68	103.22	10.47	1.30	100.62	105.81	94.97	13.37	2.23	90.45	99.50	3.43	.001**	0.31
Composite B	101.18	10.86	1.08	99.03	103.32	104.57	9.93	1.23	102.11	107.03	95.06	9.84	1.64	91.73	98.38	4.63	<.001***	0.37

Notes. Mean scores are presented with standard deviations in parentheses. SA-WASI and MVT scores are raw scores, KBIT-2 and Shipley-2 scores are standard scores. Group differences were assessed using independent-samples *t*-tests. L1 = first language. ESE = effect size estimate (in this case, Cohen's *d*). MVT = digital Multilingual Vocabulary Test. SA-WASI = 12-Item South African-adapted Wechsler Abbreviated Scale of Intelligence Vocabulary Subtest. KBIT-2 = Kaufman Brief Intelligence Scale (Second Edition).
**p* < .05. ***p* < .01. ****p* <.001. All *p*-values are two-tailed.

Table 9
Study 2: Performance on the Outcome Measures by Dominant Language (N = 101)

Measure	Total (*N* = 101)					English-dominant (*n* = 82)					Non-English-dominant (*n* = 19)					Means comparison		
				95% CI					95% CI					95% CI				
	M	*SD*	*SE*	*LL*	*UL*	*M*	*SD*	*SE*	*LL*	*UL*	*M*	SD	SE	LL	UL	*t*	*p*	*ESE*
SA-WASI	10.73	2.59	0.26	10.20	11.24	11.32	2.29	0.25	10.81	11.82	8.21	2.32	0.53	7.09	9.33	5.31	<.001***	0.57
MVT	34.22	3.62	0.36	33.50	34.93	34.33	3.76	0.42	33.50	35.16	33.74	2.98	0.68	32.30	35.17	0.64	.523	0.09
KBIT-2																		
Verbal IQ	94.67	9.84	0.98	92.73	96.61	97.41	7.81	0.86	95.70	99.13	82.84	8.98	2.06	78.51	87.17	7.12	<.001***	0.23
Non-Verbal IQ	97.79	11.92	1.19	95.44	100.14	98.16	11.66	1.29	95.60	100.72	96.21	13.13	3.03	89.85	102.57	0.64	.524	0.04
Composite	95.90	10.14	1.01	93.90	97.90	97.71	9.32	1.03	95.66	99.75	88.11	10.07	2.31	83.25	92.96	3.99	<.001***	0.10
Shipley-2																		
Verbal IQ	101.53	9.70	0.97	99.62	103.45	103.67	8.80	0.97	101.74	105.60	92.32	7.99	1.83	88.47	96.17	5.15	<.001***	0.25
Non-Verbal IQ	100.28	12.19	1.21	97.87	102.68	101.43	11.73	1.30	98.85	104.01	95.32	13.18	3.02	88.96	101.67	2.00	.048*	0.31
Composite B	101.18	10.86	1.08	99.03	103.32	103.21	10.50	1.16	100.90	105.51	92.42	7.70	1.77	88.71	96.13	4.22	<.001***	0.37

Notes. Mean scores are presented with standard deviations in parentheses. SA-WASI and MVT scores are raw scores, KBIT-2 and Shipley-2 scores are standard scores. Group differences were assessed using independent-samples *t*-tests. ESE = effect size estimate (in this case, Cohen's *d*). MVT = digital Multilingual Vocabulary Test. SA-WASI = 12-Item South African-adapted Wechsler Abbreviated Scale of Intelligence Vocabulary Subtest. KBIT-2 = Kaufman Brief Intelligence Scale (Second Edition).
**p* < .05. ***p* < .01. ****p* <.001. All *p*-values are two-tailed.

A closer look at the MVT data reveals two additional insights. First, the score distribution (shown in Figure 6) approaches a normal distribution, with most scores clustered around the mean of 34.22, and 95% of scores falling between 29.14 and 39.30. Second, even had I implemented a discontinue rule after four consecutive scores of 0 (as is the case in the WASI Vocabulary subtest), it would not have been applicable to any test-taker: The longest streak of 0-mark responses (or skipped items) was three. Hence, all scores are the sum-total of all 24 MVT items.

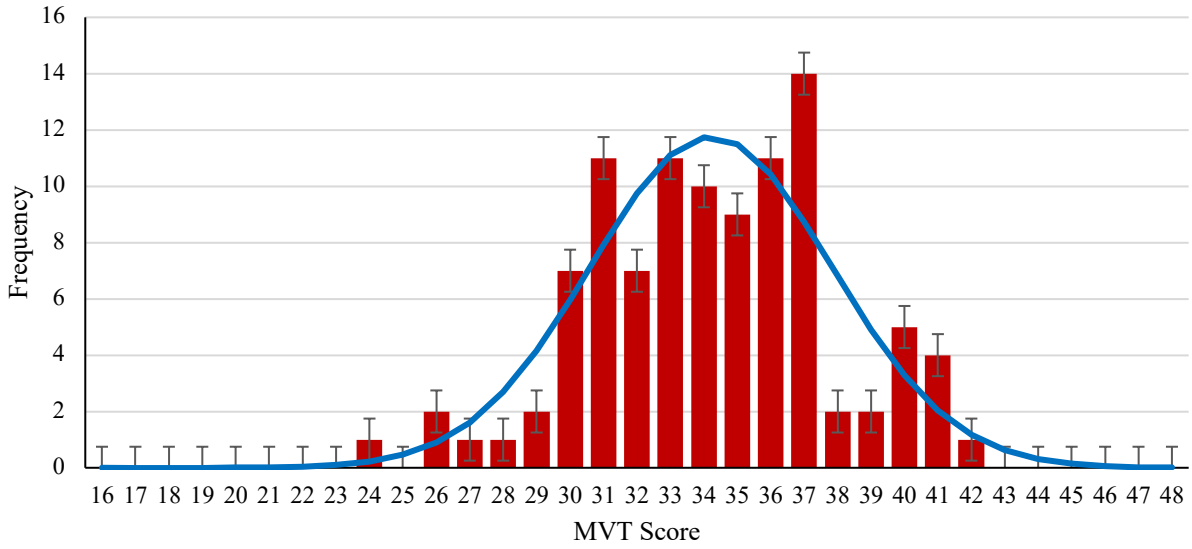


Figure 6. Distribution of MVT scores in Study 2 (N = 101), with normal curve. The entire range of scores is depicted.

MVT Psychometric Analysis

The psychometric evaluation of the MVT forms the central part of this study. Unlike most psychometric analyses, I start by providing the results of an item analysis prior to reporting on the scale's reliability, internal consistency, and criterion-related validity. The reason underlying the current approach is that having more information on each item's performance allowed me to

create a tentative abbreviated version of the instrument, which I could then analyze separately with regard to reliability and validity.

Item analysis. The item analysis presented here combines the interpretation of both CTT and IRT output, in order to provide a comprehensive evaluation of the MVT items.

Item difficulty and item-total correlations. One way of assessing item performance is a CTT item difficulty analysis. However, to conduct such an analysis in the context of the polytomous nature of the measure meant first creating an artificial dichotomous grouping of response options. The most comprehensive approach to dichotomizing was to adopt two solutions: (a) regarding 2-mark responses as correct, and all other responses as incorrect (i.e., taking a narrower definition of correctness), and (b) regarding both 1- and 2-mark responses as correct, and all other responses as incorrect (i.e., taking a wider definition of correctness). Table O-1 (Appendix O) shows the item difficulty, using both the narrower and wider definitions, and item-total correlation statistics for each of the 24 items. Figure 7 is a graphic representation of the item difficulty analysis. The three curves show, for each item, (a) the number of 2-mark-responses, (b) the number of 1-mark-responses, and (c) the total count of both 1- and 2-mark-responses. Additionally, the bars show the total sum of scores (across all participants) awarded for each item.

Overall, item difficulty as approximated using the 1-or-2-mark response rate remains fairly constant across the first two-thirds of the measure and then slightly decreases toward the end, apart from items 23 (*picture* | *prent* | *umfanekiso*) and 24 (*atoll* | *atoll* | *isiqhiti esisangqa*). However, looking at the bars showing the sum of scores awarded per item provides a more useful source of information, because they show the finer nuances with regard to item differences, which are glossed over in the cumulative 1- or 2-mark-response curve. Moreover, although the total-count curve suggests little item discriminability, this is a product of what is, perhaps, an

overly-wide definition of correctness. Indeed, when one only compares the other two curves, it is clear to see that the items do discriminate between those test-takers who selected the *most correct* response (2 marks) and those who selected one of the two *partly correct* (1 mark) responses.

Finally, item 15 stands out, as only one participant selected the 2-mark-response (in Afrikaans). This pattern of data is the result of a software error that caused the item to be displayed without the English 2-mark-response. Hence, the results for item 15 (*ambulance* | *ambulans* | *i-ambulesi*) cannot be regarded as accurate.

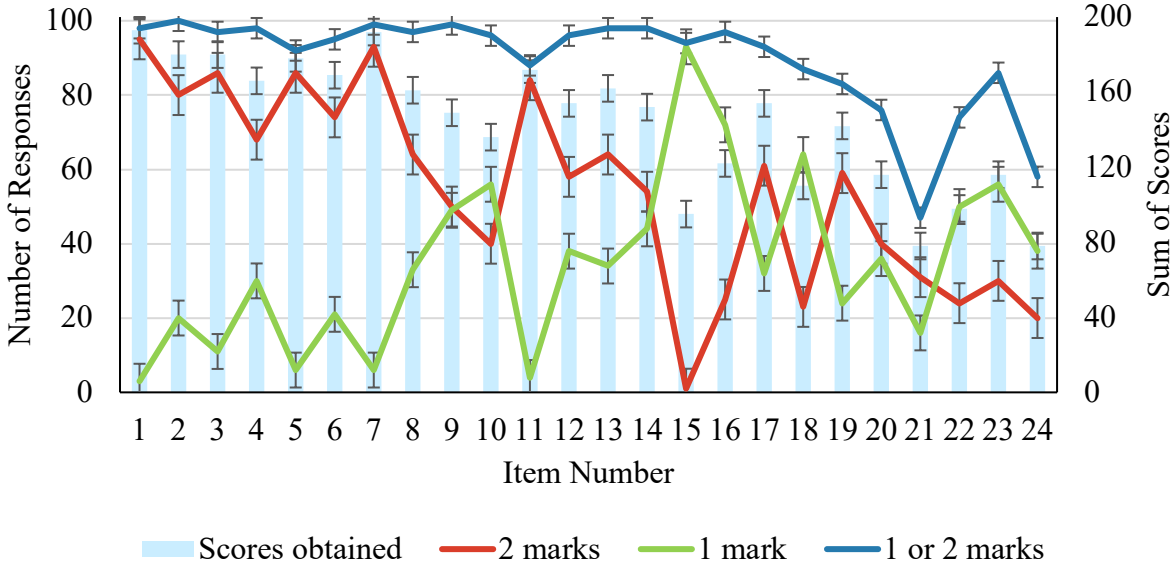


Figure 7. Relative item difficulty curves and totals of scores awarded per item of the 24 MVT items, arranged in the original order of administration ($N = 101$).

IRT item analysis. Prior to analyzing the IRCCCs, I used a likelihood ratio table to determine the best IRT model to fit the data. The unconstrained GRM proved most appropriate, $\chi^2(23) = 56.97$, $p < .001$. The complete set of IRCCCs is shown in Appendix P and should be read in conjunction with the sub-section below.

IRCCC analysis. First, for illustrative purposes, consider item 22 (Figure 8), which most closely resembles the ideal set of IRCCCs: (a) the probability of selecting the 0-mark response is

highest for test-takers with the lowest ability levels and rapidly decreases with increasing ability, (b) the probability of selecting the 1-mark response peaks at the average ability level (0) with lower probabilities for high and low ability levels, and (c) the curve indicating the probability of selecting the 2-mark-response only increases above chance levels for the highest ability levels.

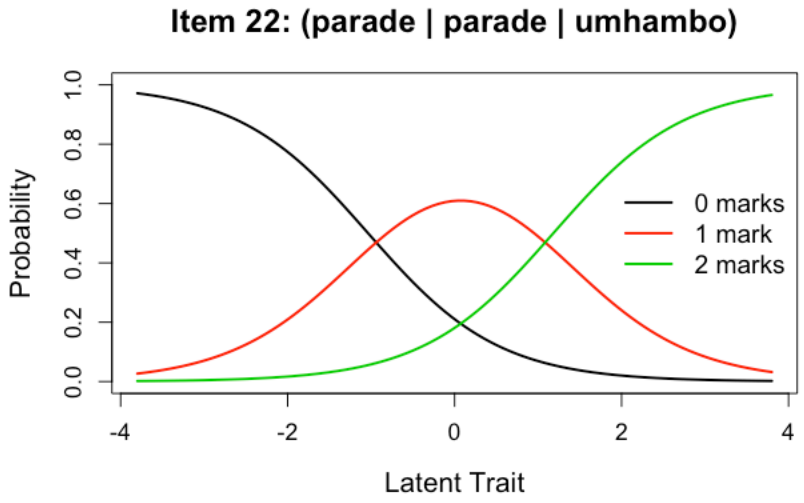


Figure 8. IRCCC of MVT item 22 in Study 2 ($N = 101$).

Along with the IRCCCs for item 22, those for items 4, 6, 7, 9, 10, 12, 16, 18, 20, and 21 also resemble the desired discrimination pattern (even though not necessarily symmetrically over an ability level of 0, but shifted either higher or lower on the *x*-axis). A screening of the remaining IRCCCs allowed discernment of two other broad types of items: (a) those with poor discriminative ability, in that the probability of selecting 0-, 1-, and 2-mark responses did not vary with ability level (items 1, 2, 3, 5, 11, 13, 17, and 19) and (b) those with a reverse response tendency, which—counterintuitively—saw the probability of selecting a 2-mark response decrease with ability level while the probability of selecting 0- and/or 1-mark-responses increased with ability level (items 8, 14, 15, 23, and 24). The latter is not necessarily an indicator of a poor word selection for this item, but primarily indicates an issue with the available response options or with their relative weighting.

Note, however, that the sections of the IRCCCs shown in Appendix P are for the range of *z*-scores (ability) from -4 to +4, which is generally recommended as the most practical and applicable (Templin, 2007). That ability range only covers 65.83% of the information the MVT can provide. Increasing the range, especially toward the lower end, increases the level of information obtained. For example, using a range of ability *z*-scores from -10 to +4 displays 96.04% of the measure's information content. The option to increase the *z*-score range was taken in Study 3.

Item information analysis. Figure 9 shows the IICs of all 24 MVT items. Recall that IICs show the amount of information an item provides for test-takers at a certain ability level. This is IRT's closest approximation of reliability, and hence provides valuable information about which items are useful to retain.

One can see that items 10, 21, and 22 provide the most information in the average ability range, and that items 1 and 15 provide relatively more information at the lower end of the ability spectrum. All other items provide very little information across all ability levels, with their curves remaining below the 0.1 level. Again, this is taken as a thought-provoking, yet non-final result, awaiting confirmation or rejection from the Study 3, which used a bigger and more heterogeneous sample.

Preliminary abbreviated version of the MVT. Following close investigation of the evidence provided in the item analyses, I proposed a tentative abbreviated version of the revised MVT. This version excludes all items with item-total correlations less than .20 (i.e., it includes 13 items, namely items 1, 3, 6, 7, 9, 10, 12, 13, 16, 19, 20, 21, and 22). In the following sections, I provide the TIC, reliability and criterion validity data for both the full 24-item MVT and the abbreviated 13-item version.

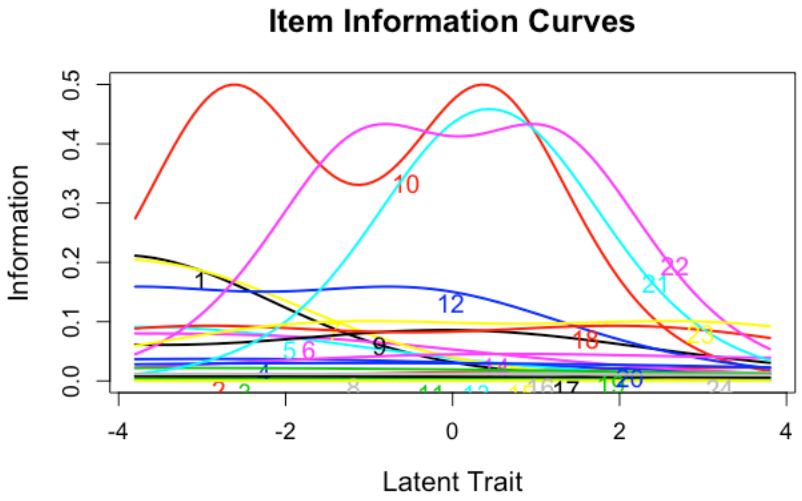


Figure 9. IICs for MVT items 1-24 in Study 2 ($N = 101$).

Test information. TIFs shows the sum of all IICs; in other words, they show how much information the instrument in its entirety reveals about test-takers at different ability levels. Figure 10 shows the TIF for the 24-item MVT. (Note that the scale of the y-axis has changed compared to Figure 9.) The slope of the curve, with its peak around the average ability level, shows that the test provides the most information for test-takers with an average ability level. Further, the MVT still provides a fair amount of information for those with lower ability levels, indicated by the relatively high position and low gradient between -4 and 0 on the *x*-axis. The discrimination and utility of the MVT does, however, decrease for those with above-average ability levels, as illustrated by the steep downward trajectory of the curve beyond its peak. Figure 11 shows the TIF for the 13-item MVT. Even though the overall trajectory resembles that of the 24-item version (both show a peak of 2.0 for the average ability level and rapidly decrease for higher ability levels), due to the reduced number of items, the amount of information provided in the lower ability range is slightly reduced.

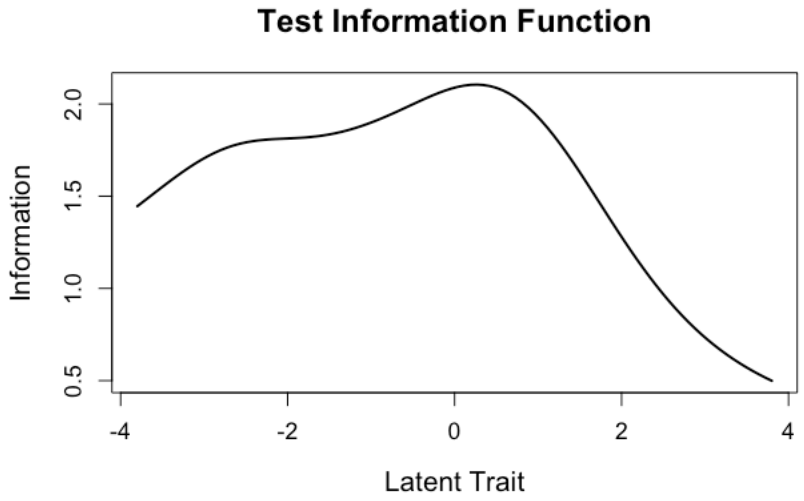


Figure 10. TIF for 24-item MVT in Study 2 (N = 101).

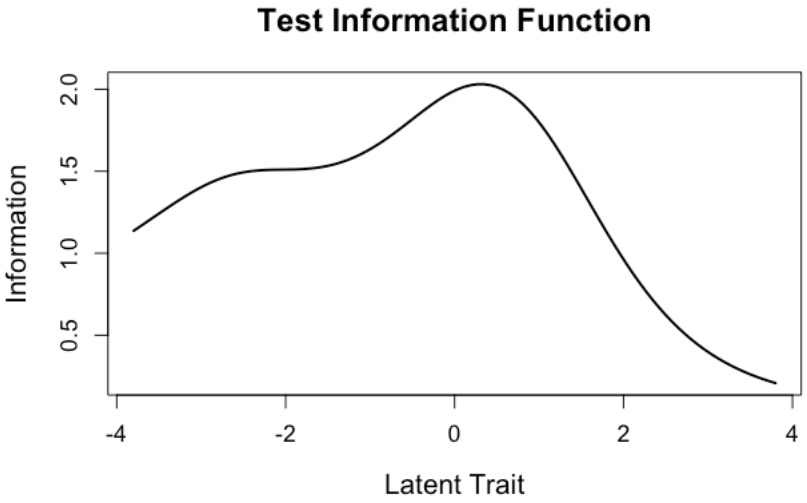


Figure 11. TIF for 13-item MVT in Study 2 (N = 101).

Scale reliability and internal consistency. Now that I have proposed a tentative shortened version of the revised MVT, the next step in the psychometric analysis is the reliability assessment. The study's design allowed for the computation of split-half reliability (including the Spearman-Brown correction) and internal consistency (using coefficients α and ω).

Split-half reliability. The initial split-half reliability coefficient (odd-even split) was $r = .33$. Corrected for the halved scale length using the Spearman-Brown formula, I obtained a coefficient of $r = .50$. This value rose to $r = .60$ when restricting the analysis to the 13 items identified as part of the shortened version. However, due to the possible influence of the way in which the test was split, I proceeded to compute internal consistency values as a more accurate reliability estimate.

Internal consistency. An initial internal consistency analysis of the full scale resulted in a Cronbach's α value of .35. However, given the well-documented shortcomings of only measuring scale reliability using Cronbach's coefficient α, a lower-bound estimate (see, e.g., Cortina, 1993; Lord & Novick, 1968; Yang & Green, 2011), I also computed congeneric reliability, which resulted in a value of McDonald's $\omega = .38$. However, given that the current sample is marginally too small to obtain an accurate congeneric reliability estimate, I interpret the McDonald's ω values obtained here cautiously and addressed this shortcoming in Study 3.

Restricting the internal validity analysis to the 13 items included in the proposed abbreviated version resulted in values of $\alpha = .55$ and $\omega = .54$. Even this moderate internal consistency coefficient suggests the need for another revision of the MVT. Having said that, in Study 3, I repeated the analysis using a bigger sample and, thus, produced a more powerful internal consistency analysis.

Criterion validity. To assess whether the MVT taps into the cognitive domain of verbal IQ and is an adequate screener for general intellectual functioning (i.e., overall IQ), I correlated test-takers' scores on the MVT (both the full 24-item version and the abbreviated, more reliable, 13-item version) with their scores (including subscale scores) on the three criterion measures (12-item SA-WASI Vocabulary subtest, KBIT-2, Shipley-2). To highlight the effect of linguistic variables on the criterion validity analysis, I first conducted the analysis for the entire

psychometric sample and, then, repeated it for language-based subsamples separated based on their L1 and dominant language.

Criterion validity results for the entire sample. Table 10 shows the results of the first criterion validity analysis, using all 101 datasets. The pattern of results is promising in terms of confirming criterion validity for the MVT. First, scores on both the full and abbreviated versions of the instrument correlated positively and strongly with scores on the verbal components of the KBIT-2 and Shipley-2, and with the verbally-based 12-Item SA-WASI Vocabulary subtest. Furthermore, these latter associations were stronger than those with the nonverbal components of the KBIT-2 and Shipley-2 (although, as was expected, they still bore moderately strong relationships to NVIQ scores). Finally, correlations between MVT scores and the overall KBIT-2 and Shipley-2 IQ scores were also of moderate strength. Correlations between the MVT and the Shipley-2 are particularly important, because using the latter as a criterion measure controls for possible effects of the multiple-choice administration format.

Table 10
Study 2: MVT Criterion-validity Analysis (N = 101)

	24-item MVT (ω = .38)		13-item MVT (ω = .54)	
Criterion measure	*r*	*p*	*r*	*p*
SA-WASI	.24	.014*	.27	.006**
KBIT-2				
Verbal IQ	.42	< .001***	.46	< .001***
Non-Verbal IQ	.20	.047*	.20	.041*
Composite	.37	< .001***	.40	< .001***
Shipley-2				
Verbal IQ	.47	< .001***	.52	< .001***
Non-Verbal IQ	.27	.007**	.29	.003**
Composite B	.45	< .001***	.50	< .001***

Notes. Data presented are Pearson's *r* correlation coefficients. MVT = digital Multilingual Vocabulary Test; SA-WASI = 12-Item South African-adapted Wechsler Abbreviated Scale of Intelligence Vocabulary subtest; KBIT-2 = Kaufman Brief Intelligence Scale-Second Edition. The abbreviated MVT includes items with item-total correlations above .20 (i.e., items 1, 3, 6, 7, 9, 10, 12, 13, 16, 19, 20, 21, and 22).
***$p < .001$. **$p < .01$. *$p < .05$. All *p*-values are two-tailed.

Comparing performance on the full version of the MVT with the proposed 13-item version, it appears that, in this context, the latter's criterion validity is slightly higher. However, in both cases the correlation coefficients do not exceed .50, which on the face of it is a less-than-optimal result when attempting to develop a sufficiently valid new instrument. Part of the reason for these relatively low high correlations with criterion measures might be insufficiently high reliability, hence the decision to conduct more powerful reliability analyses in Study 3.

Criterion validity results for the language-based subsamples. In addition to the findings presented above and given the observed significant of mean differences on all criterion measures between (a) E1-speaker and E2-speaker, and (b) those who reported English as their dominant language and those who did not, I repeated the criterion validity using language-based subsamples. Hence, I split the data by Language Acquired First, and then by Dominant Language, and repeated the correlational analyses for each of those split samples. Results of those analyses are presented in Tables 11 and 12.

Table 11
Study 2: MVT Criterion-validity Analysis by First Language (N = 101)

	E1-speakers (n = 65)				E2-speakers (n = 36)			
	24-item MVT (ω = .38)		13-item MVT (ω = .54)		24-item MVT (ω = .38)		13-item MVT (ω = .54)	
Criterion measure	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
SA-WASI	.32	.009**	.40	<.001***	.01	.932	-.12	.479
KBIT-2								
Verbal IQ	.44	<.001***	.54	<.001***	.34	.042*	.21	.218
Non-Verbal IQ	.29	.020*	.27	.028*	-.02	.885	.00	.998
Composite	.42	<.001***	.46	<.001***	.19	.279	.13	.435
Shipley-2								
Verbal IQ	.51	<.001***	.61	<.001***	.35	.035*	.27	.114
Non-Verbal IQ	.30	.017*	.28	.023*	.14	.399	.20	.247
Composite B	.47	<.001***	.52	<.001***	.34	.040*	.35	.035*

Notes. Data presented are Pearson's *r* correlation coefficients. MVT = digital Multilingual Vocabulary Test; SA-WASI = 12-Item South African-adapted Wechsler Abbreviated Scale of Intelligence Vocabulary subtest; KBIT-2 = Kaufman Brief Intelligence Scale-Second Edition. The abbreviated MVT includes items with item-total correlations above .20 (i.e., items 1, 3, 6, 7, 9, 10, 12, 13, 16, 19, 20, 21, and 22).
***p < .001. **p < .01. *p < .05. All p-values are two-tailed.

Table 12
Study 2: MVT Criterion-validity Analysis by Dominant Language (N = 101)

	English-dominant (n = 82)				Non-English-dominant (n = 19)			
	24-item MVT (ω = .38)		13-item MVT (ω = .54)		24-item MVT (ω = .38)		13-item MVT (ω = .54)	
Criterion measure	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
SA-WASI	.30	.006*	.32	.004*	-.07	.770	-.10	.685
KBIT-2								
Verbal IQ	.49	<.001***	.56	<.001***	.40	.092	.21	.377
Non-Verbal IQ	.24	.033*	.22	.044*	.00	.995	.08	.743
Composite	.41	<.001***	.44	<.001***	.20	.401	.18	.460
Shipley-2								
Verbal IQ	.52	<.001***	.57	<.001***	.37	.119	.34	.151
Non-Verbal IQ	.30	.007**	.30	.006**	.10	.684	.19	.433
Composite B	.48	<.001***	.51	<.001***	.34	.155	.46	.047*

Notes. Data presented are Pearson's *r* correlation coefficients. MVT = digital Multilingual Vocabulary Test; SA-WASI = 12-Item South African-adapted Wechsler Abbreviated Scale of Intelligence Vocabulary subtest; KBIT-2 = Kaufman Brief Intelligence Scale-Second Edition. The abbreviated MVT includes items with item-total correlations above .20 (i.e., items 1, 3, 6, 7, 9, 10, 12, 13, 16, 19, 20, 21, and 22).
***p < .001. **p < .01. *p < .05. All p-values are two-tailed

Correlations between scores on the revised MVT (both the 24- and 13-item versions) and those on the criterion measures' verbal and composite indices were statistically significant, positive in direction, and moderate in magnitude for those who reported English as their first or most dominant language. Moreover, the strength of these correlations was greater than those reported above for the entire sample. In contrast, within the subsample of participants who reported a language other than English as their first or most dominant language, correlations between MVT scores and criterion measures were, for the most part, not statistically significant. The only exceptions for non-English-dominant participants was a significant correlation between 13-item MVT performance and the Shipley-2 Composite B score. For E2 speakers, the 24-item version proved more valid (significant correlations between the MVT and KBIT-2 VIQ, Shipley VIQ, and Shipley-2 Composite B scores). Here, scores on the the 13-item version only significantly correlated with Shipley-2 Composite B scores. Although this pattern of data is unsurprising, given that the KBIT-2 and Shipley-2 were normed on E1-status samples, it does highlight the fact of how unfair the assessment process can be when tests developed and standardized in one language are administered to individuals with a different first language.

In contrast to the pattern of subgroup performance on the criterion measures, MVT scores did not differ significantly between E1- and E2-speakers, or between English-dominant and non-English-dominant individuals. Hence, it might be reasonable to conclude that the revised MVT (in both its 13- and 24-item version) is a more linguistically fair assessment tool compared to the criterion measures used in this study. That conclusion must be approached cautiously at this point, however, given that here the sample of participants with a language other than English as a first language or as a dominant language was quite small.

Proposed MVT changes. Given the relatively small sample and its homogeneity when compared to the South African population, I decided to only make minor changes to items 5, 13,

17, and 19 before launching Study 3 (these changes are detailed in the next chapter), and to delay suggestions regarding major item changes until after the analysis of Study 3 data. I did, however, rearrange the 24 items according to difficulty, with that construct defined as the sum of scores awarded per item (lower sum of scores = more difficult items). This empirically-based rearrangement resulted in a smoother item difficulty curve, and hence a refined order of administration in preparation for Study 3. Figure 12 shows the rearranged 24-item MVT (compare to Figure 7, which shows the original administration order).

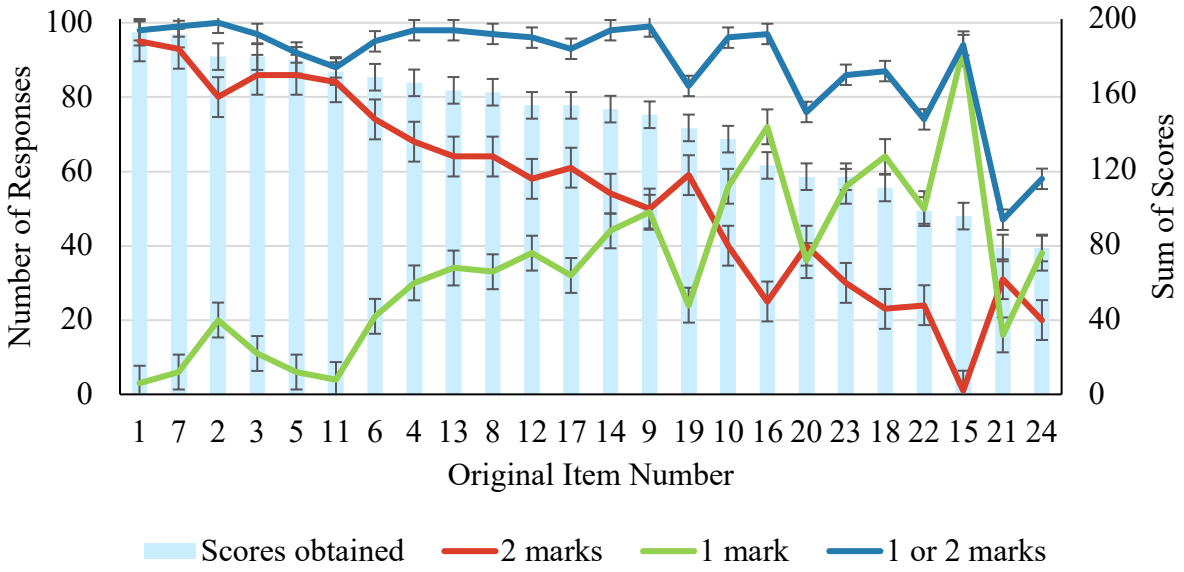


Figure 12. Relative item difficulty curves and sums of scores awarded per item for MVT items 1-24, rearranged according to sum of scores awarded per item ($N = 101$).

Linguistic Factors Predicting Test Performance

Table 13 shows the results of a series of simple linear regressions modeling the influence of different linguistic variables on verbal test performance. Consistent with the results of similar models presented in Study 1, having English as a first language, having English as the dominant language, spending more years in an English-speaking family, and scoring higher on the English semantic fluency test significantly predicted better performance on the 12-Item SA-WASI Vocabulary subtest, the Shipley-2 VIQ index, and the KBIT-2 VIQ index. None of the aforementioned linguistic variables exerted a statistically significant influence on MVT scores, however. The only statistically significant predictor of MVT performance was English phonemic fluency. However, that model accounted for only 4% of the variance in MVT scores, suggesting that, overall, MVT performance is resistant (certainly more resistant than verbally-based SA-WASI, KBIT-2, or Shipley-2 subtests) to the effects of English proficiency and familiarity.

Table 13

Study 2: Summary of Simple Linear Regression Analyses of Linguistic Factors Influencing Verbal Test Performance (N = 101)

	SA-WASI Vocabulary				KBIT-2 VIQ				Shipley-2 Vocabulary				24-item MVT			
Predictor	R^2	*F*	*β*	*p*	R^2	*F*	*β*	*p*	R^2	*F*	*β*	*p*	R^2	*F*	*β*	*p*
English dominance	.22	28.18	.47	<.001***	.34	50.69	.58	<.001***	.21	26.54	.46	<.001***	.00	.411	.06	.523
E1 status	.09	9.76	.30	.002**	.23	30.32	.48	<.001***	.14	16.02	.37	<.001***	.02	2.23	.15	.139
Number of languages	.01	1.18	-.11	.280	.05	4.78	-.22	.031*	.03	2.65	-.16	.107	.00	0.27	-.05	.606
Semantic fluency (English)	.13	14.32	.36	<.001***	.22	27.43	.47	<.001***	.06	6.61	.25	.012*	.03	2.54	.16	.114
Phonemic fluency (English)	.00	0.00	-.01	.945	.00	0.00	-.01	.948	.00	0.44	.07	.508	.04	4.27	.20	.042*
Years spent in:																
English school	.03	3.05	.18	.084	.04	3.44	.19	.067	.02	2.12	.15	.149	.01	0.70	.09	.405
English family	.10	10.42	.32	.002**	.17	19.41	.41	<.001***	.09	9.62	.31	.003**	.02	1.72	.13	.193
Education completed (years)	.02	1.59	.13	.211	.00	0.19	-.04	.661	.00	0.00	.00	.996	.01	0.50	.07	.482

Note. KBIT-2 VIQ and Shipley-2 Vocabulary Subtest scores were entered as standard scores, while 12-Item SA-WASI Vocabulary subtest and MVT scores were entered as raw scores. SA-WASI Vocabulary = 12-Item South African-adapted Wechsler Abbreviated Scale of Intelligence Vocabulary subtest; KBIT-2 = Kaufman Brief Intelligence Scale-Second Edition; MVT = digital Multilingual Vocabulary Test; E1 status = English first-language speaker.

*$p < .05$. **$p < .01$. ***$p < .001$. All *p*-values are two-tailed.

Test-Takers' MVT Experience

The brief semi-structured interview at the end of the testing session yielded a clear picture of the respondents' thoughts on the MVT. Most test-takers evaluated the instrument positively, using descriptions such as 'straightforward', 'comfortable', and 'easy to understand'. Many mentioned that they enjoyed the time-efficient nature of the MVT, as well as the fact that the test was administered on a tablet.

Moreover, although a minority of test-takers stated that they were overwhelmed by the number of options to choose from, most considered the MVT easier than the criterion measures, primarily due to its multiple-choice format. A recurring comment was that recognizing the most appropriate answer is less difficult than describing a word without having any context or cues, as, for example, one is required to do when being administered the 12-Item SA-WASI Vocabulary Subtest. Many explained this preference by referring to the presence of multiple languages, which allowed them to 'compare across languages' and to 'get clues from other languages' when responding to an item.

Many test-takers also spoke about experiencing different confidence levels when completing between the MVT and the 12-Item SA-WASI Vocabulary Subtest. Two factors appeared to be particularly important in influencing these differences in their confidence: First, the fact that they could 'check' their initial response in one language by looking across at a second/third language and, second, the fact that they felt less pressurized while completing the MVT due to the absence of a test administrator. Many interviewees preferred the tablet administration format for, among other reasons, its self-administered nature, which they stated reduced their perceived level of stress during the task. Many stated that, despite the fact that neither test employed time limits, they felt under less time pressure when completing the MVT than when completing the 12-Item SA-WASI Vocabulary subtest.

Regarding the intention to increase linguistic fairness in cognitive testing, participants—including the handful of monolingual English-speakers—had a positive attitude toward the multilingual nature of the measure, finding it to be ‘fair’, ‘inclusive’, and ‘accessible’. Many recognized the great utility and equalizing potential of the MVT in the South African context, remarking that it would allow people to ‘respond in their own language’, especially when English is not their first language. Finally, although some test-takers indicated that the presence of multiple languages within the MVT ‘did not make a difference’, ‘did not matter’, or ‘distracted’ them, nobody mentioned feeling disadvantaged. Indeed, some would have preferred to see even more languages included in the test.

In summary, the most prominent themes were the ease of the task itself, its quick and self-administered nature, a preference for the multiple-choice format, increased confidence, reduced stress levels during the task, and the MVT’s fairness and inclusivity.

Discussion

Looking at the sample’s performance on the cognitive outcome measures, one sees a much more balanced set of results, compared to Study 1 results. KBIT-2 and Shipley-2 FSIQ, VIQ, and NVIQ standard scores fall within the average range, and the only measure that showed a significant sex difference in performance is the 12-Item-SA-WASI Vocabulary Subtest. Notably, while the literature generally reports a slight female advantage in the domain of verbal functioning (see, e.g., Becker & Rindermann, 2017), the current study’s results contradict that standpoint, as men scored higher than women However, the difference in group sizes by factor 3 could be a possible explanation for this statistical difference.

The specific aim of Study 2 was to provide a comprehensive psychometric analysis of a revised 24-item MVT. Internal consistency increased from $\alpha = .24$, as observed in Study 1a, to $\alpha = .55$, yet still lagged behind the value obtained in Study 1b ($\alpha = .73$). This relatively modest

coefficient is most likely due to the smaller sample size in Study 2, an issue that is addressed in Study 3. Thanks to the use of more appropriate criterion measures, I obtained more representative criterion-related validity results. The validity analysis suggested a statistically significant and substantial relationship between participants' MVT scores and their performance on the verbal subscales of the KBIT-2 and the Shipley-2, as well as on both measures' Composite IQ indices.

Regarding linguistic predictors of cognitive performance, Study 2's results mirror the key findings from Study 1—despite the change in criterion measures from one study to the next. The most important of these findings is that the MVT is the only verbal measure that is not influenced by the test-taker's first or most dominant language, English semantic fluency performance, or amount of time spent in a predominantly English-speaking environment. On all (sub)scales other than the KBIT-2 NVIQ subscale and the MVT, I detected statistically significant group mean differences between E1- and E2-speakers, as well as between those with English and those with another language as their dominant language. This difference also manifests in the correlations between MVT and criterion performance when analyzed along these lines—correlations for E1-speakers and English-dominant individuals are higher than for their respective counterparts. This finding underpins the claim that the MVT is a linguistically fair measure and that it, unlike the standardized criterion measures used, is less susceptible to influences of test-takers' linguistic background.

CHAPTER 5:

STUDY 3—A REVISED MVT: RELIABILITY AND PREDICTORS OF PERFORMANCE

In response to some of the needs identified in Study 2, the primary aim of Study 3 was to gather additional empirical data on the 24-item MVT from a larger and more heterogeneous sample than in Study 2. This larger dataset would then allow for a more powerful (congeneric) reliability and item analysis of the MVT. Accordingly, this chapter's focus lies on the psychometric analysis, which focusses around the IRT output, while still reporting the standard CTT statistics.

A secondary aim of the study was to analyze the influence of participants' sociodemographic characteristics and linguistic profiles on MVT performance by adding predictive power to the linear regression models used in Studies 1 and 2. The reproduction of the simple linear regression analyses with increased statistical power allows for a substantiation of the MVT's status as a linguistically fair instrument.

Methods

Design and Setting

This was a cross-sectional descriptive study. It was entirely conducted online, with participants sampled from five different university student populations.

Participants

Recruitment. I approached numerous South African universities, with a focus on regions where Afrikaans and isiXhosa are prevalent languages. More specifically, I requested cooperation from the following offices, asking them to circulate among their respective student bodies an email invitation to participate in this study: the Research Ethics Committee (Human) at

Nelson Mandela University (Port Elizabeth, Eastern Cape); the Office of the Registrar at Rhodes University (Grahamstown, Eastern Cape); the Department of Student Affairs at the University of Cape Town (Cape Town, Western Cape); the Office of the Registrar at the University of Fort Hare (Alice, Eastern Cape); the Office of the Deputy Registrar for Academic Administration at the University of the Western Cape (Cape Town, Western Cape).

Eligibility criteria. These were identical to those for Study 2. Given that all study measures were administered online, participants self-reported all relevant eligibility information.

Final sample. Figure 13 presents details of participant attrition through the study procedures. Initially, 871 individuals responded to the invitation email by clicking on the hyperlink. After filtering for MVT completion, I excluded 344 datasets. Of the remainder, more than 20% had missing data in some of the key linguistic and sociodemographic sections. Due to the categorical nature of many of the linguistic and sociodemographic items and the shortfalls associated with appropriate data imputation, the sample selection process represented an attempt to strike a balance between the desire to obtain the largest possible sample while avoiding incomplete datasets. Ultimately, I decided to proceed with two distinct samples for the psychometric and regression analyses.

Accordingly, all those who had completed both the MVT and PHQ-9, and who did not suffer from untreated severe depression, were included in the psychometric analysis (hereafter, the psychometric sample; $N = 494$). The final regression sample ($n = 302$) contains those who, in addition to the above, had completed the most important sociodemographic and adapted LEAP-Q items pertaining to the key variables of age, sex, first language, and most dominant language.

Hence, the regression analysis drew on data from 203 women and 99 men aged 18 to 34 years ($M = 22.42$, 95% CI [22.00, 22.83], $SD = 3.68$). The 302 individuals who provided relevant academic information had completed between 12 and 21 years of formal education ($M = 13.35$,

95% CI [13.13, 13.57], *SD* = 1.94). A post-hoc power analysis using G*Power (Faul et al., 2009) with $n = 302$, $\alpha = .05$, 4 predictors, and an estimated effect size Cohen's $f = .25$ (corresponding to a coefficient of determination of $R^2 = .20$) computed an achieved power of .99 for multiple linear regression models and means comparisons using *t*-tests. Achieved power only dropped below .90 for effect sizes lower than $f = .05$ ($R^2 = .05$).

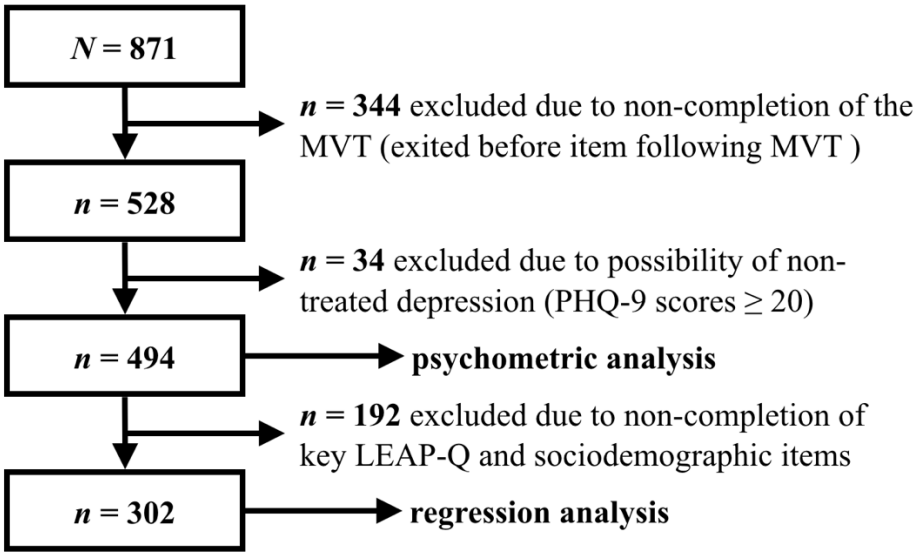


Figure 13. Participant attrition chart for Study 3.

Measures

This study used measures described in previous chapters (viz., the sociodemographic questionnaire, adapted LEAP-Q, and PHQ-9). Regarding the MVT, revisions made prior to launching this study were minor compared to those carried out between Studies 1 and 2. Specifically, I decided to maintain all 24 items from Study 2, but I made the following changes (item numbers refer to position of items in the current study):

- Item 5 (*effort* | *poging* | *umazmo*): 1-mark response (*result* | *resultaat* | *ukwenza*) changed to (*hard work* | *harde werk* | *umsebenzi onzima*);

- Item 13 (*truck* | *trok* | *itrakhi*): 1-mark response changed from (*transporter* | *vervoerder* | *isithuti*) to (*transporter* | *vervoerder* | *umthuthi*);
- Item 17 (*announce* | *aankondig* | *ukubhengeza*): item changed from (*announce* | *aankondig* | *ukwazisa*) and order of response option changed;
- Item 19 (*picture* | *prent* | *umfanekiso*): 2-mark response (*painting* | *skildery* | *ifoto*) changed to (*painting* | *skildery* | *ukuzoba*).

Additionally, items were rearranged from their Study 2 order according to item difficulty, approximated by the sums of scores awarded in Study 2, from highest to lowest. The new item order was described in the Study 2 results, and an overview of the changes of administration order across studies can be found in Appendix F.

I combined all of the abovementioned instruments into a single online survey hosted on the SurveyMonkey platform.

Procedure

All procedures were entirely self-administered, using the set of online questionnaires described above. Clicking on the link in the invitation email took prospective participants to the survey, where they were prompted to read a description of the study and to agree to voluntary participation on an informed consent document. After consenting, they were presented with the MVT instructions and, subsequently, were prompted to complete the MVT. Then, they were asked to complete the PHQ-9, followed by the sociodemographic questionnaire and the adapted LEAP-Q. The last page of the survey displayed a debriefing message and my contact details.

These procedures were approved by the Ethics Review Committee of the Humanities Faculty at UCT (Appendix K) and were accepted by all other participating universities' relevant authorities.

Statistical Analyses

The software used to analyze the data, the manner of handling assumptions underlying inferential statistical analyses, and the evaluation of output values were identical to those described in Study 2.

Preliminary analyses. I followed the steps outlined in Study 1 to provide a description of the sample characteristics and to assess for the presence of statistically significant sex differences in the key sample characteristics. Moreover, as in Study 2, I assessed for the presence of statistically significant difference in key sample characteristics pertaining to first language and dominant language.

Psychometric analysis. The psychometric analysis followed the steps described in Study 2. Although I reported and evaluated CTT output, the focus was on the IRT analysis, including ICCs, IICs, and the TIF. Given that this study design did not include criterion measures, I only report results from the internal consistency assessment and the item analysis.

Regression modelling. I repeated the set of simple linear regressions that were conducted in Studies 1 and 2.

Results

Sample Characteristics

Table 14 summarizes the regression sample's key sociodemographic and linguistic characteristics. (A similar table could not be generated for the psychometric sample because all of those participants did not complete the sociodemographic questionnaire.) The modal participant was a 22-year-old black female studying at the undergraduate level, with English as both her first and dominant language. Analyses detected no significant between-group differences on any of the evaluated variables except for age: on average, men were statistically significantly older than women.

Table 14
Study 3: Sociodemographic Characteristics of the Regression Subsample (n = 302)

Variable	Entire subsample (n = 302)	Women (n = 203)	Men (n = 99)	t / χ^2	p	ESE
Age (years)	22.42 (3.68)	22.12 (3.37)	23.02 (4.20)	2.00	.046*	0.23
Years of Education Completed[a]	13.35 (1.94)	13.37 (1.89)	13.30 (2.06)	0.27	.785	0.03
Number of Languages Spoken	3.03 (1.18)	3.07 (1.18)	2.94 (1.19)	0.895	.372	0.10
Race[b]				2.99	.394	.10
Black	111 (39.64)	68 (36.96)	43 (44.79)			
Coloured	59 (21.07)	38 (20.65)	21 (21.86)			
White	99 (35.36)	69 (37.50)	30 (31.25)			
Other/Not declared	11 (3.93)	9 (4.89)	2 (2.08)			
Dominant Language				3.56	.314	.11
Afrikaans	20 (6.62)	12 (5.91)	8 (8.08)			
English	187 (61.92)	127 (42.05)	60 (60.60)			
isiXhosa	18 (5.96)	9 (2.98)	9 (0.09)			
Other	77 (25.50)	55 (17.24)	22 (22.22)			
Language Acquired First				1.98	.578	.08
Afrikaans	28 (9.27)	17 (8.37)	11 (11.11)			
English	122 (40.40)	83 (40.89)	39 (39.39)			
isiXhosa	37 (11.49)	28 (13.79)	9 (9.09)			
Other	115 (35.71)	75 (36.95)	40 (40.40)			

Note. For the continuous variables (*Age*, *Years of Education Completed*, *Number of Languages Spoken*), means are presented with standard deviations in parentheses. For the remaining (categorical) variables, frequencies are given with percentages in parentheses. Between-group differences were assessed using independent-samples t-tests for the continuous variables and Fisher's exact tests for the categorical variables (as some of the expected cell frequencies were smaller than 5). ESE = effect size estimate (Cohen's d for continuous variables and Cramer's V for categorical variables). If percentages do not add up to 100%, it is due to rounding.
[a]n = 300 (201 women, 99 men) due to non-reported data; [b]n = 280 (184 women, 96 men) due to non-reported data.
*p < .05, two-tailed.

MVT Performance

Given the relevance of linguistic and sociodemographic data when interpreting MVT performance, this analysis of MVT performance is restricted to data from the regression subsample (n = 302). Figure 14 shows the distribution of scores, with M = 35.14, 95% CI [34.69, 35.59], SD = 3.97. Analyses detected no statistically significant sex differences in performance, (women: n = 203, M = 35.27, 95% CI [34.72, 35.81], SD = 3.93; men: n = 99, M = 34.88, 95% CI [34.07, 35.68], SD = 4.04), $t(300) = 0.80$, p = .427, Cohen's d = 0.10.

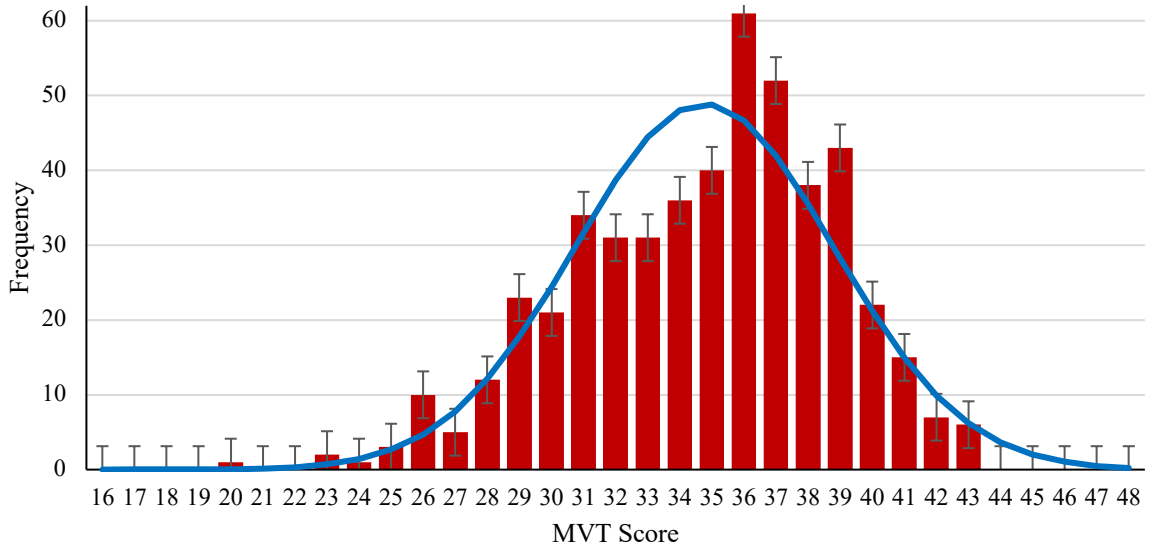


Figure 14. Distribution of MVT scores in Study 3 (N = 494), with normal curve. The entire range of scores is depicted.

A bivariate correlational analysis indicated that age was positively associated with MVT score, r = .15, p = .011. Given this apparent age effect on MVT performance and the statistically significant age difference between women and men in the sample, I conducted a univariate analysis of variance (ANOVA) that examined possible sex x age interaction effects. The analysis suggested the interaction effect was not statistically significant, $F(14, 270) = 1.04$, p = .417, η^2 = .05. This result, and especially the associated small effect size, allowed me to disregard the marginally statistically significant age difference between the sexes, as well as the age effect on MVT performance, in subsequent analyses.

MVT Psychometric Analysis

This analysis draws on data from the psychometric sample (N = 494). As noted above, I could not provide a complete description of this sample's sociodemographic and linguistic characteristics. However, for the purpose of the psychometric analysis, the increased sample size over the regression subsample, for whom I could provide such a description, increased the ability range covered by the statistical techniques employed.

Scale reliability and internal consistency. For the full 24-item MVT, split-half reliability using the Spearman-Brown correction was r = .47, and internal consistency estimates were Cronbach's α = .48 and McDonald's ω = .48. For the tentative abbreviated 13-item MVT that was proposed based on Study 2 results (i.e., items 1, 3, 6, 7, 9, 10, 12, 13, 16, 19, 20, 21, and 22) the internal consistency values were α = .36 and ω = .37. Following the item analysis of the Study 3 response sets (see below), I proposed a new 14-item short form comprising items 2, 6, 7, 8, 9, 11, 12, 13, 15, 16, 17, 20, 23, and 24. For that form, scale reliability increased to α = .58 and ω = .59.

Item analysis. As in Study 2, I report on both CTT and IRT analyses. Here, however, the focus lies on the IRT analysis.

Item-total correlations and item difficulty analysis. Table Q-1 (see Appendix Q) contains a list of item-total correlations and item difficulty values for each of the 24 MVT items. Using conventional CTT cut-off values, 14 items showed both a sufficiently high item-total correlation (> .20) and appropriate difficulty levels (between .20 and .70). Figure 15 shows the item difficulty curves and the sums of scores awarded per item for the current sample. One can observe a clearer trend of increasing difficulty, although there are still some outliers from item 18 onward. Apart from item 18 itself, the 1-or-2-mark response curve shows a slight downward trend and the 1-mark and 2-mark response curves intersect toward the middle of the measure. This pattern suggests that more participants selected the most correct response (rather than a partly correct response) in the first half of the test, but that

more selected a partly correct response (rather than the most correct one) in the second half of the test. Some smoothening of the item difficulty progression is, however, still necessary. This process is described in the later sections of this chapter.

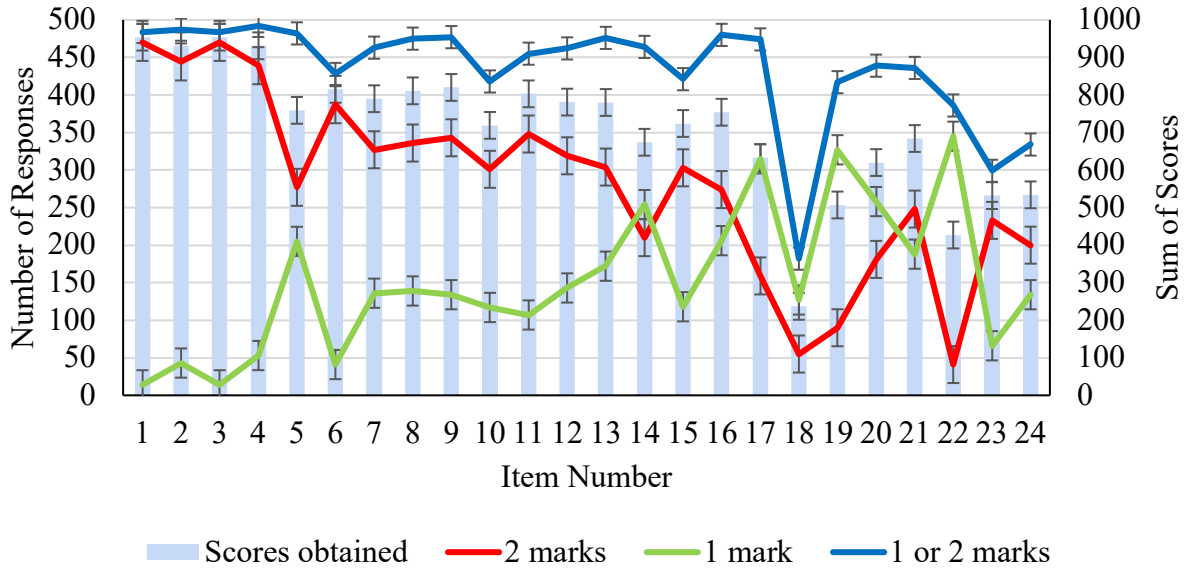


Figure 15. Relative item difficulty curves and totals of scores awarded per item for the 24 MVT items in Study 3, arranged in the original order of administration ($N = 494$).

IRCCCs. As in Study 2, the best model to describe the present data was the unconstrained GRM, $\chi^2(23) = 216.46$, $p < .001$. I computed all 24 IRCCCs (see Appendix R). Unlike in Study 2, however, I produced IRCCCs for both the conventional range of ability levels (i.e., -4 to 4) and for a wider range, from -10 to 4, given that more than 75% of the MVT's informative power falls in the below-average range (also see Appendix S).

Visual scrutiny of the curves suggests that some of the items seem to lack meaningful discrimination when screened around the average ability level (even though some hint at it toward the lowest point of the *x*-axis), but show their discriminative potential at the very low end of the ability spectrum. This phenomenon is illustrated in Figures 16 and 17, which show the IRCCCs for item 1 (*convince* | *oortuig* | *ukweyisela*) for the two different ability ranges. The same phenomenon can also be observed in the curves for items 2, 3, 6, 7, 8, 9, 11, and 13, whereas items 15, 16, 17, 20, 21, and 23 show their discriminability in the average range.

The remaining items either lack discriminability (items 5, 12, and 19) or produce negative discriminability values (items 4, 10, 14, 18, and 22).

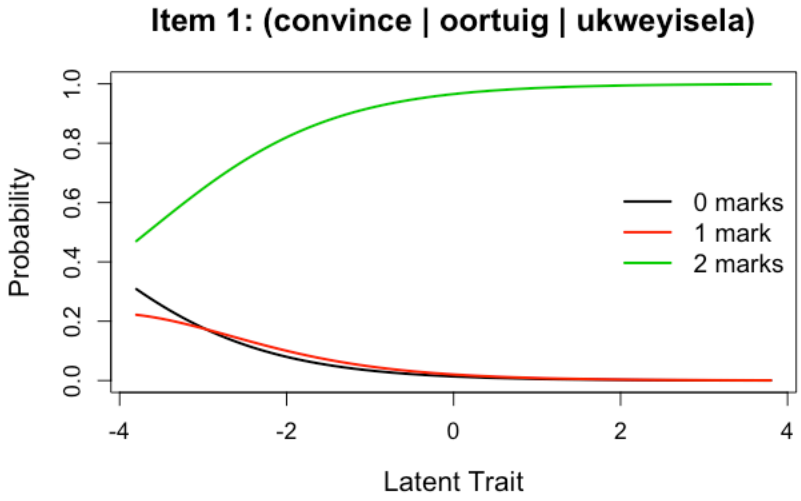


Figure 16. IRCCC for MVT item 1 in Study 3, for ability range -4 to 4 ($N = 494$).

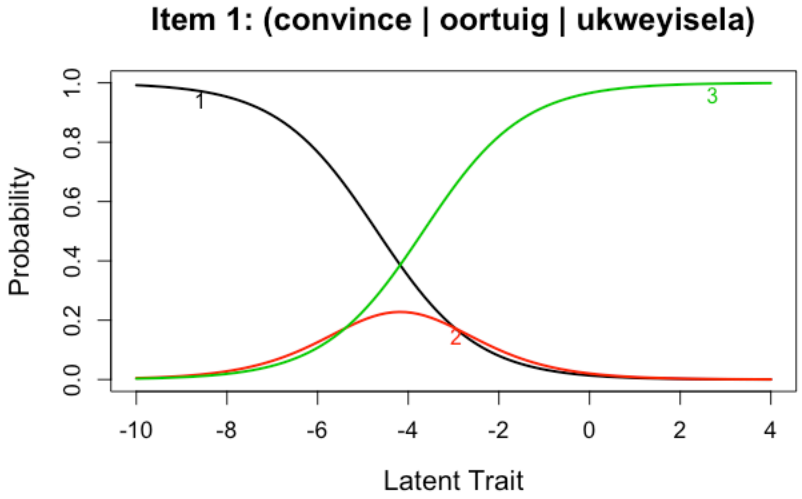


Figure 17. IRCCC for MVT item 1 in Study 3, for ability range -10 to 4 ($N = 494$).

Item information. Figure 18 shows the IICs for all 24 MVT items, using the ability range from -4 to 4. It appears that items 1, 2, and 3 provide the most information at the lower end of the ability spectrum, whereas item 23 provides the most information at the average ability level.; item 16 provide a rather consistent amount of information across the lower and middle ability range. Moreover, whereas items 7, 8, 15, 20, and 21 also provide information

above the 0.1 level, all other items provide very little information across all ability levels. (For a depiction of IICs for the broader ability range, please see Appendix T).

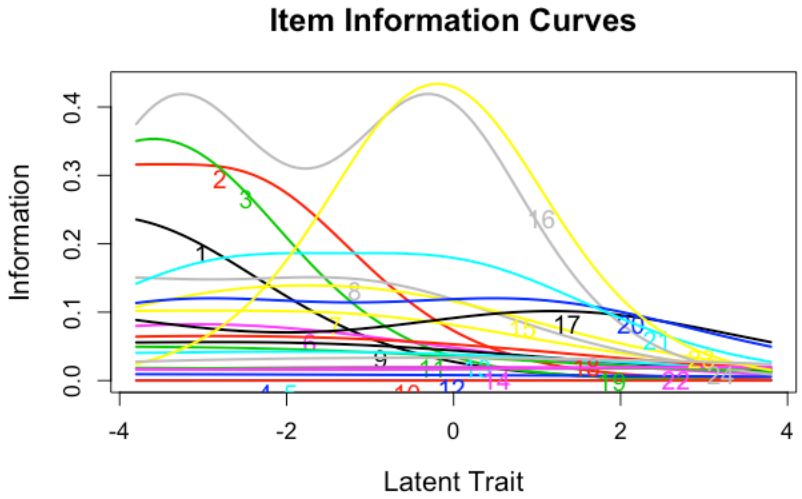


Figure 18. IICs for MVT items 1-24 in Study 3 ($N = 494$).

Test information. Figure 19 displays the TIF for the 24-item MVT. First, note that the scale of the *y*-axis is not the same as in Figure 18. The downward slope of the curve shows that the test provides most information (i.e., allows for best discrimination and most accurate measurement) in the lower half of the ability spectrum. This means that the MVT works best for individuals performing anywhere below the average ability level and that discrimination and utility decrease rapidly thereafter. Overall, 75% of the total test information was obtained between ability levels of -10 and 0 (see Appendix U for the TIF with the broader ability range), which substantiates the MVT's informative quality for below-average performance.

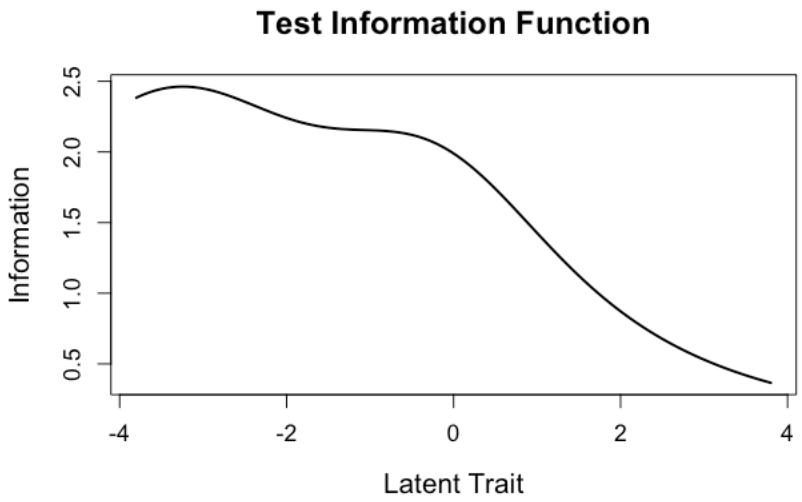


Figure 19. TIF for the 24-item MVT in Study 3 (N = 494).

Linguistic Factors Predicting Test Performance

To assess whether the MVT's resistance to personal linguistic factors (shown in Study 1a and then replicated in Study 2) would again be replicated here, in the biggest of the three samples, I repeated the simple linear regression analyses used in the previous studies (see Table 15). The linguistic factors I entered into the models were those that had a statistically significant influence on performance on the monolingual criterion measures in Studies 1a and 2, but they did not influence MVT performance in those studies. Here, at a first glance, E1 status, number of languages, years spent in an English-speaking school, and years spent in an English-speaking family were statistically significant predictors of MVT performance. It is worth noting, however, that each accounted for a very small portion of the variance in the outcome (R^2 < .075 in each case).

Table 15
Study 3: Summary of Simple Linear Regression Analyses of Linguistic Factors Influencing MVT Performance (n = 302)

Variable	R^2	F	β	p
English dominance	<.01	2.15	.12	.143
E1 status	.01	4.34	.08	.038*
Number of languages	.07	22.74	-.27	<.001***
Years spent in:				
English school[a]	.02	4.94	.13	.027*
English family[b]	.02	4.96	.13	.027*
Years of Education Completed	.03	7.72	.16	.006**

Note. *English dominance and E1 status* are dummy variables created for the purpose of the regression analyses. MVT = Multilingual Vocabulary Test
[a]n = 288 due to non-reported data. [b]n = 285 due to non-reported data.
***p < .001. **p < .01. *p < .05. All p-values are two-tailed.

I used univariate ANOVA (with independent variables being groups with 1, 2, 3, 4, and 5 or more languages) to further examine the influence of number of languages an individual had acquired on MVT performance. The analysis detected a statistically significant omnibus result, $F(4, 297) = 6.01$, $p < .001$, $\eta^2 = 0.08$. Bonferroni post-hoc pairwise tests indicated that the only significant difference was between those who had acquired 2 languages ($n = 114$) and those who had acquired 5 or more languages ($n = 53$), with a mean difference of 2.99 in favor of bilinguals, $t(165) = 5.00$, $p < .001$, and $d = 0.78$. Examining the response sets of those who reported having acquired five or more languages, almost half of them (45.28%, $n = 24$) had learned one of the MVT languages as a third language only, and almost a quarter (24.53%, $n = 13$) had learned languages other than English, Afrikaans, and isiXhosa as their first, second, and third language. This result qualifies the influence of the number of languages an individual has acquired on MVT performance.

Next, I looked at the influence of E1 status, years spent in an English-medium-of-instruction school, and years spent in an English-speaking family. These three variables were moderately correlated with one another (rs between .21 and .53, all ps < .001), which is likely due to the fact that many South Africans (regardless of L1) attend English MoI schools and, often in response to the former, use English as a family language (see, e.g., Desai, 2013; Taylor & Fintel, 2016). Nonetheless, looking only at those who reported English as their first

language (n = 122), 86.07% spent all their life in an English-speaking family and 90.16% had spent at least 12 years of education with English as the MoI. Hence, taken together, positive and/or high responses indicated a strongly English-dominated environment. Of note here, then, is that E1 status is strongly associated with high quality of education (dichotomized here as private school = high and public school = low). This finding is consistent with those of other South African studies investigating the influence of quality of education on neuropsychological test performance (see, e.g., Shuttleworth-Edwards & Kemp, 2004). Analyses indicated that E1 speakers were 1.93 times more likely to have attended a private high school than their peers, $\chi^2(1) = 10.81$, $p = .001$ (n = 278), and 2.40 times more likely to have attended a private primary school, $\chi^2(1) = 5.90$, $p = .015$ (n = 278). Hence, the current design does not allow one to conclude whether the observed effect on MVT performance is a consequence of quality of education or of an English-dominated social and educational environment. Regardless of what the underlying factors are, however, it remains important to remember the very low effect sizes associated with these predictive models.

Discussion

Responding to the need for a bigger sample identified in Study 2, the current study recruited a larger number of university students, sampled from multiple historically different universities rather than from a single institution. Hence, not only was the current sample larger, it was also more diverse in sociodemographic character (particularly with regard to linguistic and racial makeup). This study characteristic strengthened the validity of the psychometric analysis provided here, which did not replicate the reliability results produced for the 13- and 24-item version of the MVT used in Study 2. From a CTT point of view, in spite of the increased power, the MVT's psychometric properties still leave room for improvement, especially with regard to the instrument's moderate internal consistency.

However, from an IRT perspective, the MVT fulfils at least one of its intended goals: It elicits most of its information in the lower half of the ability spectrum, nearer the cut-off (to

be determined) for cognitive impairment. This characteristic of the scale is demonstrated particularly by the IICs and the TIF, especially those versions that depict data across the increased ability range. This mirrors and corroborates the findings that were already suggested in the Study 2 results. However, the fact that the current study's IIFs paint a different picture compared to those obtained in Study 2, highlights the benefit of postponing premature item changes (such as after the Study 2 analysis with a smaller sample). The information available here, however, prompted me to suggest some changes for the proposed version of the MVT, post-Study 3.

Prior to making decisions about item eliminations or changes, it is important to reiterate the intended purpose of the MVT. Responding to South African clinicians' demand for a brief, easy to score and easy to administer tool that can screen for general intellectual functioning, the instrument should be tailored to detect possible cognitive impairment, rather than provide a carefully stratified measure of intelligence. For the detection of possible cognitive impairment, two notions are of relevance: the ability to provide a roughly stratified output by category (e.g., below, at, or above average), and a focus in terms of finer differentiations on the lower half of the ability range. Following from these premises, and subsequent to the analysis of the results provided above, I planned a number of changes to the version of the MVT used in this study.

The planned changes are both on the item level and the scale level, and they constitute either item deletions or changes to the administration order. Changes to the response options or their weighting could be considered, but this would require a renewed administration and evaluation thereof. Hence, I settled on proposing the deletion of items 4, 10, 14,18, and 22 (i.e., those items displaying both low item-total correlations and negative discriminability values), as well as items 1, 3, 5, 12, and 19 (i.e., those items whose deletion increases internal consistency). Next, rearranging the remaining items in order of difficulty (approximated by sums of scores obtained across the sample) results in an abbreviated 14-item MVT, with

administration order of 2, 9, 6, 8, 11, 7, 13, 16, 15, 21, 17, 20, 23, and 24 (item numbers as labeled in the current study's administration).

Following the psychometric analysis and considering both CTT and IRT criteria, the proposed 14-item instrument constitutes the best possible version of the MVT thus far and shows improved psychometric properties compared to the version used in Study 2. Not only is it constituted by the subset of items with the highest observed internal consistency (ω = .59), but it also comprises those items that convey the greatest possible amount of information in the lower ability spectrum.

Finally, although the regression analysis showed some influence of test-takers' language profile on performance, the small effect sizes and the associations with quality of education offer some qualifications for these results. Hence, after Study 3, the MVT still stands as a linguistically fair cognitive screening tool, but with an ongoing need for psychometric refinement.

CHAPTER 6:

GENERAL DISCUSSION

This chapter starts by providing a brief summary of the development and psychometric evaluation of the Multilingual Vocabulary Test, as reported in full across the three separate empirical studies documented in Chapters 3, 4, and 5. Then, it discusses findings regarding linguistic variables that appeared to influence cognitive performance (both on the MVT and the criterion measures), highlights some of the practical and methodological challenges to the development of inherently multilingual assessment tools, and evaluates the progress of the MVT Research Project thus far. To conclude the chapter, I briefly touch on some of the project's limitations, upcoming endeavors, and suggestions for future research.

MVT Progress Report: A Summary of the Empirical Studies

A chronological summary of the three empirical studies that together comprise the core of this Master's thesis tracks the development and step-by-step refinement of the MVT as a linguistically fair screening tool for cognitive impairment in the population of South Africa's Western Cape province. The key novelty of the instrument is the implementation of an inherently multilingual testing format, which allows test-takers to choose their preferred response language for each test item, rather than having to settle on one language for the entire measure. This idea developed into a 12-item pilot version, modelled on the 12-Item SA-WASI Vocabulary subscale. Initially (i.e., at the outset of Study 1), I devised an English word list, which was then translated and back-translated into Afrikaans and isiXhosa. Only those words that successfully underwent this procedure were retained. Revisiting the needs of South African clinicians, I quickly settled on examining only the digital version of the MVT during subsequent studies (i.e., Studies 2 and 3).

In response to Study 1's results, Study 2 saw the MVT doubled in length, with 12 additional items added to the instrument. The development process for each of those

additional items was identical to that for the original items, as outlined above. After some further revisions based on Study 2 results, Study 3 produced another revision of the MVT, with 14 items and an internal consistency of $\omega = .59$. Even though, from an IRT point of view, the IRCCCs and IIFs analyses highlighted remaining potential for item improvements in order to increase the instrument's informational power, the TIF showed that the MVT produces most of its information about test-takers' ability in the lower ability spectrum—as is required by a screening tool designed to quickly detect cognitive impairment in resource-stricken settings (regardless of its linguistic properties). The development of an instrument with this characteristic, as well as with the ability to fairly accommodate multilinguals, was the central practical objective of this project.

Moreover, despite the methodological difficulties outlined below, criterion validity analyses of the 13-item abbreviated MVT developed over the course of Study 2 detected correlations of .46 and .52 with the KBIT-2 and Shipley-2 Verbal IQ indices, respectively (.54 and .61, respectively, for English first-language (E1) speakers). These statistically significant, positive in direction, and moderate in magnitude correlations suggest a substantial relationship between the constructs underlying the three measures. The fact that criterion validity differed between different first-language (L1) groups suggests that either the criterion measures exhibit a bias toward E1 speakers, or that there is a more general influence of linguistic variables on MVT performance. However, given the significantly greater likelihood of E1-speakers to have grown up in an English-dominant environment and to have received a higher quality education (as documented in the Study 3 data analysis), I can still confidently conclude that the MVT displays a vastly reduced linguistic bias toward E2- and non-English-dominant-speakers when compared to the KBIT-2 and Shipley-2. In any event, these results shed light on the influence of test-takers' linguistic background on their test performance, a key driver of testing bias (Griessel, 2005; Oller et al., 2007).

Influence of Linguistic Variables on Cognitive Assessment Results

An ancillary aim of this thesis was to identify linguistic variables that influence cognitive performance, with a particular interest in how that influence varied between the MVT and the standardized criterion measures. The rich data from the sociodemographic questionnaire and the adapted LEAP-Q allowed examination of how test-takers' language profile correlated with their performance on the various measures. Across the three studies, results suggested that the two most striking influences were E1-status/English dominance and the number of languages individuals had acquired.

E1 Status and English Dominance

A common thread throughout the three empirical studies was the sensitivity of the verbal criterion measures to test-takers' linguistic background, with the variables of E1 status (i.e., whether or not participants were English first-language speakers) and English dominance (i.e., whether or not participants reported English as their dominant language) being particularly significant in this regard. In Studies 1a and 2, being an E1-speaker and having reported English as the dominant language were, independently, statistically significant predictors of criterion measure performance (12-Item SA-WASI Vocabulary subtest), with R^2s between .11 and .22. Additionally, in Study 2 the influence of E1 status and English dominance extended to performance on the other verbal criterion measures (KBIT-2 and Shipley-2 VIQ scores), with even higher R^2s of between .14 and .34.

In contrast, neither E1 status nor English dominance exerted a statistically significant influence on MVT performance in Studies 1a, 1b, and 2. Study 3's analyses detected a statistically significant, yet small, effect (R^2 = .10) of E1 status. Even though this effect size was small, I established that the high correlations between E1 status and higher quality of education within that sample likely suggest that the underlying reason for this significant result is the increased educational quality, rather than the difference in L1. This explanation is

likely to also hold true for the statistically significant (yet small, R^2 less than or equal to .02) effects of the time spent in an English-dominant family or school.

These findings confirm the oft-reported bias against second- or additional-language English-speakers in cognitive measures normed on E1 samples (see, e.g., Cockcroft et al., 2015; Karlsson et al., 2015; Peviani, Scarpa, Toraldo, & Bottini, 2016). Especially in nations with a history of English hegemonic and colonial rule, and consequent English-language dominance over indigenous languages (e.g., South Africa, Australia, or India), this bias translates to great potential for the majority of the population to be misdiagnosed simply due to not being E1 speakers. This bias against non-English speaking individuals is exacerbated in LAMICs' clinical contexts that are marked by struggles with economic and infrastructural resources. In such contexts, clinicians often resort to using Western tests without appropriate adaptations, or to interpreting performance on poorly-adapted tests relative to foreign norms (see, e.g., Branson et al., 2012; Manly, 2008; Peviani et al., 2016).

Number of Languages

Another important linguistic variable, and one that is closer to the core of multilingualism, is the number of languages an individual has acquired. In the current research, this variable was operationalized as the number of languages participants self-reported they had acquired, regardless of proficiency levels. Hence, this variable is helpful in detecting a test bias for monolingual individuals and against multilingual individuals, who, in the South African context are likely to be those who have a distinct first and/or dominant language other than English.

Prior to delving into commentary on the effects of number of languages on test performance, I would like to point out that, wherever a statistically significant effect manifested, it indicated a negative correlation between number of languages and test performance (i.e., performance was poorer in individuals who self-reported having acquired more languages). Of course, for the monolingual English criterion measures, the effects

observed here are likely to mirror those reported in the previous section—especially for the non-balanced multilinguals. Hence, it is unsurprising that, in Study 1a, analyses suggested that number of languages was a statistically significant negative predictor of 12-Item SA-WASI Vocabulary subtest performance (R^2 = .11). Similarly, Study 2 analyses detected a statistically significant negative relationship between KBIT-2 VIQ score and number of languages (R^2 =.05), yet not for the other verbal criterion measures.

In contrast, analyses of the MVT data did not detect any statistically significant associations between number of languages and performance on that test in Studies 1a, 1b, and 2. In Study 3, however, a simple linear regression detected a small (R^2 = .07) predictive effect of number of languages on MVT performance. Even though the coefficient of determination was small, I deemed it necessary to conduct Bonferroni post-hoc tests comparing groups based on the number of languages they had acquired. Those analyses suggested that the only statistically significant difference occurred between bilinguals and those having reported to have acquired five or more languages; the latter performed statistically significantly more poorly than bilinguals on the MVT. The linguistic profile of those who had acquired five or more languages showed that most had not acquired any of the MVT languages as a first or second language, which explains their disadvantage: The set of languages in the MVT did not match any of their most proficient languages.

Overall, the consideration of test-takers' linguistic profiles proved difficult, given the historical difference between South African population groups. Even though, on the surface level, Apartheid segregation was enforced purely on the basis of race, this brought with it a segregation of different L1 groups. Therefore, results of South Africa's racial inequality (such as unequal quality of education) often-times appear as differences between L1 groups (see, e.g., Cockcroft et al., 2015; Shuttleworth-Edwards & Kemp, 2004). Due to these effects manifesting in seeming linguistic differences, disentangling historically-based educational and other sociodemographic factors from linguistic ones is a difficult undertaking. It remains

important, though, to not always attribute differential performance to individuals' L1, but to take into consideration the correlates of that L1 status.

Methodological Considerations When Measuring Linguistic Fairness

One of the main methodological challenges encountered in the course of this research project was the identification of an appropriate design and sampling frame to measure the linguistic fairness of the MVT. This difficulty, at least in part, hinges on the question of which criterion measures to choose.

Study 1 used the APM. I made this choice because the instrument is nonverbal—and, thus, supposedly more culturally fair (Mackintosh, 1998; Strauss et al., 2006)— and because it was designed for high-performing populations and has been successfully tested in similar African university settings (Rushton & Skuy, 2000). However, the APM's use in the current research proved unsuccessful, as the sample's mean score was significantly lower than expected for a healthy university student population (see Rushton et al., 2004). This was probably due to the fact that the instrument's level of difficulty exceeded the ability range of this study's South African undergraduate student population. In Study 2, the choice of criterion measures (the KBIT-2 and Shipley-2) appeared more appropriate as (a) performance of the sample on each instrument's composite score index fell within the average range, and (b) both measures provided a closer emulation of the verbal component, as they had distinct verbal subscales, but they still possessed a non-verbal 'control' component, which provided a closer approximation of *g*.

Notwithstanding, getting to the crux of the issue, the project's goal was to suggest a linguistically fair instrument (an alternative to the currently used, and inherently biased, monolingual English instruments) that could be administered to multilingual individuals in the Western Cape province. Yet, if those English-focused verbal instruments are regarded as the reference criterion, how can such a new tool be deemed valid? Naturally, if the verbal subscales of widely used intelligence tests, such as the Shipley-2 or KBIT-2, disadvantage

those who do not speak English as a first or dominant language, correlations between scores on those measures and those on the instrument under investigation cannot be expected to be high. If they were, both the criterion measures and the MVT would share the same bias.

On the other hand, of course, I am not attempting to use this methodological conundrum of how to assess the MVT's validity as a verbal screening tool for overall cognitive functioning to advocate for the issuing of a blank check to the MVT as a valid tool. Acknowledging the still-standing psychometric weaknesses of the revised MVT, I rather draw attention to the methodological difficulty of assessing the psychometric properties of radically new assessment tools. One potential avenue to explore, beyond the scope of this thesis, would be to initially validate the MVT (or, in order to keep the discussion more general, of any similarly cross-linguistic tool) using an E1 sample and then testing whether an E2-/multilingual sample performs equally on said measure. In that way, validity and linguistic fairness would be assessed in two distinct steps, with the benefit of avoiding the aforementioned flawed validity assessment.

I already hinted at this strategy of a two-step validity assessment by examining correlations between the MVT and criterion measures separately within samples of, for instance, E1 speakers and English-dominant participants. Examining E1 speakers only is a first step in trying to match the normative samples of the KBIT-2 and Shipley-2 (A. S. Kaufman & Kaufman, 2004; Shipley et al., 2009). And, in line with my predictions, these correlations exceeded those obtained when using the full, linguistically diverse, sample. However, the degree of English-dominance, the overall language profile, and the sociodemographic characteristics of South African E1 speakers differs from the individuals who comprised the KBIT-2 and Shipley-2 normative samples. Therefore, as a next step, those findings would have to be analyzed separately for the different L1 groups.

However, potential performance differences across different L1 groups need not indicate that the MVT is inherently biased against certain language groups. Due to the

parallelism of the lines separating groups of different races, different L1s, and different educational standards, analyzed in light of the historical aftermath of the inequality and racial segregation enforced by the Apartheid regime, render the comparison of subsamples grouped according to their L1 a complicated and flawed undertaking. And, even once linguistic background factors have been controlled for, this still leaves the influence of cultural differences and the inappropriateness of foreign norms. To add to this complexity, linguistic fairness of a multilingual measure—building on the assumption that multilingual individuals draw on their knowledge distributed across their languages—must not be assessed by looking at languages separately (Barbosa et al., 2016; Nell, 2000). Rather, the two-step validation suggested above is likely to produce the most accurate validity results, as long as the use of L1 means comparisons only serves to validate the measure under scrutiny against criterion measures in that language and as long as one takes into consideration variables likely to correlate with L1.

Nonetheless, it remains to note that the issues regarding inappropriate application of foreign normative data remain. A critical step in concluding MVT criterion validity analyses would be to gather South African norms for the current set of criterion measures, or to use South African-developed criterion measures, and to then repeat the correlational analyses. Overall, however, the current analyses (especially those describing a significant, positive, and moderate relationship with KBIT-2 and Shipley-2 verbal subscales) suggested promising trends for MVT criterion validity.

Limitations and Suggestions for Future Research

Despite the extensive scope and the refinements made over the course of the three empirical studies reported on in this thesis, I recognize some limitations of this work, which I present alongside recommendations and plans on how to address them in future studies. Most prominently, with regard to its external validity, the MVT needs to be administered to a wider population. Sampling beyond the university realm and including individuals from different

geographical regions (which, by the nature of contemporary South Africa, means different cultures and languages), age groups, and socioeconomic classes is necessary in order to truly establish the measure's appropriateness for multilingual population. Besides, using a multilingual sample matched on quality and level of education would increase confidence in the analysis of linguistic effects by ruling out L1-educational associations, as they were observed in the above analyses.

Moreover, in the long term, and following the encouragement received during the post-test interviews, it is desirable and—based on the already-established test development process—feasible to increase the number of languages in the MVT. This would not only bring fairer IQ screening to an even broader population, but it would also allow the test administrator to compile a test using, for example, a set of three languages preferred by the test-takers, without sacrificing comparability to individuals with other language profiles. Even if the actual combination of items will have to differ between different sets of different languages, a careful IRT analysis could still ensure comparability across different versions of the MVT.

Furthermore, although the use of English as a baseline language and developing translations from that language is appropriate in South Africa (as in the development process of the initial MVT, described here) and many other regions of the world, it is not globally acceptable. In some countries, one would have to anchor the MVT in the locally most relevant supercentral language (De Swaan, 2001), which could be Mandarin, Arabic, Spanish, or German. This approach is likely to yield the greatest utility and comparability from the instrument.

Next, from a psychometric point of view, the analysis would benefit from two additional pieces of information. The first is IRCCCs by response category *and* language, rather than by response category only. Other than the imprecise participants' self-report, this could not be examined within the empirical studies described here. Gaining such information

would be helpful in evaluating the value of each individual language within a given item, as some items might perform differently in this regard.

However, such an analysis by language would still not reveal what language route(s) a test-taker used, given that they could have gotten clues to the correct response in one language, but selected the response in another one. Although this is only indirectly relevant to the MVT analysis, as the response language is not relevant to the weighting of the response, having this information would yield information about how multilingual individuals combine (or do not combine) knowledge from their various lexical knowledge pools when completing vocabulary tasks. During the Study 2 post-test interviews, for instance, some participants indicated they had used multiple languages in various orders of reference to settle on a response option, before using the English response option as the 'default' based on their daily routine of English-language assessment at the university. Therefore, in order to bring to light the likely multilingual thought process covered up by the bias toward responding in English, future research on the MVT should attempt to find a way to assess the actual language use, as opposed to merely the selected response option—for instance, by using eye-tracking or functional neuroimaging techniques during the testing procedures.

The second piece of additional information from which the analysis would benefit is definite knowledge about the reason why a test-taker has skipped an item. Thus far, however, the MVT format neither employed a forced-choice scenario, nor did it offer a neutral response. To remedy this situation, one might consider adding a response option reading "I do not know". However, the value of this addition would have to be weighed up against the possibility of test-takers accepting this neutral response option rather than committing to a definite answer.

CHAPTER 7:

CONCLUSION

The primary intentions of this thesis were twofold: On a practical level, I aimed to develop an inherently multilingual—and, hence, linguistically fair—assessment tool for overall cognitive functioning, which can help clinicians in under-resourced settings detect possible cognitive impairment. On a more theoretical level, I intended to start a conversation around and provide a first suggestion on how to address the incorporation of more than one language into a single assessment tool. And although the project was born out of a dire need in South Africa's Western Cape province, it responded, on a bigger scale, to the issues that arise with the global increase in multilingualism (European Commission, 2007).

With the MVT Research Project as described in this Master's thesis, I have made a case for linguistically fair assessment and, at the same time, have demonstrated one possible way of going about the development of inherently multilingual tools for cognitive screening purposes. Despite the room and need for an improvement of the MVT's psychometric properties, I hope to have prompted researchers and clinicians to consider multilingualism in assessment and to have opened new ways of thinking about how to do so. Such endeavors are not just innovative ways of thinking about linguistically fair assessment, but they are also a long overdue development in an increasingly multilingual world and a necessary component of fair and equal assessment for all.

How did the MVT Research Project accomplish these aims? First, the multilingual nature of the MVT elegantly circumvents the question of which of a multilingual's languages to use for the assessment. Second, the simultaneous presentation of stimulus items in English, Afrikaans, and isiXhosa allows test-takers to draw on their knowledge of multiple languages in order to access a given concept. Third, the option to use a different response language for each item caters for individual patterns of multilingualism. Importantly, the digital format

reduces the need for trained administration personnel and allows for automated scoring, thus making the MVT the kind of resource- and cost-effective initial screening tool that is eagerly sought-after by South African clinicians, as well as those in other LAMICs.

Beyond proposing a solution to "one of the most serious challenges facing the field of neuropsychology" (Razani et al., 2007, p. 107), if anything, this work has reiterated the seriousness of this challenge and stressed the need for linguistically fair cognitive tools. The regression models showed that the 12-Item SA-WASI Vocabulary subtest, the KBIT-2, and the Shipley-2 produce different results for L1-speakers of different languages. Although this statistically significant difference might not always reflect in rough screening results (seeing that both groups' scores fell into the average range), this difference can carry implications when more finely stratified results are of interest in, for example, the context of placement tests.

To conclude, with the MVT, I have addressed the long-avoided issues of accommodating linguistic diversity and multilingualism in neuropsychological screening for overall cognitive performance. This thesis therefore adds to the literature on cross-cultural neuropsychology by expanding its focus to explicitly address assessment needs in cross-linguistic and multilingual populations. Thus, as a blueprint, the MVT Research Project paves the way toward more fairness in cognitive assessments, in general, and provides a promising first step toward addressing one of South African neuropsychologists' greatest needs—that of a quick and easy-to-administer, yet linguistically fair, cognitive screening tool.

References

Abutalebi, J., & Clahsen, H. (2016). Bimodal bilingualism: Language and cognition. *Bilingualism: Language and Cognition*, *19*(2), 221–222. http://doi.org/10.1017/S1366728916000158

Adewuya, A. O., Ola, B. A., & Afolabi, O. O. (2006). Validity of the patient health questionnaire (PHQ-9) as a screening tool for depression amongst Nigerian university students. *Journal of Affective Disorders*, *96*(1), 89–93. http://doi.org/10.1016/j.jad.2006.05.021

Alexander, N. (2012). The centrality of the language question in post-apartheid South Africa: Revisiting a perennial issue. *South African Journal of Science*, *108*(9/10). http://doi.org/10.4102/sajs.V108i9/10.1443

Alexander, N. (2013). Language Policy and National Unity in South Africa/Azania. Cape Town, South Africa: The Estate of Neville Edward Alexander.

Anastasi, A., & Urbina, S. (2002). *Psychological testing*. New York, NY: Prentice Hall.

Aronin, L., & Singleton, D. (2008). Multilingualism as a New Linguistic Dispensation. *International Journal of Multilingualism*, *5*(1), 1–16. http://doi.org/10.2167/ijm072.0

Bain, S. K., & Jaspers, K. E. (2010). Review of Kaufman Brief Intelligence Test, Second Edition. *Journal of Psychoeducational Assessment*, *28*(2), 167–174. http://doi.org/10.1177/0734282909348217

Barac, R., Bialystok, E., Castro, D. C., & Sanchez, M. (2014). The cognitive development of young dual language learners: A critical review. *Early Childhood Research Quarterly*, *29*(4), 699–714. http://doi.org/10.1016/j.ecresq.2014.02.003

Barbosa, P., Nicoladis, E., & Keith, M. (2016). Bilingual children's lexical strategies in a narrative task. *Journal of Child Language*, *44*(04), 829–849. http://doi.org/10.1017/s030500091600026x

Becker, D., & Rindermann, H. (2017). Cognitive Sex Differences: Evolution and History. *Mankind Quarterly*, *58*(1), 83–92.

Bennett, J., & Verney, S. P. (2018). Linguistic factors associated with phonemic fluency performance in a sample of bilingual Hispanic undergraduate students. *Applied Neuropsychology: Adult*, *4*(4), 1–14. http://doi.org/10.1080/23279095.2017.1417309

Bhana, A., Rathod, S. D., Selohilwe, O., Kathree, T., & Petersen, I. (2015). The validity of the Patient Health Questionnaire for screening depression in chronic care patients in primary health care in South Africa. *BMC Psychiatry*, *15*(1), 859. http://doi.org/10.1186/s12888-015-0503-0

Bialystok, E. (2001). *Bilingualism in Development: Language, Literacy, and Cognition*. Cambridge, England: Cambridge University Press.

Bialystok, E. (2009). Bilingualism: The good, the bad, and the indifferent. *Bilingualism: Language and Cognition*, *12*(1), 3–11. http://doi.org/10.1017/s1366728908003477

Bialystok, E., & Craik, F. I. M. (2010). Cognitive and Linguistic Processing in the Bilingual Mind. *Current Directions in Psychological Science*, *19*(1), 19–23. http://doi.org/10.1177/0963721409358571

Bialystok, E., Craik, F. I. M., & Luk, G. (2012). Bilingualism: consequences for mind and brain. *Trends in Cognitive Sciences*, *16*(4), 240–250. http://doi.org/10.1016/j.tics.2012.03.001

Bialystok, E., Craik, F. I. M., Green, D. W., & Gollan, T. H. (2009). Bilingual Minds. *Psychological Science in the Public Interest*, *10*(3), 89–129. http://doi.org/10.1177/1529100610387084

Bilingualism and Psycholinguistics Research Group. (2017). LEAP-Questionnaire. Retrieved May 24, 2017, from http://www.bilingualism.northwestern.edu/leapq/

Blumenfeld, H. K., Bobb, S. C., & Marian, V. (2016). The role of language proficiency, cognate status and word frequency in the assessment of Spanish–English bilinguals' verbal fluency. *International Journal of Speech-Language Pathology*, *18*(2), 190–201.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*(1), 29–51. http://doi.org/10.1007/bf02291411

Branson, N., Garlick, J., Lam, D., & Leibbrandt, M. (2012). *Education and Inequality: The South African Case*. A Southern Africa Labour and Development Research Unit Working Paper Number 75. Cape Town, South Africa.

Branson, N., Hofmeyr, C., & Lam, D. (2014). Progress through school and the determinants of school dropout in South Africa. *Development Southern Africa*, *31*(1), 106–126. http://doi.org/10.1080/0376835X.2013.853610

Breuer, E., De Silva, M. J., Shidaye, R., Petersen, I., Nakku, J., Jordans, M. J. D., et al. (2015). Planning and evaluating mental health services in low- and middle-income countries using theory of change. *The British Journal of Psychiatry*, *208*(s56), s55–s62. http://doi.org/10.1192/bjp.bp.114.153841

Brickman, A. M., Cabo, R., & Manly, J. J. (2006). Ethical Issues in Cross-Cultural Neuropsychology. *Applied Neuropsychology*, *13*(2), 91–100. http://doi.org/10.1207/s15324826an1302_4

Burns, J. K. (2015). Poverty, inequality and a political economy of mental health. *Epidemiology and Psychiatric Sciences*, *24*(2), 107–113. http://doi.org/10.1017/S2045796015000086

Cawthra, T. A. (2016). *A South African-Adapted WASI Vocabulary Subtest: Construct Validity and Screening Tool Potential*. (Unpublished Honours dissertation). University of Cape Town, South Africa.

Cenoz, J. (2013). Defining Multilingualism. *Annual Review of Applied Linguistics*, *33*, 3–18. http://doi.org/10.1017/S026719051300007X

Cholera, R., Gaynes, B. N., Pence, B. W., Bassett, J., Qangule, N., Macphail, C., et al. (2014). Validity of the patient health questionnaire-9 to screen for depression in a high-HIV burden primary healthcare clinic in Johannesburg, South Africa. *Journal of Affective Disorders*, *167*, 160–166. http://doi.org/10.1016/j.jad.2014.06.003

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. http://doi.org/10.1037/1040-3590.7.3.309

Cockcroft, K., Alloway, T., Copello, E., & Milligan, R. (2015). A cross-cultural comparison between South African and British students on the Wechsler Adult Intelligence Scales Third Edition (WAIS-III). *Frontiers in Psychology*, *6*, 297. http://doi.org/10.3389/fpsyg.2015.00297

Cohen, J. (1988). *Statistical Power Analyses for the Behavioral Sciences* (2nd ed.). New York, NY: Academic Press.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *The Journal of Applied Psychology*, *78*(1), 98–104. http://doi.org/10.1037/0021-9010.78.1.98

Court, J. H., & Raven, J. C. (1993). *Manual for Raven's Progressive Matrices and Vocabulary Scales – Section 1: General overview*. Oxford, England: Oxford Psychologists Press.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. http://doi.org/10.1007/bf02310555

Das-Munshi, J., Lund, C., Mathews, C., Clark, C., Rothon, C., & Stansfeld, S. (2016). Mental Health Inequalities in Adolescents Growing Up in Post-Apartheid South Africa: Cross-

Sectional Survey, SHaW Study. *PLoS ONE*, *11*(5), e0154478. http://doi.org/10.1371/journal.pone.0154478

Daugherty, J. C., Puente, A. E., Fasfous, A. F., Hidalgo-Ruzzante, N., & Pérez-García, M. (2016). Diagnostic mistakes of culturally diverse individuals when using North American neuropsychological tests. *Applied Neuropsychology: Adult*, *24*(1), 16–22. http://doi.org/10.1080/23279095.2015.1036992

Davies, M. (2000). *Google Books (American English) Corpus (155 billion words, 1810-2009)*. Retrieved from https://googlebooks.byu.edu/

Davies, M. (2013). *Corpus of News on the Web (NOW)*. Retrieved from https://corpus.byu.edu/now/

De Swaan, A. (2001). *Words of the world: The global language system*. Cambridge, UK: Polity Press.

Desai, Z. (2013). Local languages: Good for the informal marketplace but not for the formal classroom? *Education as Change*, *17*(2), 193–207. http://doi.org/10.1080/16823206.2013.803659

European Commission. (2007). *Final report: High level group on multilingualism*. Luxembourg, Luxembourg: European Commission.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. http://doi.org/doi:10.3758/BRM.41.4.1149

Ferrett, H. L. (2011). *The Adaptation and Norming of Selected Psychometric Tests for 12- to 15-year-old Urbanized Western Cape Adolescents.* (Unpublished Doctoral dissertation). University of Stellenbosch, South Africa.

Ferrett, H. L., Carey, P. D., Baufeldt, A. L., Cuzen, N. L., Conradie, S., Dowling, T., et al. (2014). Assessing Phonemic Fluency in Multilingual Contexts: Letter Selection Methodology and Demographically Stratified Norms for Three South African Language

Groups. *International Journal of Testing*, *14*(2), 143–167. http://doi.org/10.1080/15305058.2013.865623

Finchilescu, G. (2013). Measurement. In C. G. Tredoux & K. Durrheim (Eds.), *Numbers, Hypotheses & Conclusions* (pp. 210–229). Cape Town: Juta & Company Ltd.

Foxcroft, C. D. (1997). Psychological testing in South Africa: Perspectives regarding ethical and fair practices. *European Journal of Psychological Assessment*, *13*(3), 229–235. http://doi.org/10.1027/1015-5759.13.3.229

Foxcroft, C. D., & Aston, S. (2006). Critically examining language bias in the South African adaptation of the WAIS-III. *SA Journal of Industrial Psychology*, *32*(4), 97–102. http://doi.org/10.4102/sajip.v32i4.243

Foxcroft, C. D., Roodt, G., & Abrahams, F. (2005). Psychological testing: a brief retrospective overview. In C. D. Foxcroft & G. Roodt (Eds.), *An Introduction to Psychological Assessment in the South African Context* (pp. 8–23). Cape Town, South Africa: Oxford University Press.

Friederici, A. D., & Gierhan, S. M. (2013). The language network. *Current Opinion in Neurobiology*, *23*(2), 250–254. http://doi.org/10.1016/j.conb.2012.10.002

Friesen, D. C., Luo, L., Luk, G., & Bialystok, E. (2013). Proficiency and control in verbal fluency performance across the lifespan for monolinguals and bilinguals. *Language, Cognition and Neuroscience*, *30*(3), 238–250. http://doi.org/10.1080/23273798.2014.918630

Griessel, L. (2005). Administering psychological assessment measures. In C. D. Foxcroft & G. Roodt (Eds.), *An Introduction to Psychological Assessment in the South African Context* (pp. 83–98). Cape Town, South Africa: Oxford University Press.

Grieve, K. W. (2005). Factors affecting assessment results. In C. D. Foxcroft & G. Roodt (Eds.), *An Introduction to Psychological Assessment in the South African Context* (pp. 224–241). Cape Town, South Africa: Oxford University Press.

Grieve, K. W., & Viljoen, S. (2000). An Exploratory Study of the Use of the Austin Maze in South Africa. *South African Journal of Psychology*, *30*(3), 14–18. http://doi.org/10.1177/008124630003000303

Grosjean, F. (1989). Neurolinguists, beware! The bilingual is not two monolinguals in one person. *Brain and Language*, *36*(1), 3–15. http://doi.org/10.1016/0093-934X(89)90048-5

Grosjean, F. (2008). Studying Bilinguals. Oxford, UK: Oxford University Press.

Hall, B. J., Puffer, E., Murray, L. K., Ismael, A., Bass, J. K., Sim, A., & Bolton, P. A. (2014). The Importance of Establishing Reliability and Validity of Assessment Instruments for Mental Health Problems: an Example from Somali Children and Adolescents Living in Three Refugee Camps in Ethiopia. *Psychological Injury and Law*, *7*(2), 153–164. http://doi.org/10.1007/s12207-014-9188-9

Hambleton, R. K. (2000). Emergence of Item Response Modeling in Instrument Development and Data Analysis. *Medical Care*, *38*(9), II60–II65.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. New York, NY: Springer Science+Business Media.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Hamel, R., & Schmittmann, V. D. (2006). The 20-minute version as a predictor of the Raven Advanced Progressive Matrices Test. *Educational and Psychological Measurement*, *66*(6), 1039–1046. http://doi.org/10.1177/0013164406288169

Hebben, N., & Milberg, W. (2009). *Essentials of Neuropsychological Assessment* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Herbert, R. K., & Bailey, R. (2002). The Bantu languages: Sociohistorical perspectives. In R. Mesthrie (Ed.), *Language in South Africa* (pp. 50–78). Cambridge: Cambridge University Press.

Higby, E., Kim, J., & Obler, L. K. (2013). Multilingualism and the Brain. *Annual Review of Applied Linguistics*, *33*, 68–101. http://doi.org/10.1017/S0267190513000081

Huang, F. Y., Chung, H., Kroenke, K., Delucchi, K. L., & Spitzer, R. L. (2006). Using the patient health questionnaire-9 to measure depression among racially and ethnically diverse primary care patients. *Journal of General Internal Medicine*, *21*(6), 547–552. http://doi.org/10.1111/j.1525-1497.2006.00409.x

jamovi project. (2018). jamovi (Version 0.9). Retrieved from https://www.jamovi.org

Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological Testing: Principles, Applications and Issues.* Belmont, CA: Wadsworth, Cengage Learning.

Karlsson, L. C., Soveri, A., Räsänen, P., Kärnä, A., Delatte, S., Lagerström, E., et al. (2015). Bilingualism and Performance on Two Widely Used Developmental Neuropsychological Test Batteries. *PLoS ONE*, *10*(4), e0125867. http://doi.org/10.1371/journal.pone.0125867

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Brief Intelligence Test, Second Edition*. Bloomington, MN: Pearson, Inc.

Klein, D., Mok, K., Chen, J.-K., & Watkins, K. E. (2014). Age of language learning shapes brain structure: A cortical thickness study of bilingual and monolingual individuals. *Brain and Language*, *131*, 20–24. http://doi.org/10.1016/j.bandl.2013.05.014

Knoetze, J., Bass, N., & Steele, G. (2005). The Raven's Coloured Progressive Matrices: Pilot norms for isiXhosa-speaking primary school learners in peri-urban Eastern Cape. *South African Journal of Psychology*, *35*(2), 175–194. http://doi.org/10.1177/008124630503500202

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. http://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kroll, J. F., Gullifer, J. W., & Rossi, E. (2013). The Multilingual Lexicon: The Cognitive and Neural Basis of Lexical Comprehension and Production in Two or More Languages.

Annual Review of Applied Linguistics, *33*, 102–127. http://doi.org/10.1017/s0267190513000111

Lachenicht, L. (2013). Correlation. In C. G. Tredoux & K. Durrheim (Eds.), *Numbers, Hypotheses & Conclusions* (pp. 181–200). Cape Town: Juta & Company Ltd.

Leong, F. T. L., Park, Y. S., & Leach, M. M. (2013). Ethics in psychological testing and assessment. In K. F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology* (pp. 265–282). Washington, DC: American Psychological Association.

Lezak, M. D., Howieson, D. B., Bigler, E. D., & Tranel, D. (2012). *Neuropsychological Assessment* (5th ed.). Oxford, England: Oxford University Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Mackintosh, N. J. (1998). *IQ and Human Intelligence*. New York, NY: Oxford University Press.

Magno, C. (2009). Demonstrating the difference between classical test theory and item response theory using derived test data. *The International Journal of Educational and Psychological Assessment*, *1*(1), 1–11.

Manly, J. J. (2008). Critical issues in cultural neuropsychology: Profit from diversity. *Neuropsychology Review*, *18*(3), 179–183. http://doi.org/10.1007/s11065-008-9068-8

Manly, J. J., Byrd, D. A., Touradji, P., & Stern, Y. (2004). Acculturation, reading level, and neuropsychological test performance among African American elders. *International Journal of Speech-Language Pathology*, *11*(1), 37–46.

Manly, J. J., Jacobs, D. M., Touradji, P., Small, S. A., & Stern, Y. (2002). Reading level attenuates differences in neuropsychological test performance between African American and White elders. *Journal of the International Neuropsychological Society*, *8*(3). http://doi.org/10.1017.S135561770102015X

Marian, V., & Spivey, M. (2003). Competing activation in bilingual language processing: Within- and between-language competition. *Bilingualism: Language and Cognition*, *6*(2), 97–115. http://doi.org/10.1017/s1366728903001068

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Resarch*, *50*(4), 940–967. http://doi.org/1092-4388/07/5004-0940

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing Among Parametric Item Response Models for Polychotomous Ordered Data. *Applied Psychological Measurement*, *18*(3), 245–256. http://doi.org/10.1177/014662169401800305

Menken, K., & Shohamy, E. (2015). Invited colloquium on negotiating the complexities of multilingual assessment, AAAL Conference 2014. *Language Teaching*, *48*(3), 421–425. http://doi.org/10.1017/S0261444815000166

Moore, D., & Gajo, L. (2009). Introduction – French voices on plurilingualism and pluriculturalism: theory, significance and perspectives. *International Journal of Multilingualism*, *6*(2), 137–153. http://doi.org/10.1080/14790710902846707

Msila, V. (2014). Transforming Society through Quality Primary Education in South Africa: Lessons from Two Decades after Apartheid. *Mediterranean Journal of Social Sciences*, *5*(6), 339. http://doi.org/10.5901/mjss.2014.v5n6p339

Muñoz-Sandoval, A. F., Cummins, J., Alvarado, C. G., & Ruef, M. L. (2005). *Bilingual verbal ability tests: Normative update*. Rolling Meadows, IL: The Riverside Publishing Company.

Muraki, E. (1992). *A Generalized Partial Credit Model: Application of an EM Algorithm*. Princeton, NJ: Educational Testing Service.

Nell, V. (1994). Interpretation and misinterpretation of the South African Wechsler-Bellevue Adult Intelligence Scale: A history and a prospectus. *South African Journal of Psychology*, *24*(2), 100–109. http://doi.org/10.1177/008124639402400208

Nell, V. (1999). Standardising the WAIS-III and the WMS-III for South Africa: Legislative, psychometric, and policy issues. *South African Journal of Psychology*, *29*(3), 128–137. http://doi.org/10.1177/008124639902900305

Nell, V. (2000). *Cross-cultural neuropsychological assessment: Theory and practice*. London, England: Lawrence Erlbaum Associates.

Oller, D. K., Pearson, B. Z., & Cobo-Lewis, A. B. (2007). Profile effects in early bilingual language and literacy. *Applied Psycholinguistics*, *28*(2), 191–230. http://doi.org/10.1017/s0142716407070117

Ostini, R., & Nering, M. L. (2006). *Polytomous Item Response Theory Models*. Thousand Oaks, CA: Sage Publications.

Perani, D., & Abutalebi, J. (2005). The neural basis of first and second language processing. *Current Opinion in Neurobiology*, *15*(2), 202–206. http://doi.org/10.1016/j.conb.2005.03.007

Peviani, V., Scarpa, P., Toraldo, A., & Bottini, G. (2016). Accounting for ethnic-cultural and linguistic diversity in neuropsychological assessment of patients with drug-resistant epilepsy: A retrospective study. *Epilepsy & Behavior*, *64*, 94–101. http://doi.org/10.1016/j.yebeh.2016.09.011

Portocarrero, J. S., Burright, R. G., & Donovick, P. J. (2007). Vocabulary and verbal fluency of bilingual and monolingual college students. *Archives of Clinical Neuropsychology*, *22*(3), 415–422. http://doi.org/10.1016/j.acn.2007.01.015

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Ralarala, M. K. (2012). A compromise of rights, rights of language and rights to a language in Eugene Terre“Blanche”s (ET) trial within a trial: evidence lost in translation. *Stellenbosch Papers in Linguistics*, *41*(1), 55–70. http://doi.org/10.5774/41-0-43

Raven, J. C. (2000). The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology*, *41*(1), 1–48. http://doi.org/10.1006/cogp.1999.0735

Raven, J., Raven, J. C., & Court, J. H. (1998). *Advance Progressive Matrices: Sets I & II. Manual for Ravens Progressive Matrices and Vocabulary Scales*. San Antonio, TX: Pearson.

Razani, J., Murcia, G., Tabares, J., & Wong, J. (2007). The effects of culture on WASI test performance in ethnically diverse individuals. *The Clinical Neuropsychologist*, *21*(5), 776–788. http://doi.org/10.1080/13854040701437481

Rizopoulos, D. (2006). ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses. *Journal of Statistical Software*, *17*(5), 1–25.

Roberge, P. T. (2002). Afrikaans: Considering origins. In R. Mesthrie (Ed.), *Language in South Africa* (pp. 79–103). Cambridge: Cambridge University Press.

Rosselli, M., & Ardila, A. (2003). The impact of culture and education on non-verbal neuropsychological measurements: A critical review. *Brain and Cognition*, *52*(3), 326–333. http://doi.org/10.1016/S0278-2626(03)00170-2

Rushton, J. P., & Skuy, M. (2000). Performance on Raven's Matrices by African and White University Students in South Africa. *Intelligence*, *28*(4), 251–265. http://doi.org/10.1016/S0160-2896(00)00035-0

Rushton, J. P., Skuy, M., & Bons, T. A. (2004). Construct Validity of Raven's Advanced Progressive Matrices for African and Non-African Engineering Students in South Africa. *International Journal of Selection and Assessment*, *12*(3), 220–229. http://doi.org/10.1111/j.0965-075X.2004.00276.x

Sabanathan, S., Wills, B., & Gladstone, M. (2015). Child development assessment tools in low-income and middle-income countries: how can we use them more appropriately? *Archives of Disease in Childhood*, *100*(5), 1–7. http://doi.org/10.1136/archdischild-2014-308114

Salisbury, T. (2016). Education and inequality in South Africa: Returns to schooling in the post-apartheid era. *International Journal of Educational Development*, 43–52. http://doi.org/10.1016/j.ijedudev.2015.07.004

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, *34*(S1), 1–97. http://doi.org/10.1007/bf03372160

Sanchez, S. V., Rodriguez, B. J., Soto-Huerta, M. E., Villarreal, F. C., Guerra, N. S., & Flores, B. B. (2013). A Case for Multidimensional Bilingual Assessment. *Language Assessment Quarterly*, *10*(2), 160–177. http://doi.org/10.1080/15434303.2013.769544

Schwartz, S. J., Benet-Martínez, V., Knight, G. P., Unger, J. B., Zamboanga, B. L., Rosiers, Des, S. E., et al. (2014). Effects of language of assessment on the measurement of acculturation: Measurement equivalence and cultural frame switching. *Psychological Assessment*, *26*(1), 100–114. http://doi.org/10.1037/a0034717

Shipley, W. C., Gruber, C. P., Martin, T. A., & Klein, A. M. (2009). *Shipley-2 Manual*. Los Angeles, CA: Western Psychological Services.

Shohamy, E. (2011). Assessing Multilingual Competencies: Adopting Construct Valid Assessment Policies. *The Modern Language Journal*, *95*(3), 418–429. http://doi.org/10.1111/j.1540-4781.2011.01210.x

Shuttleworth-Edwards, A. B. (2016). Generally representative is representative of none: Commentary on the pitfalls of IQ test standardization in multicultural settings. *The Clinical Neuropsychologist*, *30*(7), 975–998. http://doi.org/10.1080/13854046.2016.1204011

Shuttleworth-Edwards, A. B. (2017). Countrywide norms declared obsolete: Best practice alert for IQ testing in a multicultural context. *South African Journal of Psychology*, *47*(1), 3–6. http://doi.org/10.1177/0081246316684465

Shuttleworth-Edwards, A. B., & Kemp, R. D. (2004). Cross-cultural effects on IQ test performance: A review and preliminary normative indications on WAIS-III test performance. *Journal of Clinical and Experimental Neuropsychology*, *26*(7), 903–920. http://doi.org/10.1080/13803390490510824

Singh, N. C., Rajan, A., Malagi, A., Ramanujan, K., Canini, M., Rosa, Della, P. A., et al. (2017). Microstructural anatomical differences between bilinguals and monolinguals. *Bilingualism: Language and Cognition*, *83*, 1–14. http://doi.org/10.1017/s1366728917000438

Slocum-Gori, S. L., & Zumbo, B. D. (2011). Assessing the Unidimensionality of Psychological Scales: Using Multiple Criteria from Factor Analysis. *Social Indicators Research*, *102*(3), 443–461. http://doi.org/10.1007/s11205-010-9682-8

Spaull, N. (2013a). Poverty & privilege: Primary school inequality in South Africa. *International Journal of Educational Development*, *33*(5), 436–447. http://doi.org/10.1016/j.ijedudev.2012.09.009

Spaull, N. (2013b). *South Africa's Education Crisis: The quality of education in South Africa 1994-2011* (pp. 1–65). Centre for Development & Enterprise.

Spaull, N., & Kotze, J. (2015). Starting behind and staying behind in South Africa. *International Journal of Educational Development*, *41*, 13–24. http://doi.org/10.1016/j.ijedudev.2015.01.002

Spearman, C. (1904a). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *15*(1), 72–101. http://doi.org/10.2307/1412159

Spearman, C. (1904b). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292. http://doi.org/10.2307/1412107

Spitzer, R. L., Kroenke, K., Williams, J. B. W., Patient Health Questionnaire Primary Care Study Group. (1999). Validation and Utility of a Self-report Version of PRIME-MD: The PHQ Primary Care Study. *Jama*, *282*(18), 1737–1744.

Spitzer, R. L., Williams, J. B. W., Kroenke, K., Linzer, M., deGruy, F. V., Hahn, S. R., et al. (1994). Utility of a New Procedure for Diagnosing Mental Disorders in Primary Care: The PRIME-MD 1000 Study. *Jama*, *272*(22), 1749–1756. http://doi.org/10.1001/jama.1994.03520220043029

Statistics South Africa. (2012a). *Census 2011: Census in brief*. Pretoria, South Africa: Statistics South Africa.

Statistics South Africa. (2012b). *Income and expenditure of households 2010/2011*. Pretoria, South Africa: Statistics South Africa.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of Neurological Tests: Administration, Norms, and Commentary* (3rd ed.). Oxford, England: Oxford University Press.

Taylor, S., & Fintel, von, M. (2016). Estimating the impact of language of instruction in South African primary schools: A fixed effects approach. *Early Childhood Research Quarterly*, *50*, 75–89. http://doi.org/10.1016/j.econedurev.2016.01.003

Templin, J. (2007). *Test Reliability & Development Using IRT*. Lawrence, KS: University of Kansas.

Thierry, G. (2016). Neurolinguistic Relativity: How Language Flexes Human Perception and Cognition. *Language Learning*, *66*(3), 690–713. http://doi.org/10.1111/lang.12186

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. http://doi.org/10.4102/sajip.v30i4.169

van Dulm, O., & Southwood, F. (2013). Child language assessment and intervention in multilingual and multicultural South Africa: Findings of a national survey. *Stellenbosch Papers in Linguistics*, *42*, 55–76. http://doi.org/10.5774/42-0-147

van Wijk, C. H., & Meintjes, W. (2015). Grooved Pegboard for adult employed South Africans: Normative data and human immunodeficiency virus associations. *South African Journal of Psychology*, *45*(4), 521–535. http://doi.org/10.1177/0081246315587692

van Wyhe, K. (2012). *Wechsler Abbreviated Scale of Intelligence: Preliminary normative data for 12-15-year-old English- and Afrikaans-speaking Coloured learners in the Western Cape*. Cape Town, South Africa.

Verhallen, M., & Schoonen, R. (1998). Lexical Knowledge in L1 and L2 of Third and Fifth Graders. *Applied Linguistics*, *19*(4), 452–470. http://doi.org/10.1093/applin/19.4.452

Walker, A. J., Batchelor, J., & Shores, A. (2009). Effects of education and cultural background on performance on WAIS-III, WMS-III, WAIS-R and WMS-R measures: Systematic review. *Australian Psychologist*, *44*(4), 216–223.

Watts, A. D., & Shuttleworth-Edwards, A. B. (2016). Neuropsychology in South Africa: Confronting the challenges of specialist practice in a culturally diverse developing country. *The Clinical Neuropsychologist*, *30*(8), 1305–1324. http://doi.org/10.1080/13854046.2016.1212098

Weber, R. C., Johnson, A., Riccio, C. A., & Liew, J. (2015). Balanced bilingualism and executive functioning in children. *Bilingualism: Language and Cognition*, *19*(2), 425–431. http://doi.org/10.1017/s1366728915000553

Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). Baltimore, MD: Williams & Wilkins.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale* (4th ed.). London, England: Pearson Assessment.

Wechsler, D. (2014). *Wechsler Adult Intelligence Scale: Fourth South African Edition*. Johannesburg, South Africa: JvR Psychometrics (Pty) Ltd.

Wechsler, D., & Zhou, X. (2011). *WASI-II: Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: The Psychological Corporation.

Wei, L. (2008). Research perspectives on bilingualism and multilingualism. In L. Wei & M. G. Moyer (Eds.), *The Blackwell Guide to Research Methods in Bilingualism and Multilingualism* (pp. 1–17). Malden, MA: Blackwell Publishing.

Wei, M., Joshi, A. A., Zhang, M., Mei, L., Manis, F. R., He, Q., et al. (2015). How age of acquisition influences brain architecture in bilinguals. *Journal of Neurolinguistics*, *36*, 35–55. http://doi.org/10.1016/j.jneuroling.2015.05.001

Wong, B., Yin, B., & O'Brien, B. (2016). Neurolinguistics: Structure, Function, and Connectivity in the Bilingual Brain. *BioMed Research International*, *2016*(16), 1–22. http://doi.org/10.1155/2016/7069274

Yang, Y., & Green, S. B. (2011). Coefficient Alpha: A Reliability Coefficient for the 21st Century? *Journal of Psychoeducational Assessment*, *29*(4), 377–392. http://doi.org/10.1177/0734282911406668

Zieff, M. R. (2017). *A Psychometric Evaluation of a Xhosa Translation of the SA-WASI Vocabulary Subtest*. (Unpublished honours dissertation). University of Cape Town, South Africa.

Appendix A
Sociodemographic Questionnaire

Sociodemographic Questionnaire
ACSENT Laboratory
University of Cape Town

Participant ID:

1. Demographics

1.1 Age:
1.2 Sex:
1.3 Race*:

2. Education

2.1 Are you currently studying? (please tick) O Yes O No
- 2.1.1 If yes, what year are you in?
- 2.1.2 If yes, what degree are you enrolled for?
- 2.1.3 What are your majors?
- 2.1.4 What language are you studying in?

2.2 What is your highest qualification?
2.3 How many years of education have you completed?
2.4 These questions pertain to your primary school:
- 2.4.1 Was it in a rural or urban setting? O Rural O Urban
- 2.4.2 What was the name of the school?
- 2.4.3 Was it a public or a private school?
- 2.4.4 What was the language of instruction?

2.5 These questions pertain to your high school:
- 2.5.1 Was it in a rural or urban setting? O Rural O Urban
- 2.5.2 What was the name of the school?
- 2.5.3 Was it a public or a private school?
- 2.5.4 What was the language of instruction?

3. General Information

3.1 What area did you live in while growing up?
3.2 Have you ever been or are you currently diagnosed with a psychological, psychiatric, neurological or learning disorder? If yes, please specify:
3.3 Are you currently taking any psychiatric/chronic medications? If yes, please specify:

*This will help us to better distinguish between the different language experiences different racial groups tend to show as first-language speakers of a given language.

Appendix B
Adapted Language Experience And Profile Questionnaire (LEAP-Q)

Adapted Language Experience And Profile Questionnaire (LEAP-Q)
Part A

Participant ID:	

1. Please list all the languages you know **in order of dominance**:

1._____ 2._____ 3._____ 4._____ 5._____

2. Please list all the languages you know **in order of acquisition** (your native language first):

1._____ 2._____ 3._____ 4._____ 5._____

3. Please list what percentage of the time you are **currently** and **on average** exposed to each language (*Your percentages should add up to 100%):*

Language:				
Percentage:				

4. When choosing to read a text available in all your languages, in what percentage of cases would you choose to read it in each of your languages? Assume the original was written in another language, which is unknown to you (*Your percentages should add up to 100%):*

Language:				
Percentage:				

5. When choosing to speak with a person who is equally fluent in all your languages, what percentage of time would you choose to speak each language? Please report the percentage of total time (*Your percentages should add up to 100%):*

Language:				
Percentage:				

6. Please name the cultures with which you identify. On a scale **from zero to ten**, please rate the extent to which you identify with each culture. (Examples of possible cultures are *black, South African, christian, etc.):*

Culture:				
Rank:				

Based on: Marian, Blumenfeld, & Kaushanskaya (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*(4), 940-96.

Adapted Language Experience And Profile Questionnaire (LEAP-Q)
Part B (to be filled in for each language)

Participant ID:
Language:

1. Age when you… …this language.

began acquiring	became fluent in	began reading in	became fluent reading in

2. Please list the number of years and months you spent in each language environment.

	years	months
A province where this language is spoken:		
A family where this language is spoken:		
A school/workplace where this language is spoken:		

3. On a scale from 0 to 10, please select your level of **proficiency** in speaking, understanding, and reading this language *(circle the appropriate number):*

	None					*Adequate*					*Perfect*
Speaking:	0	1	2	3	4	5	6	7	8	9	10
Understanding:	0	1	2	3	4	5	6	7	8	9	10
Reading:	0	1	2	3	4	5	6	7	8	9	10

4. On a scale from 0 to 10, please select how much the following factors contributed to you learning this language (*circle the appropriate number):*

	Not a contributor					*Moderate*					*Most important*
Interacting with friends:											
	0	1	2	3	4	5	6	7	8	9	10
Interacting with family:											
	0	1	2	3	4	5	6	7	8	9	10
Reading:											
	0	1	2	3	4	5	6	7	8	9	10
Language tapes/self-instruction:											
	0	1	2	3	4	5	6	7	8	9	10
Watching TV:											
	0	1	2	3	4	5	6	7	8	9	10
Listening to the radio:											
	0	1	2	3	4	5	6	7	8	9	10

5. Please rate to what extent you are currently exposed to this language in the following contexts:

	Never				*Half of the time*					*Always*
Interacting with friends:										
	0	1	2	3	4	5	6	7	8	9	10
Interacting with family:										
	0	1	2	3	4	5	6	7	8	9	10
Watching TV:										
	0	1	2	3	4	5	6	7	8	9	10
Listening to radio/music:										
	0	1	2	3	4	5	6	7	8	9	10
Reading:										
	0	1	2	3	4	5	6	7	8	9	10
Language-lab/self-instruction:										
	0	1	2	3	4	5	6	7	8	9	10

6. In your perception, how much of a foreign accent do you have in this language:

None				*Moderate*					*Pervasive*
0	1	2	3	4	5	6	7	8	9	10

7. Please rate how frequently others identify you as a non-native speaker *based on your accent* in this language:

Never				*Half of the time*					*Always*
0	1	2	3	4	5	6	7	8	9	10

Based on: Marian, Blumenfeld, & Kaushanskaya (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research, 50*(4), 940-967.

Appendix C
12-Item SA-WASI Vocabulary Subtest

South African-Adapted Wechsler Abbreviated Scale of Intelligence
12-Item Vocabulary Subtest

Participant ID:

Instructions: Start at item 1 and administer all items. Stop testing after discontinuance point (5 consecutive scores of 0). Score items up to discontinuance point.

Item		Response	Score
1	Bird		/2
2	Calendar		/2
3	Complicated		/2
4	Haste		/2
5	Entertain		/2
6	Impulse		/2
7	Cart		/2
8	Ruminate		/2
9	Intermittent		/2
10	Formidable		/2
11	Impertinent		/2
12	Tirade		/2
Total:			**/24**

Appendix D
Multilingual Vocabulary Test (pen-and-paper version)

Multilingual Vocabulary Test (MVT)

Participant ID: _____
Examiner: _____
Date: _____

Instructions: Start at item 1 and administer all items. Stop testing after discontinuance point (5 consecutive scores of 0). Score items up to discontinuance point.

Item		Response	Score
1	E: **horse**		
	A: **perd**		
	X: **ihashe**		/2
2	E: **picture**		
	A: **prent**		
	X: **umfanekiso**		/2
3	E: **train**		
	A: **trein**		
	X: **uloliwe**		/2
4	E: **announce**		
	A: **aankondig**		
	X: **ukwazisa**		/2
5	E: **suggest**		
	A: **voorstel**		
	X: **ukucebisa**		/2
6	E: **convince**		
	A: **oortuig**		
	X: **ukweyisela**		/2

Multilingual Vocabulary Test (MVT)
12-Items (continued)

Item		Response	Score
7	E: **excellence**		
	A: **uitnemendheid**		
	X: **ukugqwesa**		
8	E: **recurrent**		
	A: **terugkerend**		
	X: -**phindaphindayo**		/2
9	E: **impetuous**		
	A: **oorhastig**		
	X: **-dyuduzayo**		/2
10	E: **deliberation**		
	A: **deliberasie**		
	X: **ukucamngca**		/2
11	E: **effort**		
	A: **poging**		
	X: **umzamo**		/2
12	E: **tumult**		
	A: **rumoer**		
	X: **isidubedube**		/2
Total:			**/24**

Appendix E
Multilingual Vocabulary Test (digital version)

Multilingual Vocabulary Test (MVT)

Please provide the closest meaning of the word below.
Kies asseblief die naaste betekening van die word onder.
Khetha elona intsingiselo echanekileyo ehambelana nalamagama.

horse	**perd**	**ihashe**
□ riding animal	□ rybare dier	□ silwanyana esikhwelwayo
□ farm animal	□ plaas dier	□ isilwanyana sasekhaya
□ hoofed animal	□ gehoefde dier	□ isilwanyana esikhabayo
□ big animal	□ groot dier	□ isilwanyana esikhulu
□ strong animal	□ sterk dier	□ isilwanyana esinamandla

On-screen representation resembles the above. An item list is provided below.

d-MVT Items and Response Options (Study 1a)

Item	English	Afrikaans	isiXhosa	Score
1	**horse**	**perd**	**ihashe**	
	○ riding animal	○ riding animal	○ isilwanyana esikhwelwayo	2
	○ farm animal	○ plaas dier	○ isilwanyana sasekhaya	1
	○ hoofed animal	○ gehoefde dier	○ isilwanyana esikhabayo	1
	○ big animal	○ groot dier	○ isilwanyana esikhulu	0
	○ strong animal	○ sterk dier	○ isilwanyana esinamandla	0
2	**picture**	**prent**	**umfanekiso**	
	○ painting	○ skildery	○ ifoto	2
	○ artwork	○ kunswerk	○ umzobo	1
	○ still	○ stillewe	○ omboniso	1
	○ caption	○ opskrif	○ isazobe	0
	○ show	○ skou	○ umabonwakude	0
3	**train**	**trein**	**uloliwe**	
	○ locomotive	○ lokomotief	○ inqwelo enamakhareji	2
	○ carriage	○ wa	○ igutsi	1
	○ railway	○ spoorlyn	○ ingqwelo ende	1
	○ vehicle	○ motor	○ ingqwelo	0
	○ transport	○ vervoer	○ imoto	0
4	**announce**	**aankondig**	**ukwazisa**	
	○ proclaim	○ verkondig	○ ukuvakalisa	2
	○ make known	○ bekend maak	○ ukudumisa umba	1
	○ state	○ verklaar	○ ukusasaza iindaba	1
	○ communicate	○ kommunikeer	○ ukuthetha	0
	○ talk	○ praat	○ ukucacisa	0
5	**suggest**	**voorstel**	**ukucebisa**	
	○ propose	○ aanbeveel	○ ukuveza iimbono	2
	○ argue	○ argumenteer	○ ukubonisa	1
	○ imply	○ impliseer	○ ukunceda umntu	1
	○ say	○ sê	○ ukuyalela	0
	○ scream	○ skree	○ ukuthetha	0
6	**convince**	**oortuig**	**ukweyisela**	
	○ persuade	○ oorreed	○ ukuphembelela	2

	○ conclude ○ tempt ○ win ○ vindicate	○ gevolgtrekking ○ versoek ○ oorwin ○ verdedig	○ ukubonisana ngento ○ ukuqhubela phambili ○ ukuqiqa ○ ukubona	1 1 0 0
7	**excellence**	**uitnemendheid**	**ukugqwesa**	
	○ brilliance ○ greatness ○ sufficiency ○ performance ○ difference	○ briljant ○ grootheid ○ genoegsaamheid ○ werkverrigting ○ verskil	○ ukuphumelela ngaphambili ○ ukwenza kakuhle kakhulu ○ ukwenza ngokufanelekileyo ○ ukulunga ○ ukuphumelela	2 1 1 0 0
8	**recurrent**	**terugkerend**	**-phindaphindayo**	
	○ repetitive ○ frequent ○ regular ○ respected ○ recent	○ herhalend ○ frekwent ○ gereeld ○ gerespekteerd ○ onlangs	○ ukwenza izidlandlo ezininzi ○ ukwenza kwakhona ○ ukumana ukhumbula ○ ukukhumbula ○ iinkumbulo	2 1 1 0 0
9	**impetuous**	**oorhastig**	**-dyuduzayo**	
	○ impulsive ○ imprudent ○ uncontrolled ○ considered ○ disciplined	○ impulsief ○ onverstandig ○ onbeheersd ○ orweeg ○ gedissiplineerd	○ ukwenza into ngokungxama ○ ukwenza ngaphandle kokucinga ○ ukwenza into ngokungathali ○ ukonqena ○ ukukhathala	2 1 1 0 0
10	**deliberation**	**deliberasie**	**ukucamngca**	
	○ consideration ○ carefulness ○ thinking ○ freedom ○ communication	○ oorweging ○ versigtigheid ○ dink ○ Vryheid ○ kommunikasie	○ ukucingisisa nzulu ○ ukucinga kakhulu ○ ukucinga ngento ○ ukuqwalasela ○ ukuphonononga	2 1 1 0 0
11	**effort**	**poging**	**umzamo**	
	○ attempt ○ achievement ○ result ○ victory ○ competence	○ probeerslag ○ prestasie ○ resultaat ○ oorwinning ○ bevoegheid	○ ukuzabalaza ○ ukwenza amatiletile ○ ukwenza ○ umsebenzi ○ ukutsala nzima	2 1 1 0 0
12	**tumult**	**rumoer**	**isidubedube**	
	○ commotion ○ trouble ○ chaos ○ tantrum ○ temper	○ oproer ○ moeilikheid ○ chaos ○ vloermoer ○ humeur	○ umbhodamo ○ isiphithiphithi ○ isigxumgxum ○ abantu abaninzi ○ ingxolo eninzi	2 1 1 0 0

Appendix F

Table F-1
Administration Order of MVT Items Across Studies

	Position					
Item (*English* | *Afrikaans* | *isiXhosa*)	Study 1a	Study 1b	Study 2 Pilot	Study 2	Study 3	Final Proposed
convince | oortuig | ukweyisela	6	1	8	1	**1**	--
dinner | dinee | idinala	--	--	23	7	**2**	1
decade | decade | ishumi leminyaka	--	--	18	3	**3**	--
suggest | voorstel | ukucebisa	5	2	3	2	**4**	--
effort | poging | umzamo	11	6	5	5	**5**	--
value | waarde | ixabiso	--	--	13	11	**6**	3
excellence | uitnemendheid | ukugqwesa	7	5	21	6	**7**	6
probability | waarskynlikheid | ithuba	--	--	22	4	**8**	4
recurrent | terugkerend | -phindaphindayo	8	4	14	13	**9**	2
horse | perd | ihashe	1	3	24	8	**10**	--
impetuous | oorhastig | -dyuduzayo	9	8	7	17	**11**	5
habit | gewoonte | umkhuba	--	--	16	12	**12**	--
truck | trok | itrakhi	--	--	1	14	**13**	7
conversion | omskakeling | ukuguqula	--	--	19	9	**14**	--
deliberation | deliberasie | ukucamngca	10	7	17	19	**15**	9
tendency | neiging | isiqhelo	--	--	20	10	**16**	8
announce | aankondig | ukubhengeza	4	11	12	16	**17**	11
train | trein | uloliwe	3	10	2	20	**18**	--
picture | prent | umfanekiso	2	9	4	23	**19**	--
tumult | rumoer | isidubedube	12	12	6	18	**20**	12
parade | parade | umhambo	--	--	10	22	**21**	10
ambulance | ambulans | i-ambulensi	--	--	11	15	**22**	--
pretentious | pretensieus | ukuzenzisa	--	--	9	21	**23**	13
atoll | atoll | isiqhiti esisangqa	--	--	15	24	**24**	14

Notes. Listed according to administration order in Study 3.

Appendix G
Study 1a: Preliminary Scoring Rubric (English) for the p-MVT

Multilingual Vocabulary Test (MVT)—Preliminary Scoring Rubric				
This preliminary scoring rubric serves as a guideline of how to evaluate responses. In general, the more abstract and comprehensive a response, the higher the score should be.				
Item			**Score**	**Response**
1	E: **horse**		**0**	Animal, big animal, strong animal
	A: **perd**		**1**	Mammal, used for riding
	X: **ihashe**		**2**	Hoofed riding animal
2	E: **picture**		**0**	Something you take, with your phone
	A: **prent**		**1**	Drawing, photo, documentation
	X: **umfanekiso**		**2**	Can be painting/photographed, a captured moment
3	E: **train**		**0**	Transports people, takes people to work
	A: **trein**		**1**	Railway, public transport, vehicle
	X: **uloliwe**		**2**	Public transport on railways
4	E: **announce**		**0**	Tell people, say something to someone
	A: **aankondig**		**1**	Put out a notice, report
	X: **ukwazisa**		**2**	Proclaim, make known
5	E: **suggest**		**0**	Argue, tell your opinion
	A: **voorstel**		**1**	Put forward an idea, show
	X: **ukucebisa**		**2**	Propose, imply, insinuate
6	E: **convince**		**0**	Say, prove sth., argue
	A: **oortuig**		**1**	Make s.o. do sth., win over
	X: **ukweyisela**		**2**	Persuade, induce, sway s.o.
7	E: **excellence**		**0**	Good, nice, great work
	A: **uitnemendheid**		**1**	Accomplishment, achievement,
	X: **ukugqwesa**		**2**	Outstanding performance, brilliance, superiority
8	E: **recurrent**		**0**	Happening, once-off, now and then, always there
	A: **terugkerend**		**1**	Ongoing, keeps coming back
	X: **-phindaphindayo**		**2**	Repetitive, returning, reiterative,

Multilingual Vocabulary Test (MVT)—Preliminary Scoring Rubric (continued)				
Item			**Score**	**Response**
9	E: **impetuous**		**0**	Doing sth. quickly, fast
	A: **oorhastig**		**1**	Hasty, reckless, w/o thinking, hurry
	X: **-dyuduzayo**		**2**	Impulsive, impromptu, spur-of-the-moment
10	E: **deliberation**		**0**	Thinking, willingness
	A: **deliberasie**		**1**	Thinking deeply, discussing, consultation
	X: **ukucamngca**		**2**	Rumination, reflection
11	E: **effort**		**0**	Energy, power, making/doing sth.
	A: **poging**		**1**	Try, hard work
	X: **umzamo**		**2**	Attempt, achievement, accomplishment
12	E: **tumult**		**0**	Turmoil, confusion
	A: **rumoer**		**1**	Loud event, happening
	X: **isidubedube**		**2**	Commotion, chaotic and loud group of people

Appendix H
Study 1a: Open-ended Questions Used to Obtain Test Takers' Feedback

Post-test Interview Questions

1. How did you like the MVT?
2. How was your testing experience?
3. What aspects did you like about it?
4. What aspects did you not like about it?
5. How did it feel compared to the English-only measure (*referring to the 12-Item SA-WASI Vocabulary subtest*)?

Appendix I
Study 1: Ethical Approval Letter

UNIVERSITY OF CAPE TOWN

Department of Psychology

University of Cape Town Rondebosch 7701 South Africa
Telephone (021) 650 3417
Fax No. (021) 650 4104

09 June 2017

Julian Siebert
Department of Psychology
University of Cape Town
Rondebosch 7701

Dear Julian

I am pleased to inform you that ethical clearance has been given by an Ethics Review Committee of the Faculty of Humanities for your study, *Developing a Linguistically Fair IQ Screening Tool Appropriate to the Multilingual Reality of South Africa.* The reference number is PSY2017 -020.

I wish you all the best for your study.

Yours sincerely

University of Cape Town
PSYCHOLOGY DEPARTMENT
Upper Campus
Rondebosch

Lauren Wild (PhD)
Associate Professor
Chair: Ethics Review Committee

Appendix J
Study 1a: Performance on Outcome Measures by Sex ($N = 65$)

Table J-1
Study 1a: Performance on Outcome Measures by Sex (N = 65)

	Total sample ($N = 65$)					Women ($n = 46$)					Men ($n = 19$)					Means comparison		
				95% CI					95% CI					95% CI				
	M	*SD*	*SE*	*LL*	*UL*	*M*	*SD*	*SE*	*LL*	*UL*	*M*	SD	SE	LL	UL	*t*	*p*	*ESE*
p-MVT [a]	15.27	2.71	0.46	14.27	15.20	15.27	2.70	0.49	14.26	16.28	14.80	3.03	1.36	11.03	18.57	0.35	.727	0.17
MVT [b]	17.60	2.39	0.44	16.71	17.60	17.56	2.37	0.59	16.30	18.82	17.64	2.50	0.67	16.20	19.09	0.09	.929	0.03
12-Item-SA-WASI Vocabulary Subtest	12.08	3.97	0.49	11.09	12.08	11.61	3.91	0.58	10.45	12.77	13.21	3.98	0.91	11.29	15.13	1.50	.410	0.41
APM	17.29	4.11	0.51	16.27	17.29	17.07	3.73	0.55	15.96	18.17	17.84	4.99	1.15	15.44	20.25	0.69	.493	0.19

Notes. Mean scores are presented with standard deviations in parentheses. 12-Item SA-WASI Vocabulary subtest and MVT scores are raw scores, KBIT-2 and Shipley-2 are standard scores. Group differences were assessed using independent-samples *t*-tests. ESE = effect size estimate (in this case, Cohen's *d*). MVT: digital Multilingual Vocabulary Test. SA-WASI: South African-adapted Wechsler Abbreviated Scale of Intelligence. KBIT-2: Kaufman Brief Intelligence Scale (Second Edition).
[a] Total n = 35, 30 women and 5 men; [b] Total n = 30, 16 women and 14 men

Appendix K
Studies 2 and 3: Ethical Approval Letter

UNIVERSITY OF CAPE TOWN

Department of Psychology

University of Cape Town Rondebosch 7701 South Africa
Telephone (021) 650 3417
Fax No. (021) 650 4104

20 March 2018

Julian Siebert
Department of Psychology
University of Cape Town
Rondebosch 7701

Dear Julian

I am pleased to inform you that ethical clearance has been given by an Ethics Review Committee of the Faculty of Humanities for your study, Toward Linguistically Fair IQ Screening: The Multilingual Vocabulary test. The reference number is PSY2018-009

I wish you all the best for your study.

Yours sincerely

Lauren Wild (PhD)
Associate Professor
Chair: Ethics Review Committee

Appendix L
Study 2: Post-test Interview Schedule to Obtain Qualitative Feedback on the MVT

Qualitative Feedback
Participant ID: _____ Examiner: _____ Date: _____
Instructions: - *Ask the participant the following questions (read them out as written below).* - *Begin by saying:* ***Before we end, I would like to ask you some quick questions about your experience of taking the test on the tablet.*** - *Clarify which test you are referring to (the d-MVT, i.e. the one where they had to select the meaning of a word from the choices on the screen).* - *Ask the questions one after the other and write down their responses in as much detail as possible. The idea of this is to find out how they*
1. Please tell me about your experience of taking the test on the tablet. What was it like?
2. Please tell me how it compared to the other tasks you did, particularly to the one where you had to tell me the meaning of words and I wrote them down.
3. How did you like the test on the tablet? What did you like about it and what didn't you like about it?
4. What did you think about the fact that there were multiple languages in the test and what languages did you use?

Appendix M
Study 2: Sociodemographic Characteristics by First and Dominant Language ($N = 101$)

Table M-1
Study 2: Sociodemographic Characteristics by First Language (N = 101)

		First language				
Variable	Total ($N = 101$)	English ($n = 65$)	Other ($n = 36$)	t / χ^2	p	ESE
Age (years)	19.53 (1.97)	19.60 (2.30)	19.42 (1.16)	0.45	.656	0.09
Years of Education Completed	12.73 (2.64)	13.15 (1.47)	13.28 (1.26)	0.34	.671	0.07
Number of Languages Spoken	2.53 (0.94)	2.22 (0.61)	3.11 (1.14)	5.11	<.001***	1.06
Race				49.28	<.001***	0.70
Black	34 (34.65)	6 (9.23)	28 (77.78)			
Coloured	46 (45.54)	41 (63.08)	5 (13.89)			
White	12 (11.88)	10 (15.38)	2 (5.56)			
Other/Not declared	9 (8.91)	8 (12.31)	1 (2.78)			
Sex				1.42	.233	0.12
Female	77 (76.24)	52 (80.00)	25 (69.44)			
Male	24 (23.76)	13 (20.00)	11 (30.56)			
Dominant Language				35.79	<.001***	0.60
Afrikaans	1 (1.00)	---	1 (2.78)			
English	82 (81.19)	64 (98.46)	18 (50.00)			
isiXhosa	11 (10.90)	1 (1.54)	10 (27.78)			
Other	7 (6.93)	---	7 (19.45)			

Notes. For the continues variables (*Age*, *Years of Education Completed*, *Number of Languages Spoken*), means are presented with standard deviations in parentheses. For the remaining (categorical) variables, frequencies are given with percentages in parentheses. Group differences were assessed using independent-samples *t*-tests for the continuous variables and Fisher's exact tests for the categorical variables (as some of the expected cell frequencies were smaller than 5). L1 = first language. ESE: Effect size estimate (Cohen's *d* for continuous variables and Cramer's *V* for categorical variables). If percentages do not add up to 100%, it is due to rounding.
***$p < .001$, two-tailed.

Table M-2
Study 2: Sociodemographic Characteristics by Dominant Language (N = 101)

Variable	Total (N = 101)	Dominant language: English (n = 82)	Dominant language: Other (n = 19)	t / χ^2	p	ESE
Age (years)	19.53 (1.97)	19.63 (2.11)	19.11 (1.15)	1.06	.293	0.27
Years of Education Completed	12.73 (2.64)	13.29 (1.47)	12.79 (0.92)	1.05	.294	0.27
Number of Languages Spoken	2.53 (0.94)	2.39 (0.81)	3.16 (1.21)	3.35	.001**	0.85
Race				32.94	<.001***	0.57
Black	34 (34.65)	17 (20.73)	17 (89.47)			
Coloured	46 (45.54)	45 (54.88)	1 (5.26)			
White	12 (11.88)	11 (13.41)	1 (5.26)			
Other/Not declared	9 (8.91)	9 (10.98)	---			
Sex				0.09	.758	0.03
Female	77 (76.24)	62 (75.61)	15 (78.95)			
Male	24 (23.76)	20 (24.39)	4 (21.05)			
Language Acquired First				38.34	<.001***	0.62
Afrikaans	6 (5.94)	4 (4.88)	2 (10.53)			
English	65 (64.36)	64 (78.05)	1 (5.26)			
isiXhosa	18 (17.82)	8 (9.76)	10 (52.63)			
Other	11 (10.89)	5 (6.10)	6 (31.58)			

Notes. For the continues variables (*Age*, *Years of Education Completed*, *Number of Languages Spoken*), means are presented with standard deviations in parentheses. For the remaining (categorical) variables, frequencies are given with percentages in parentheses. Group differences were assessed using independent-samples *t*-tests for the continuous variables and Fisher's exact tests for the categorical variables (as some of the expected cell frequencies were smaller than 5). ESE: Effect size estimate (Cohen's *d* for continuous variables and Cramer's *V* for categorical variables). If percentages do not add up to 100%, it is due to rounding. ***$p < .001$. **$p < .01$ All *p*-values are two-tailed.

Appendix N
Study 2: Performance on Outcome Measures by Sex (N = 101)

Table N-1
Study 2: Performance on Outcome Measures by Sex (N = 101)

	Total (N = 101)					Women (n = 77)					Men (n = 24)					Means comparison		
				95% CI					95% CI					95% CI				
Measure	*M*	*SD*	*SE*	*LL*	*UL*	*M*	*SD*	*SE*	*LL*	*UL*	*M*	SD	SE	LL	UL	*t*	*p*	*ESE*
12-Item-SA-WASI Vocabulary Subtest	10.73	2.59	0.26	10.20	11.24	10.39	2.41	0.27	9.84	10.94	11.83	2.90	0.59	10.61	13.06	2.44	.016*	0.57
MVT	34.22	3.62	0.36	33.50	34.93	34.14	3.79	0.43	33.28	35.00	34.46	3.05	0.62	33.17	35.75	0.37	.711	0.09
KBIT-2																		
Verbal IQ	94.67	9.84	0.98	92.73	96.61	94.13	9.62	1.10	91.95	96.31	96.42	10.51	2.15	91.98	100.85	0.99	.322	0.23
Non-Verbal IQ	97.79	11.92	1.19	95.44	100.14	97.90	12.69	1.45	95.02	100.78	97.46	9.23	1.88	93.56	101.36	0.16	.876	0.04
Composite Score	95.90	10.14	1.01	93.90	97.90	95.66	10.69	1.22	93.24	98.09	96.57	8.29	1.69	93.17	100.17	0.42	.674	0.10
Shipley-2																		
Verbal IQ	101.53	9.70	0.97	99.62	103.45	100.95	9.17	1.05	98.87	103.03	103.42	11.25	2.30	98.67	108.17	1.09	.278	0.25
Non-Verbal IQ	100.28	12.19	1.21	97.87	102.68	99.38	11.66	1.33	96.73	102.02	103.17	13.60	2.78	97.42	108.91	1.34	.185	0.31
Composite B	101.18	10.86	1.08	99.03	103.32	100.23	10.24	1.17	97.91	102.56	104.21	12.38	2.53	98.98	109.44	1.58	.118	0.37

Notes. Mean scores are presented with standard deviations in parentheses. 12-Item SA-WASI Vocabulary subtest and MVT scores are raw scores, KBIT-2 and Shipley-2 are standard scores. Group differences were assessed using independent-samples *t*-tests. ESE = effect size estimate (in this case, Cohen's *d*). MVT: digital Multilingual Vocabulary Test. SA-WASI: South African-adapted Wechsler Abbreviated Scale of Intelligence. KBIT-2: Kaufman Brief Intelligence Scale (Second Edition).
*p < .05, two-tailed.

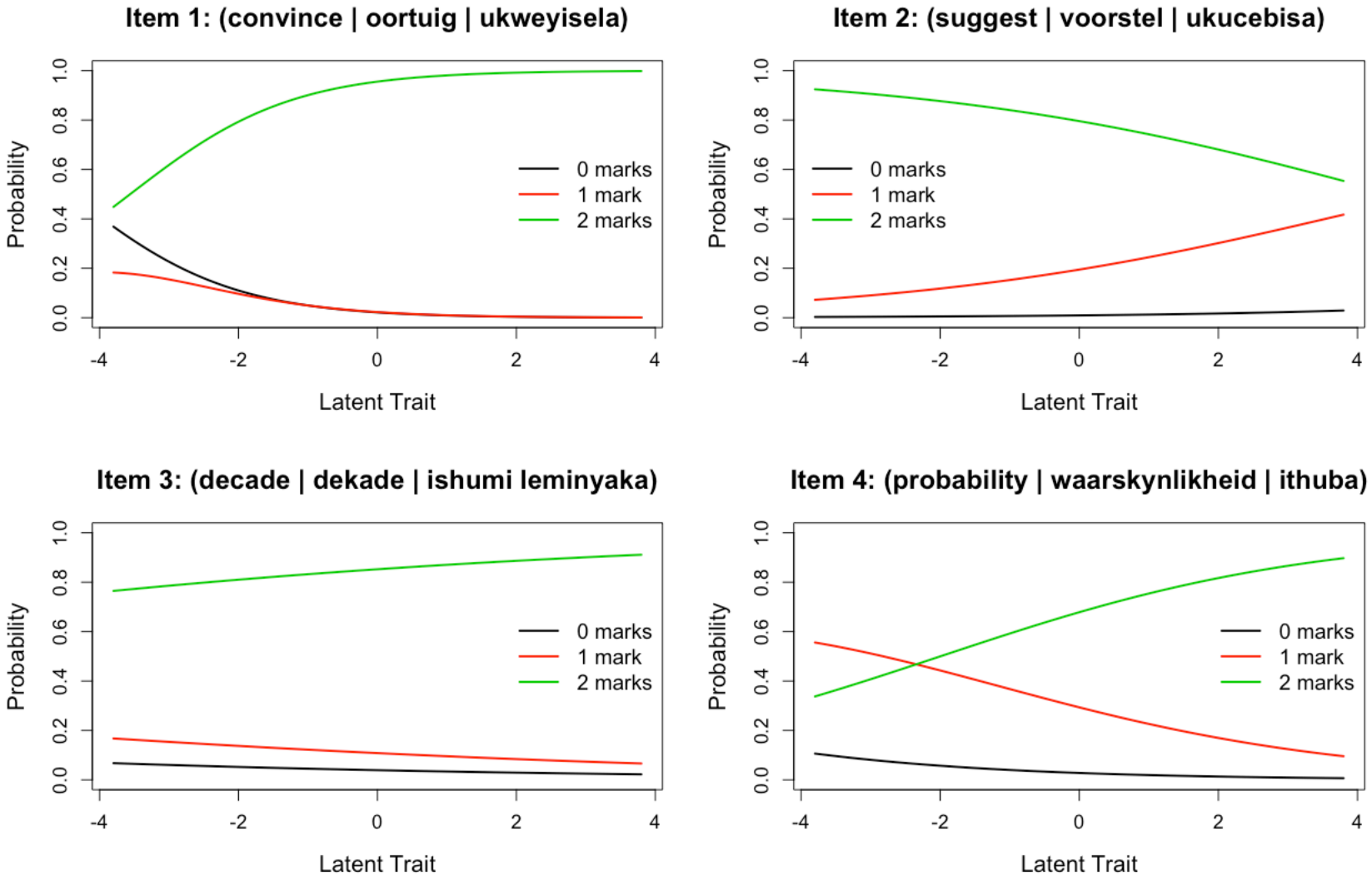
Appendix O
Study 2: Item Difficulty and Item-total Correlations for 24-item MVT ($N = 101$)

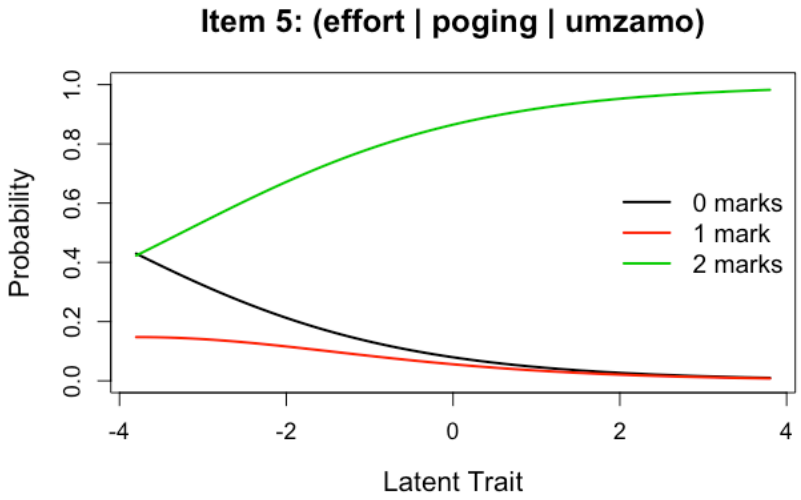
Table O-1
Study 2: Item Difficulty and Item-total Correlations for 24-item MVT (N = 101)

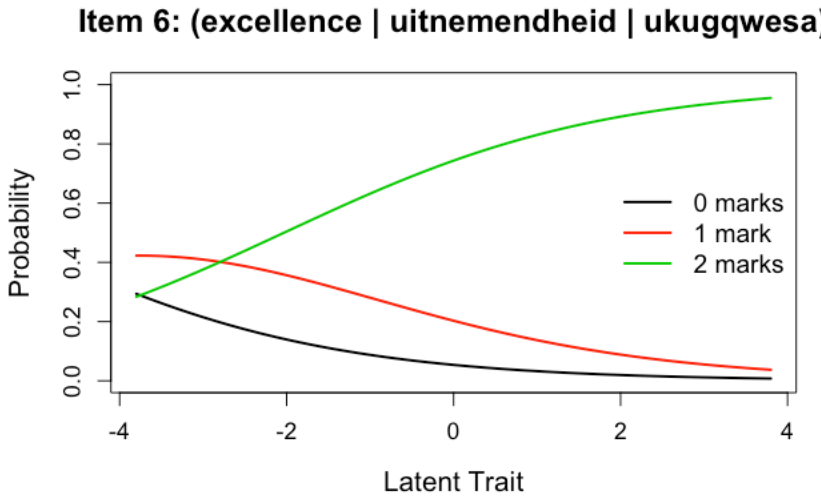
Item (*English* | *Afrikaans* | *isiXhosa*)	Item difficulty		Item-total correlations
	Wide definition	Narrow definition	
Item 1 (convince | oortuig | ukweyisela)	0.97	0.94	0.26
Item 2 (suggest | voorstel | ukucebisa)	0.99	0.79	-0.01
Item 3 (decade | decade | ishumi leminyaka)	0.96	0.85	0.26
Item 4 (probability | waarskynlikheid | ithuba)	0.97	0.67	0.17
Item 5 (effort | poging | umzamo)	0.91	0.85	0.14
Item 6 (excellence | uitnemendheid | ukugqwesa)	0.94	0.73	0.73
Item 7 (dinner | dinee | idinala)	0.98	0,92	0.26
Item 8 (horse | perd | ihashe)	0.96	0.63	0.10
Item 9 (conversion | omskakeling | ukuguqula)	0.08	0.50	0.38
Item 10 (tendency | neiging | isiqhelo)	0.95	0.40	0.42
Item 11 (value | waarde | ixabiso)	0.87	0.83	0.19
Item 12 (habit | gewoonte | umkhuba)	0.95	0.57	0.38
Item 13 (recurrent | terugkerend | -phindaphindayo)	0.97	0.63	0.25
Item 14 (truck | trok | itrakhi)	0.97	0.53	0.09
Item 15 (ambulance | ambulans | i-ambulensi)	0.93	0.01	0.14
Item 16 (announce | aankondig | ukubhengeza)	0.96	0.25	0/30
Item 17 (impetuous | oorhastig | -dyuduzayo)	0.92	0.60	0.18
Item 18 (train | trein | uloliwe)	0.86	0.23	0.18
Item 19 (deliberation | deliberasie | ukucamngca)	0.82	0.58	0.33
Item 20 (tumult | rumoer | isidubedube)	0.75	0.40	0.28
Item 21 (pretentious | pretensieus | ukuzenzisa)	0.47	0.31	0.54
Item 22 (parade | parade | umhambo)	0.73	0.24	0.49
Item 23 (picture | prent | umfanekiso)	0.85	0.30	0.09
Item 24 (atoll | atoll | isiqhiti esisangqa)	0.57	0.20	0.15
Mean	0.89	0.54	0.25

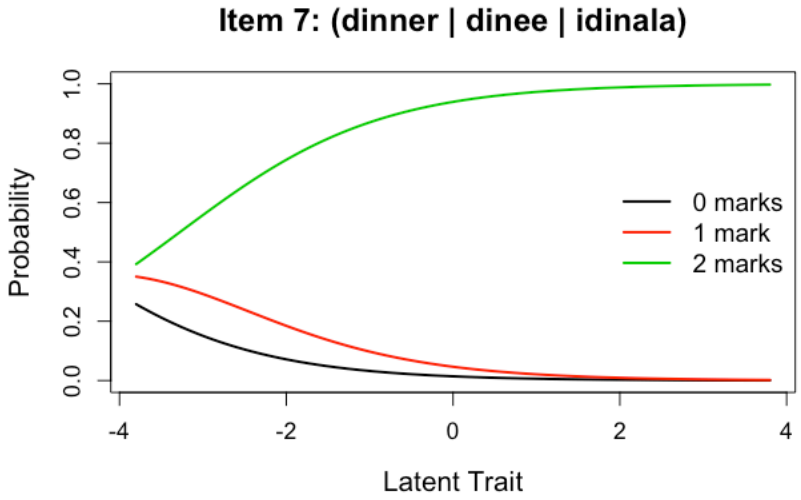
Notes. Item numbers represent the Study 2 administration order. For item difficulty: Wide definition = both 1- and 2-mark responses regarded as correct. Narrow definition = only 2-mark responses regarded as correct. MVT = digital Multilingual Vocabulary Test.

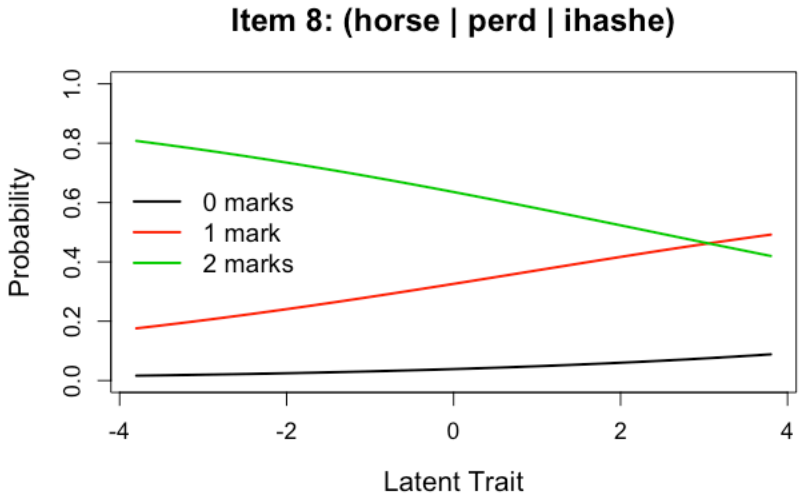
Appendix P
Study 2: IRCCCs for MVT Items 1-24

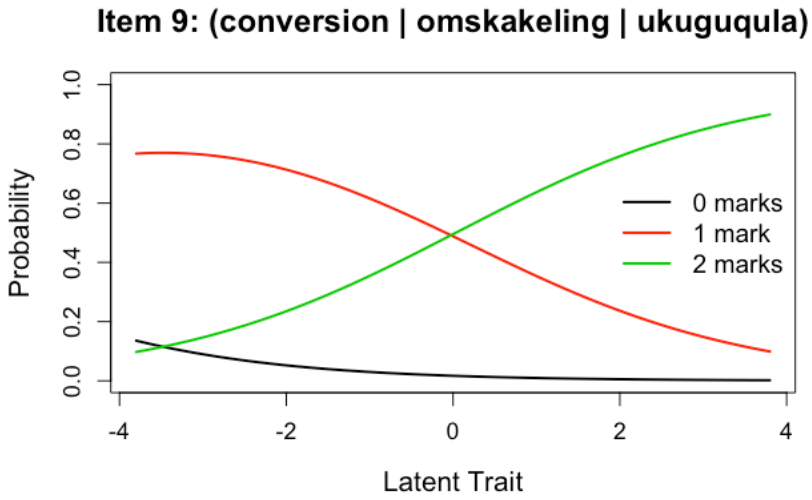


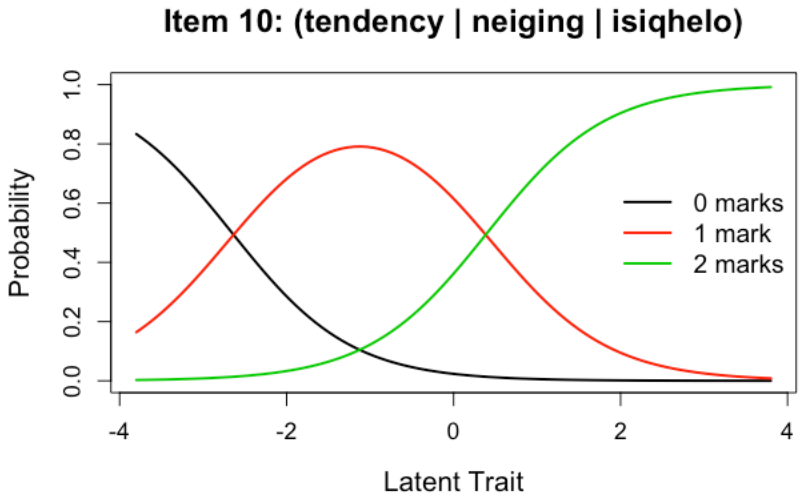

Item 5: (effort | poging | umzamo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-4
-2
0
2
4
Latent Trait

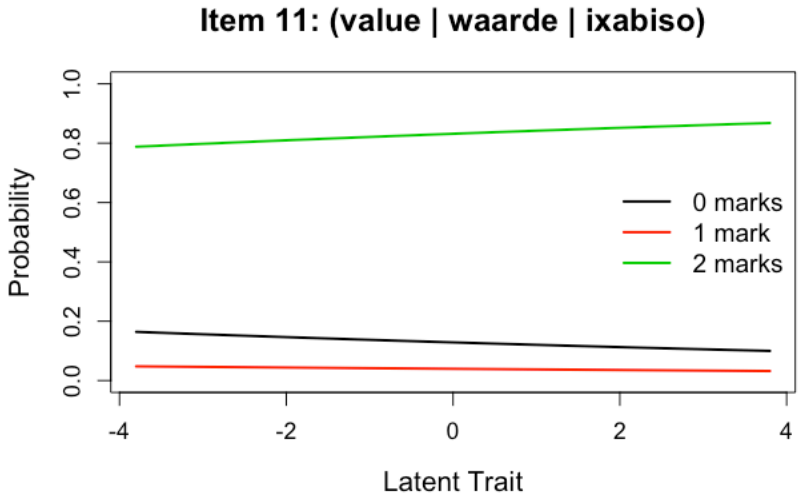

Item 6: (excellence | uitnemendheid | ukugqwesa)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-4
-2
0
2
4
Latent Trait

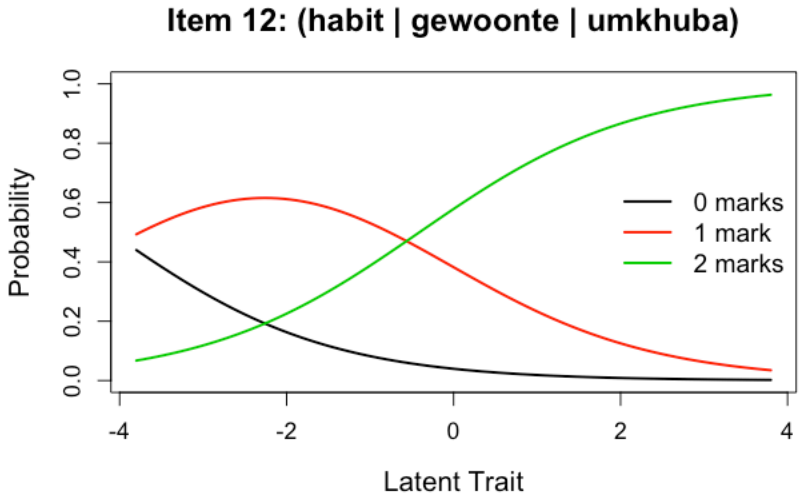

Item 7: (dinner | dinee | idinala)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-4
-2
0
2
4
Latent Trait

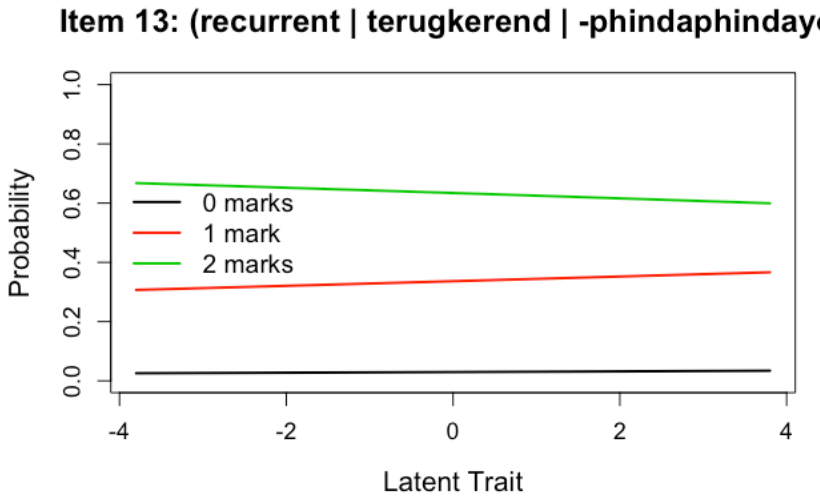

Item 8: (horse | perd | ihashe)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-4
-2
0
2
4
Latent Trait

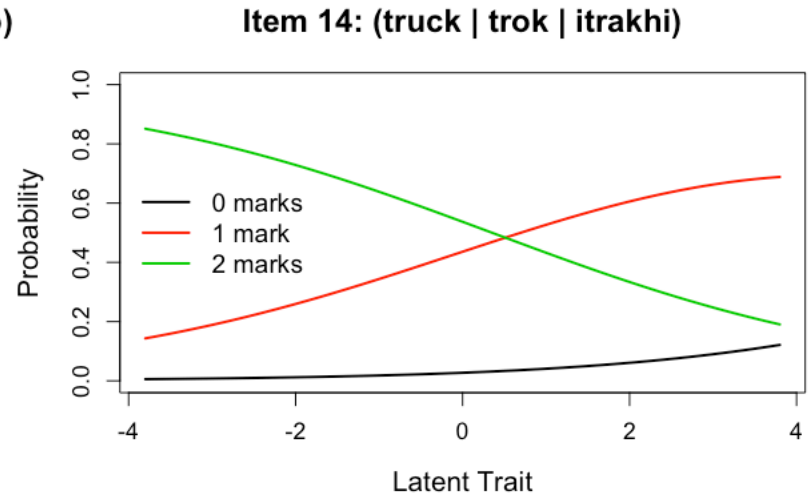

Item 9: (conversion | omskakeling | ukuguqula)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

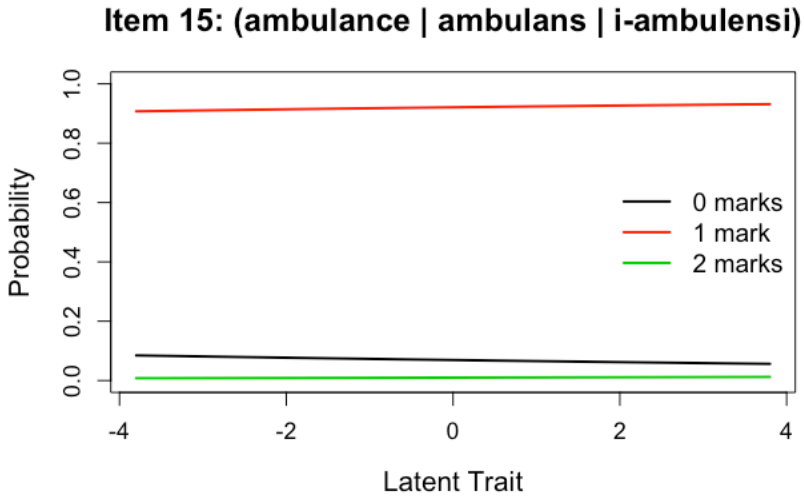

Item 10: (tendency | neiging | isiqhelo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

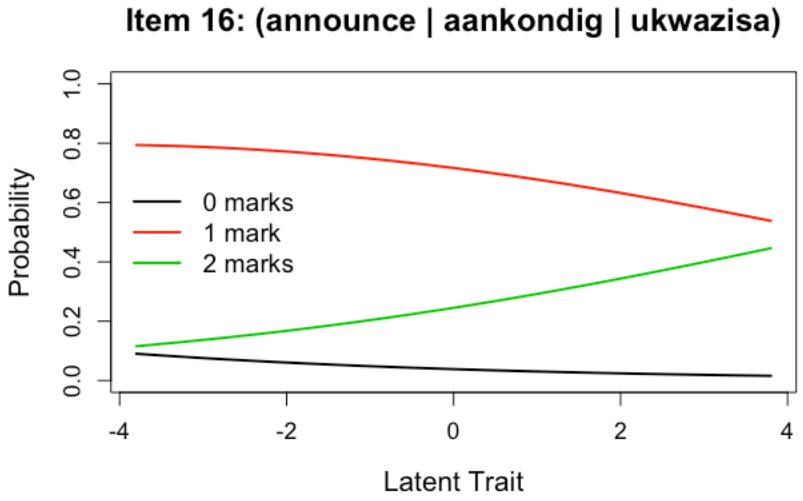

Item 11: (value | waarde | ixabiso)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

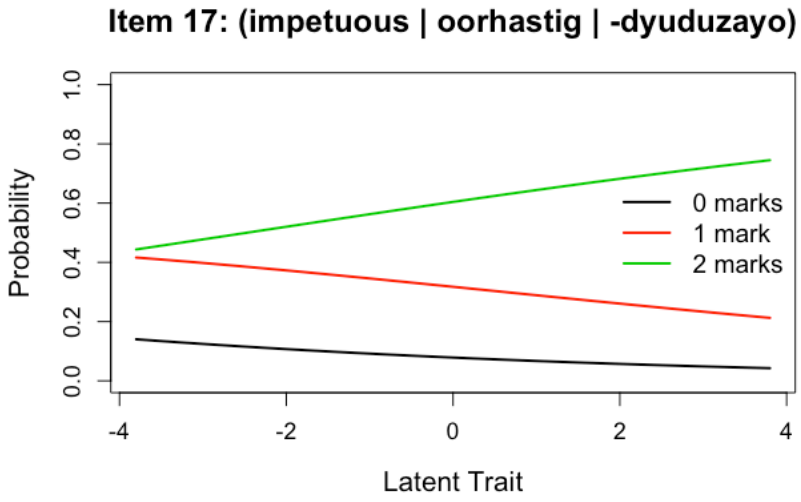

Item 12: (habit | gewoonte | umkhuba)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

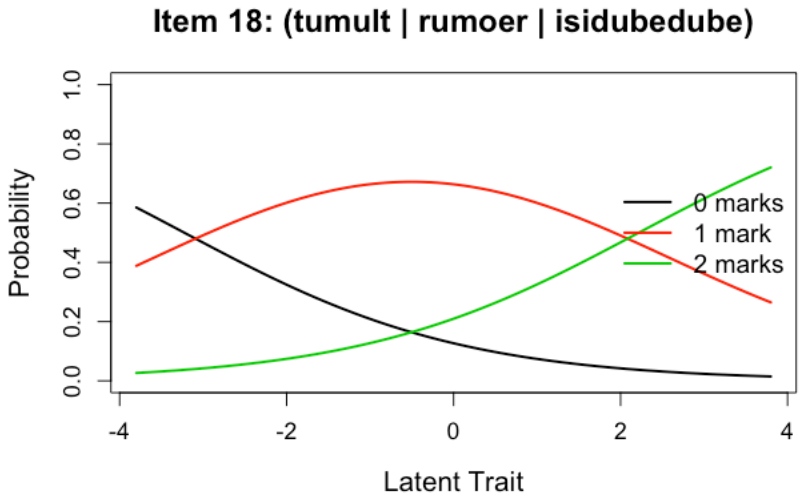

Item 13: (recurrent | terugkerend | -phindaphindayo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-4
-2
0
2
4
Latent Trait

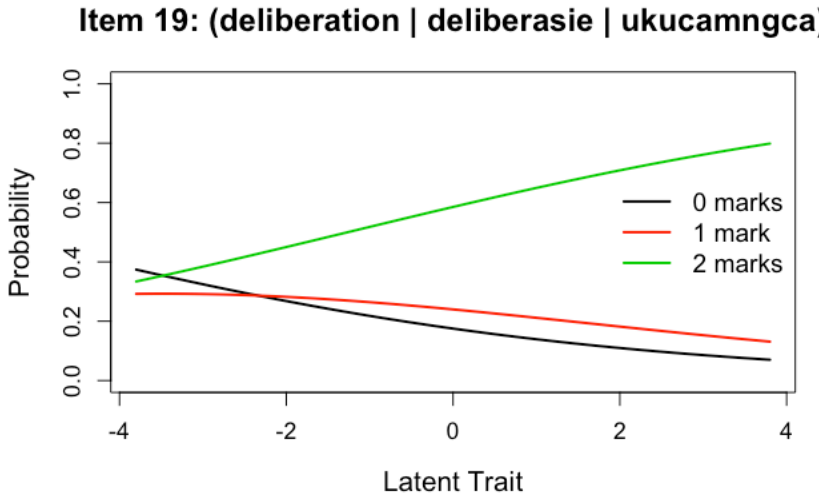

Item 14: (truck | trok | itrakhi)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-4
-2
0
2
4
Latent Trait

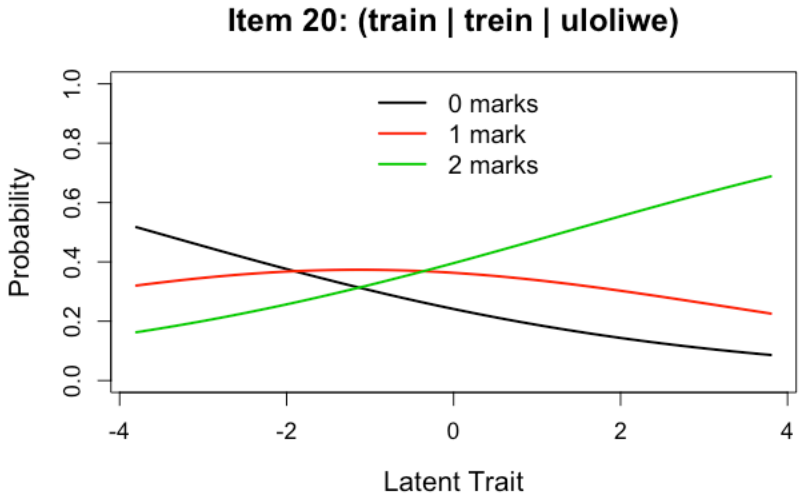

Item 15: (ambulance | ambulans | i-ambulensi)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-4
-2
0
2
4
Latent Trait

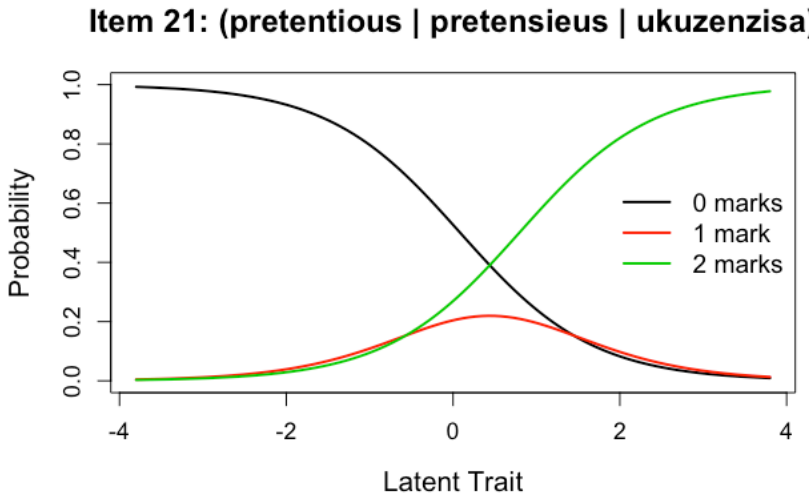

Item 16: (announce | aankondig | ukwazisa)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-4
-2
0
2
4
Latent Trait

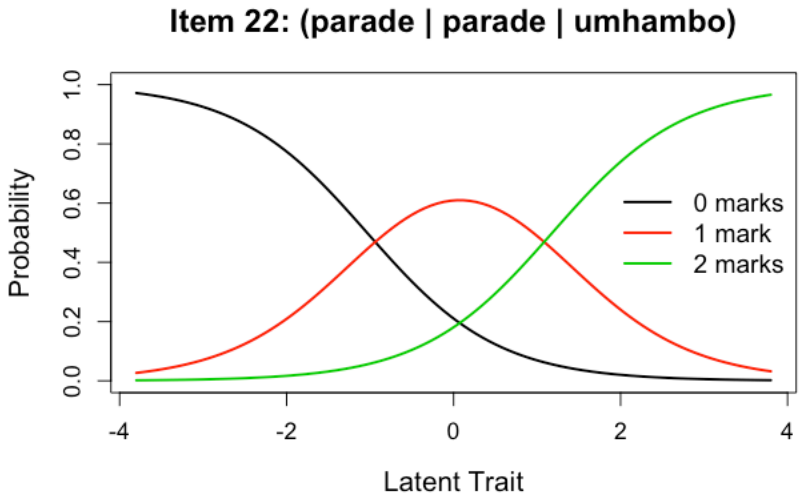

Item 17: (impetuous | oorhastig | -dyuduzayo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

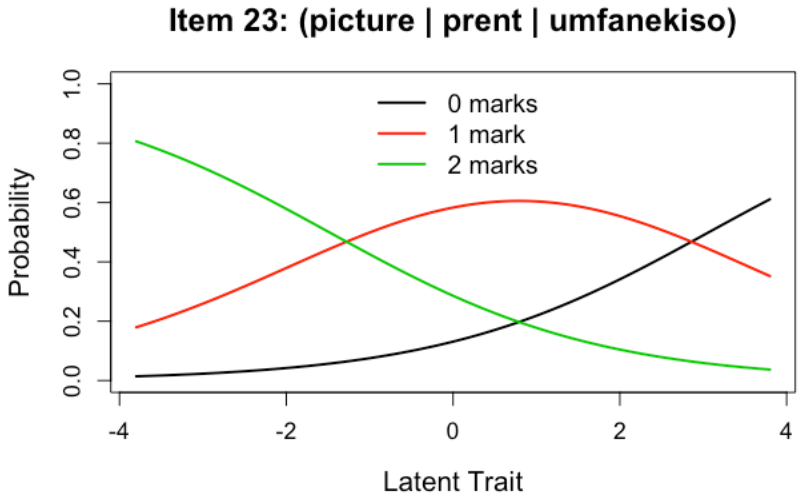

Item 18: (tumult | rumoer | isidubedube)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

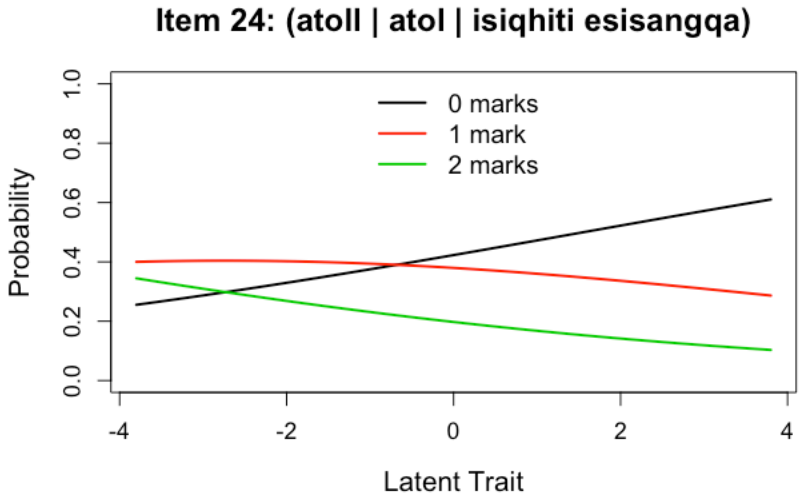

Item 19: (deliberation | deliberasie | ukucamngca)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks


Item 20: (train | trein | uloliwe)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks


Item 21: (pretentious | pretensieus | ukuzenzisa)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks


Item 22: (parade | parade | umhambo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks


Item 23: (picture | prent | umfanekiso)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks


Item 24: (atoll | atol | isiqhiti esisangqa)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

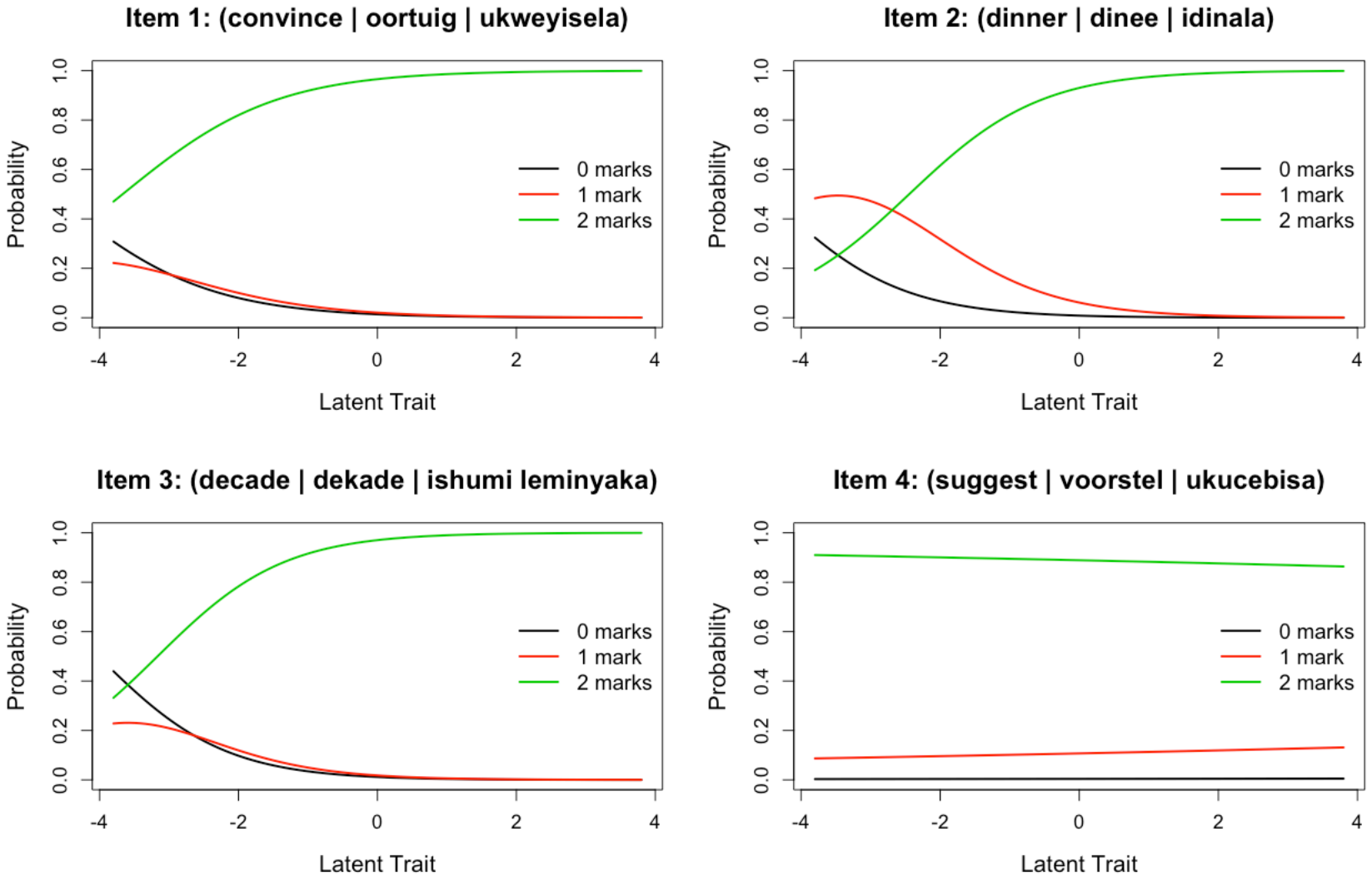
Appendix Q
Study 3: Item Difficulty and Item-total Correlations for 24-item MVT (N = 494)

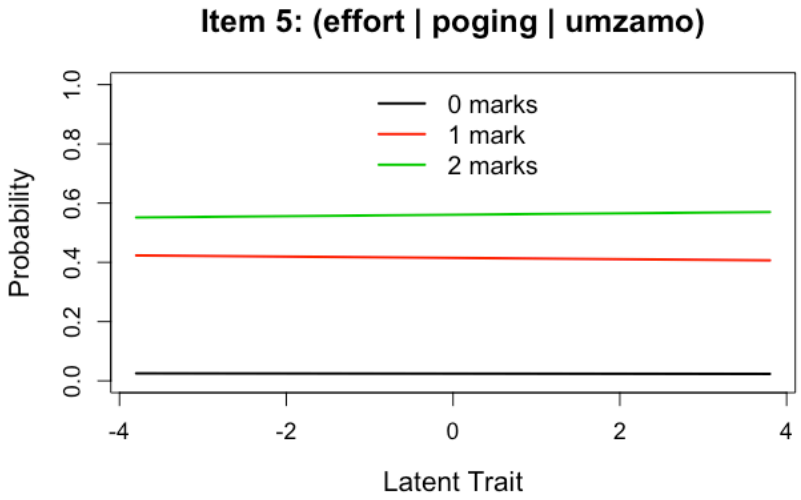
Table Q-1
Study 3: Item Difficulty and Item-total Correlations for 24-item MVT (N = 494)

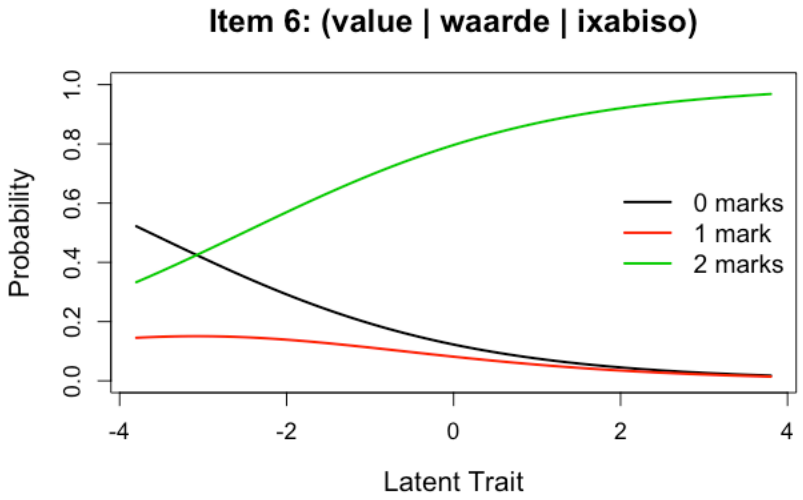
Item (*English* | *Afrikaans* | *isiXhosa*)	Item difficulty		Item-total correlations
	Wide definition	Narrow definition	
Item 1 (convince | oortuig | ukweyisela)	0.98	0.95	0.15
Item 2 (dinner | dinee | idinala)	0.99	0.90	0.27
Item 3 (decade | decade | ishumi leminyaka)	0.98	0.95	0.20
Item 4 (suggest | voorstel | ukucebisa)	1.00	0.89	0.08
Item 5 (effort | poging | umzamo)	0.98	0.56	0.18
Item 6 (value | waarde | ixabiso)	0.87	0.78	0.33
Item 7 (excellence | uitnemendheid | ukugqwesa)	0.94	0.66	0.37
Item 8 (probability | waarskynlikheid | ithuba)	0.96	0.68	0.28
Item 9 (recurrent | terugkerend | -phindaphindayo)	0.97	0.69	0.26
Item 10 (horse | perd | ihashe)	0.85	0.61	0.22
Item 11 (impetuous | oorhastig | -dyuduzayo)	0.92	0.70	0.33
Item 12 (habit | gewoonte | umkhuba)	0.94	0.65	0.25
Item 13 (truck | trok | itrakhi)	0.96	0.62	0.31
Item 14 (conversion | omskakeling | ukuguqula)	0.04	0.43	0.10
Item 15 (deliberation | deliberasie | ukucamngca)	0.85	0.61	0.36
Item 16 (tendency | neiging | isiqhelo)	0.97	0.55	0.44
Item 17 (announce | aankondig | ukubhengeza)	0.96	0.32	0.31
Item 18 (train | trein | uloliwe)	0.37	0.11	0.10
Item 19 (picture | prent | umfanekiso)	0.84	0.18	0.27
Item 20 (tumult | rumoer | isidubedube)	0.89	0.37	0.41
Item 21 (parade | parade | umhambo)	0.88	0.50	0.35
Item 22 (ambulance | ambulans | i-ambulensi)	0.78	0.08	0.03
Item 23 (pretentious | pretensieus | ukuzenzisa)	0.61	0.47	0.49
Item 24 (atoll | atoll | isiqhiti esisangqa)	0.68	0.40	0.39
Mean	0.88	0.57	0.27

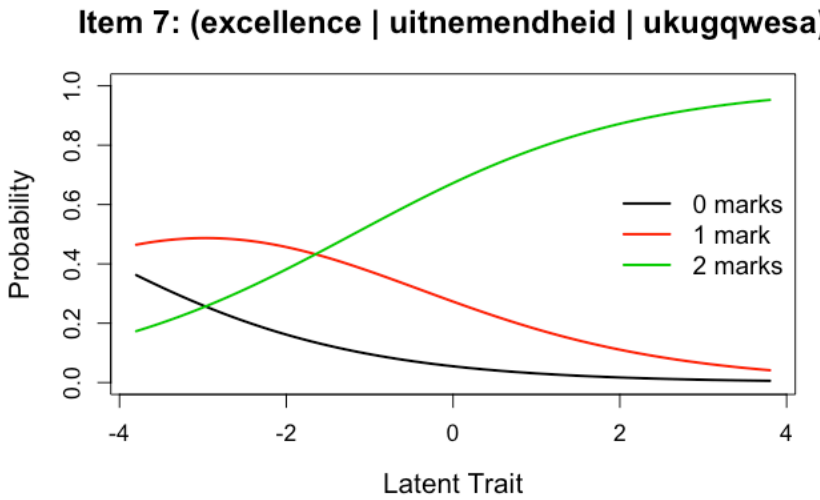
Notes. Item numbers represent the Study 2 administration order. For item difficulty: Wide definition = both 1- and 2-mark responses regarded as correct. Narrow definition = only 2-mark responses regarded as correct. MVT = digital Multilingual Vocabulary Test.

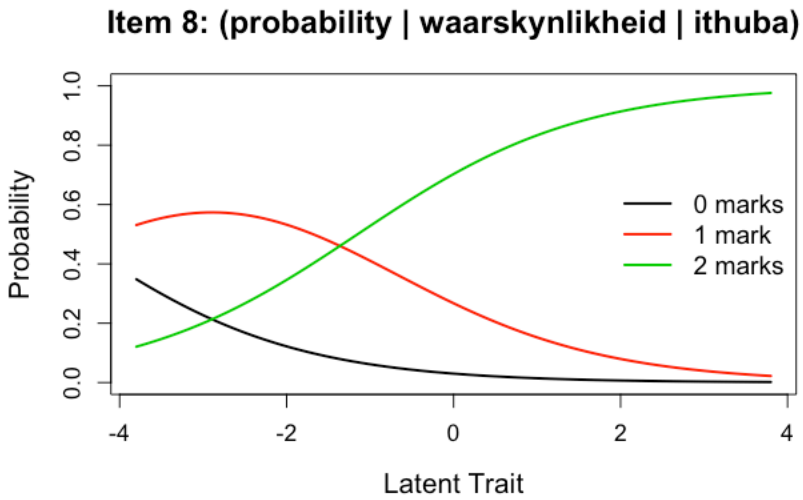
Appendix R
Study 3: IRCCCs for MVT Items 1-24 for Ability Range from -4 to +4 ($N = 494$)

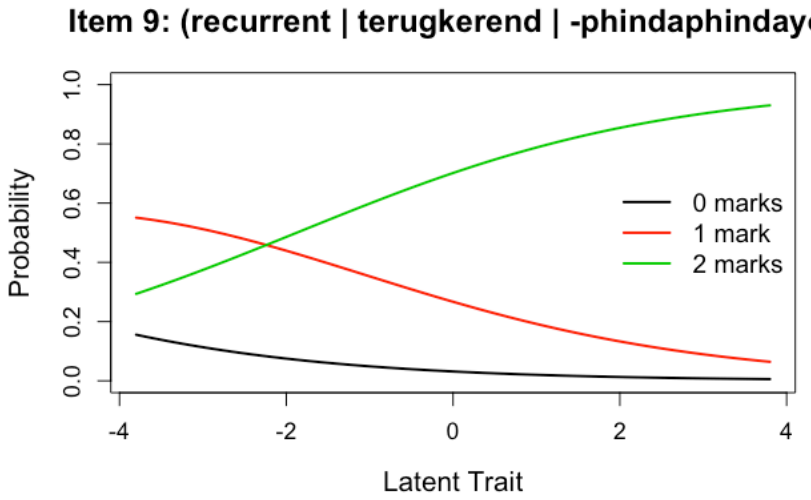


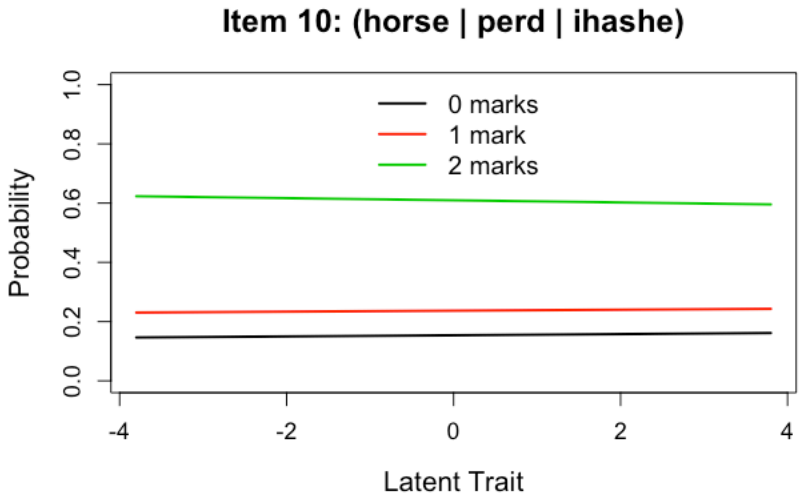

Item 5: (effort | poging | umzamo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-4
-2
0
2
4
Latent Trait

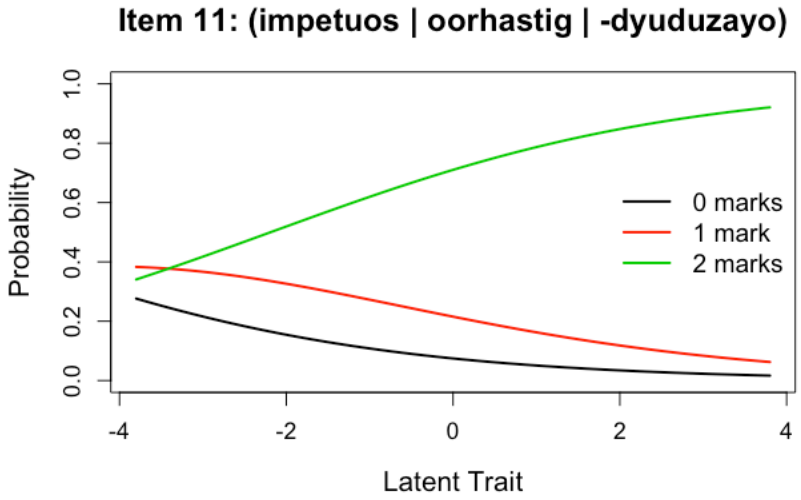

Item 6: (value | waarde | ixabiso)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-4
-2
0
2
4
Latent Trait

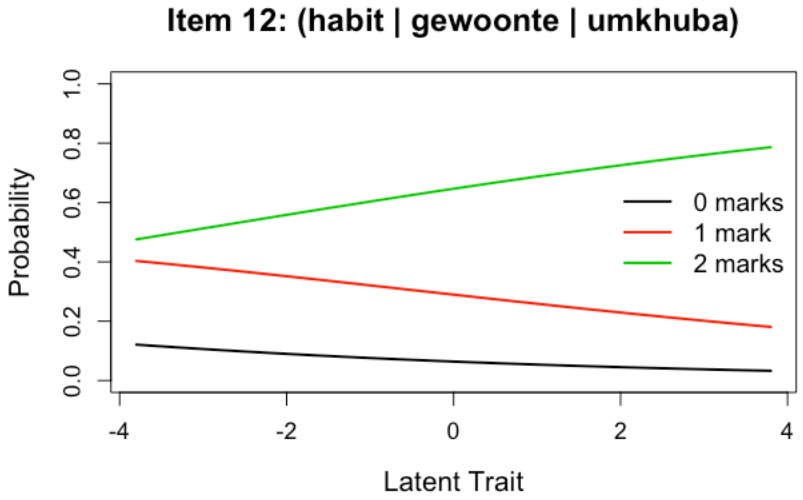

Item 7: (excellence | uitnemendheid | ukugqwesa)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-4
-2
0
2
4
Latent Trait

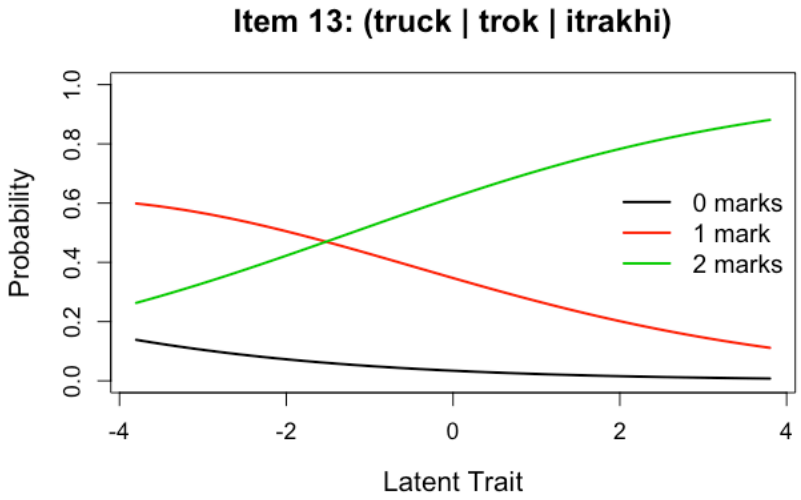

Item 8: (probability | waarskynlikheid | ithuba)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-4
-2
0
2
4
Latent Trait

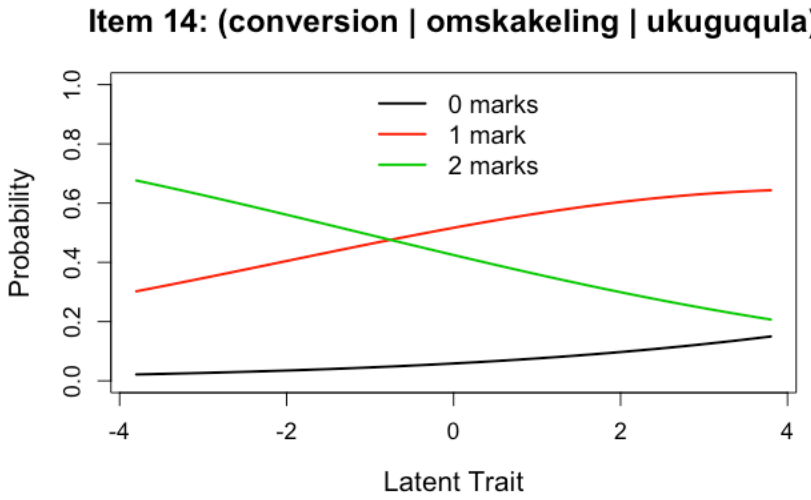

Item 9: (recurrent | terugkerend | -phindaphindayo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

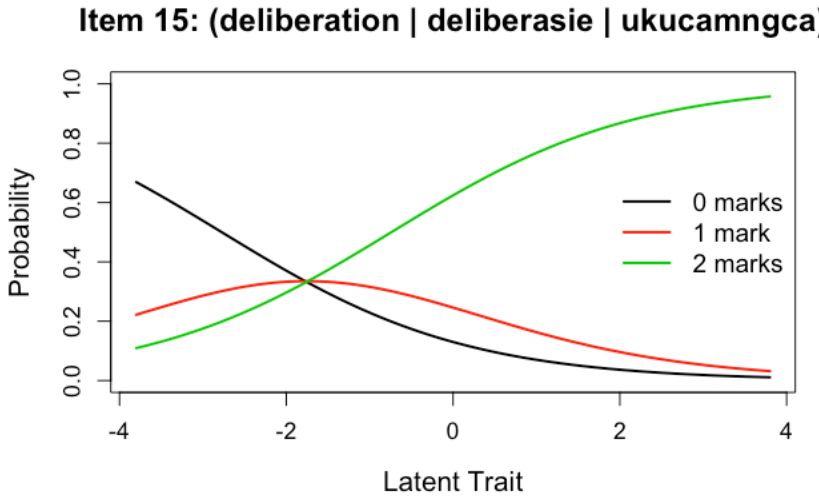

Item 10: (horse | perd | ihashe)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

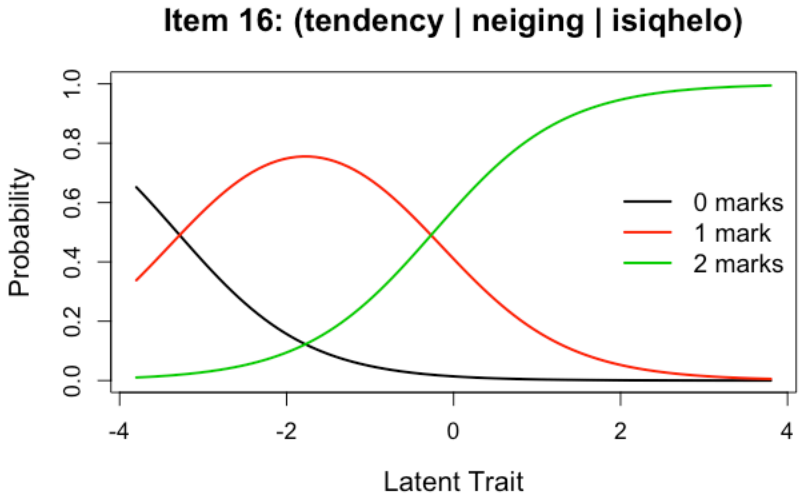

Item 11: (impetuos | oorhastig | -dyuduzayo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

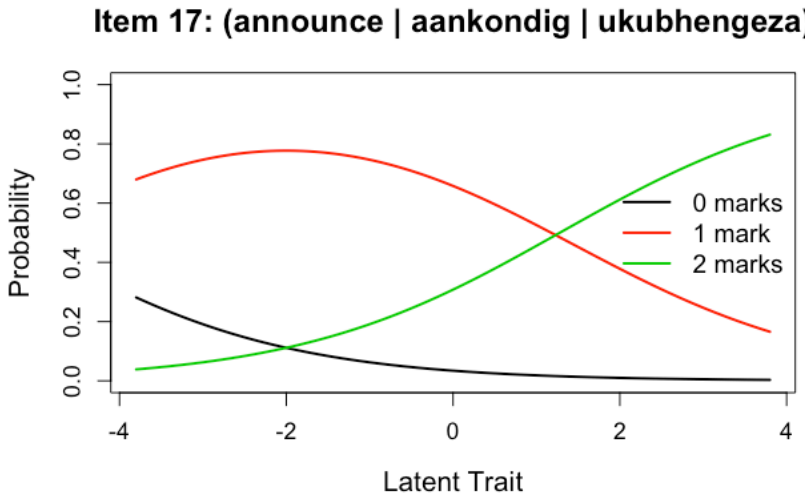

Item 12: (habit | gewoonte | umkhuba)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

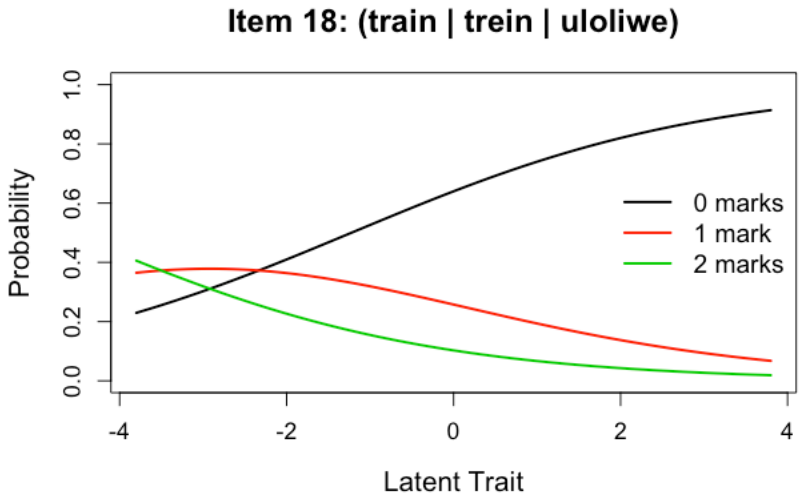

Item 13: (truck | trok | itrakhi)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

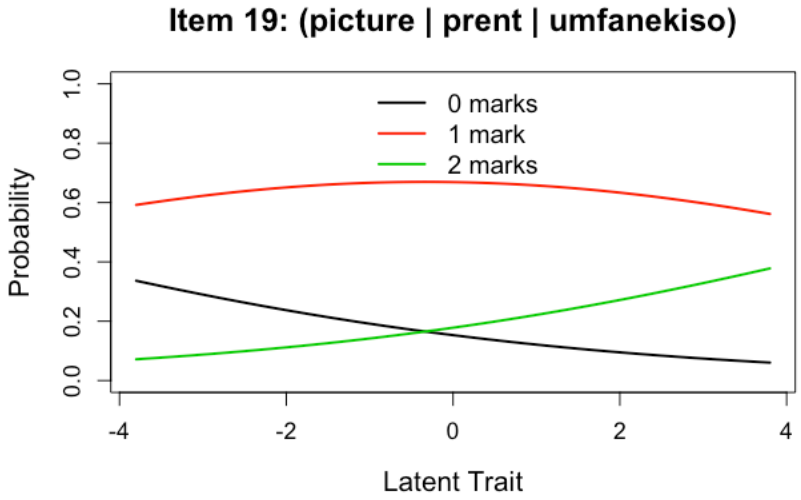

Item 14: (conversion | omskakeling | ukuguqula)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

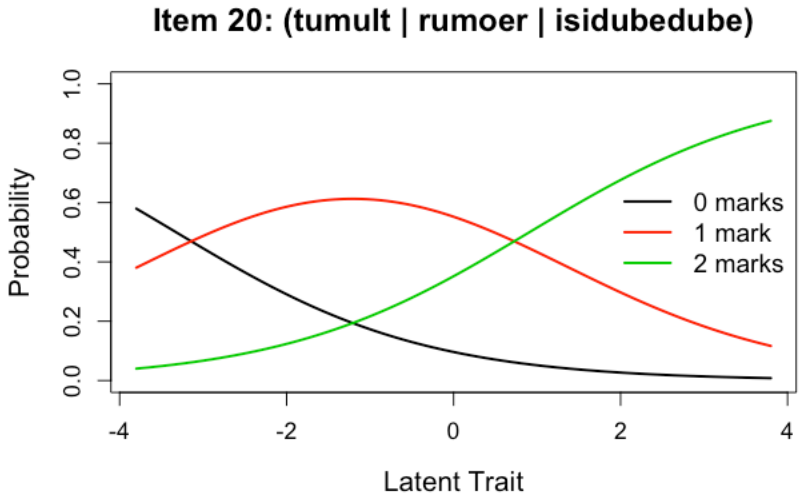

Item 15: (deliberation | deliberasie | ukucamngca)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

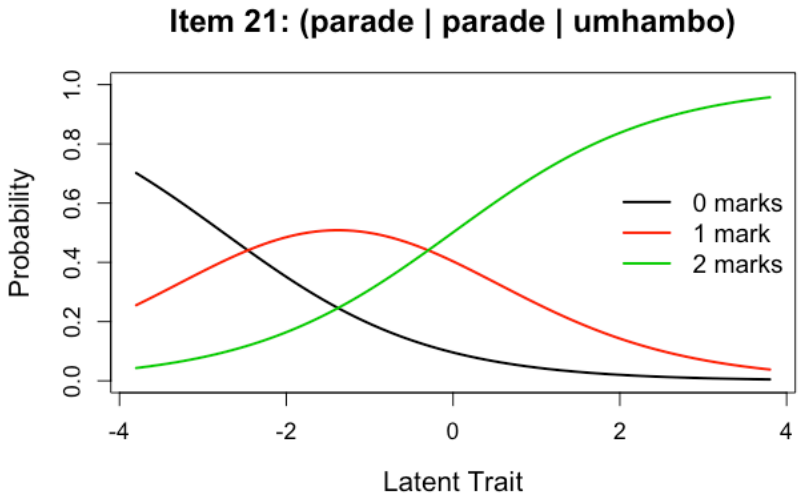

Item 16: (tendency | neiging | isiqhelo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

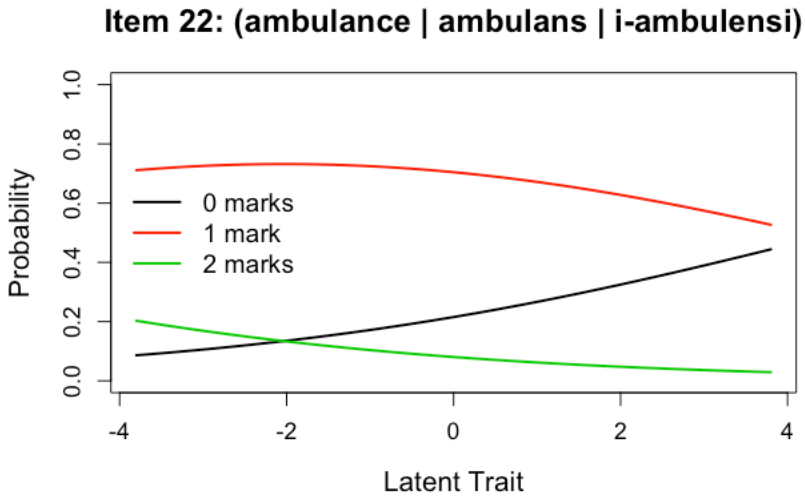

Item 17: (announce | aankondig | ukubhengeza)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

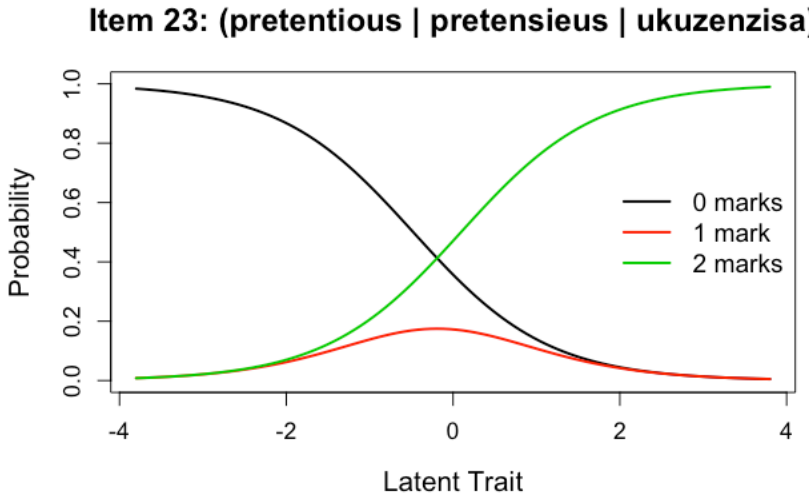

Item 18: (train | trein | uloliwe)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

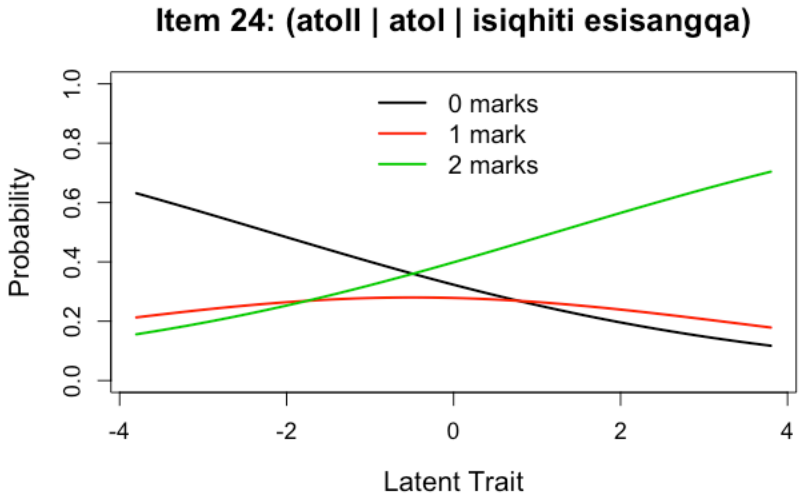

Item 19: (picture | prent | umfanekiso)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks


Item 20: (tumult | rumoer | isidubedube)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

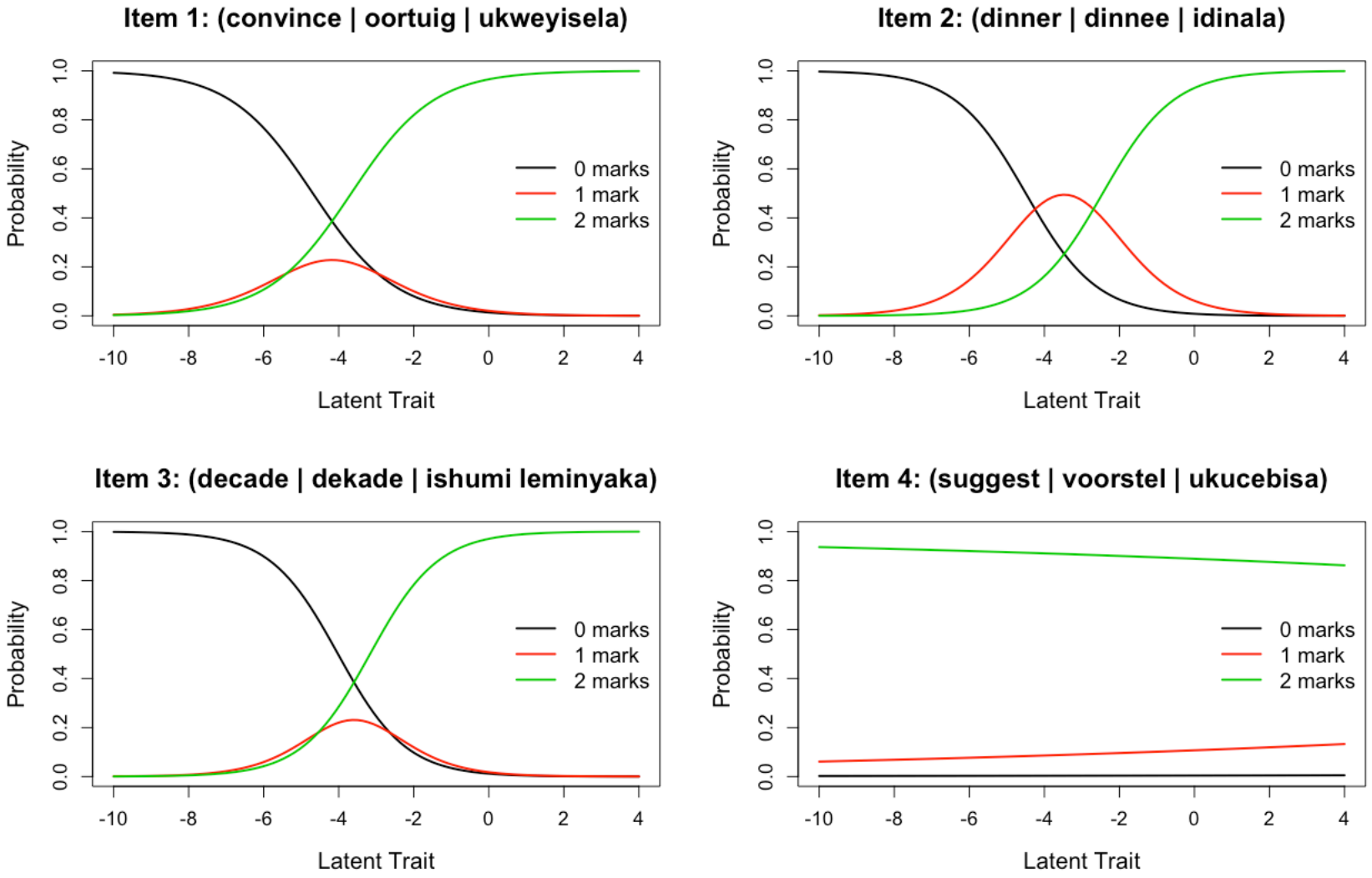

Item 21: (parade | parade | umhambo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

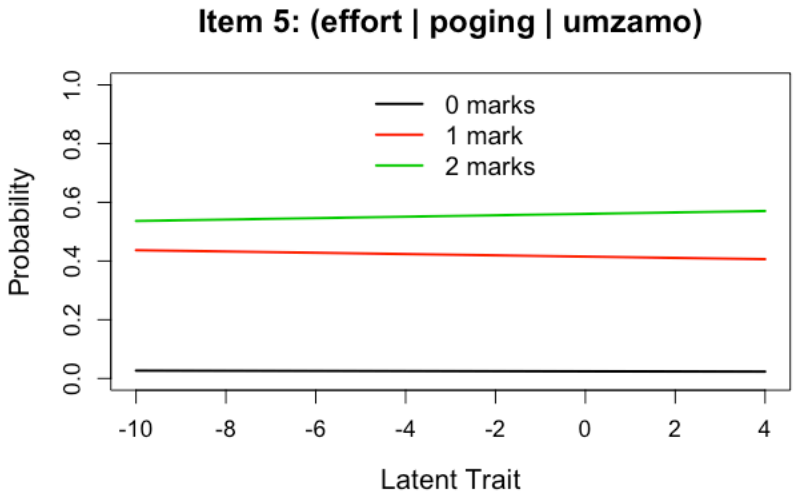

Item 22: (ambulance | ambulans | i-ambulensi)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

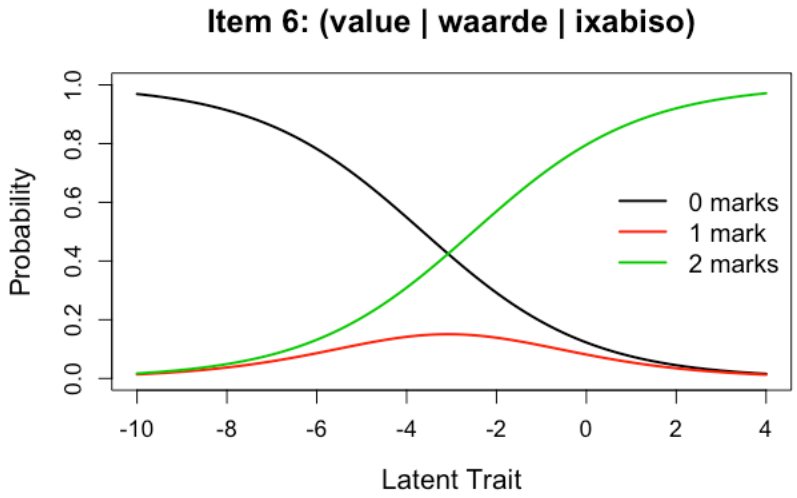

Item 23: (pretentious | pretensieus | ukuzenzisa)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

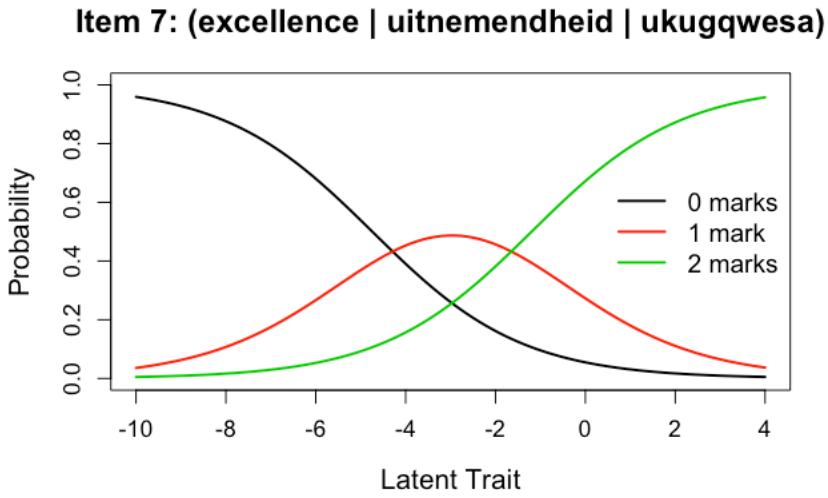

Item 24: (atoll | atol | isiqhiti esisangqa)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

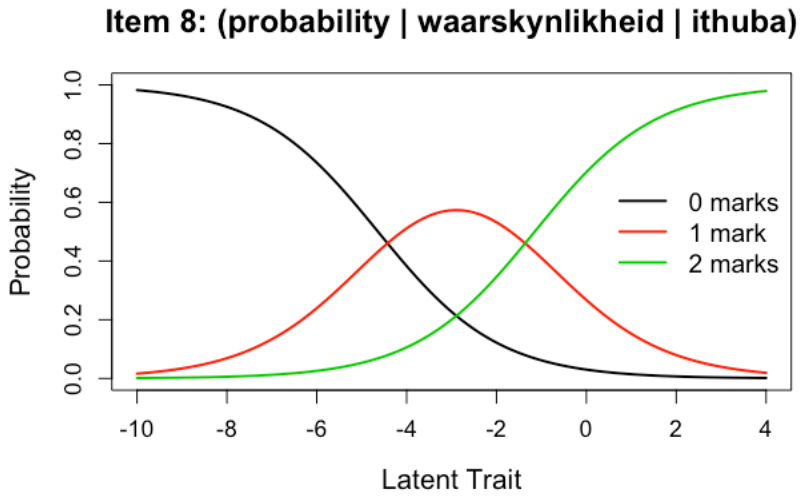
Appendix S
Study 3: IRCCCs for MVT Items 1-24 for Ability Range from -10 to +4 ($N = 494$)

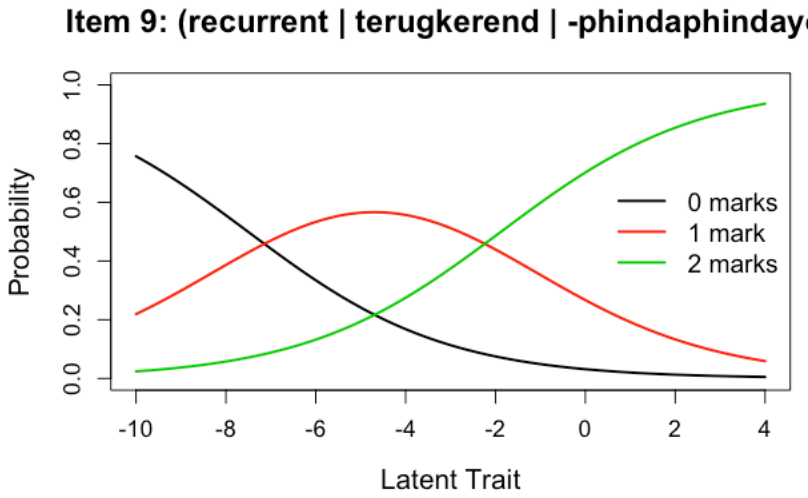


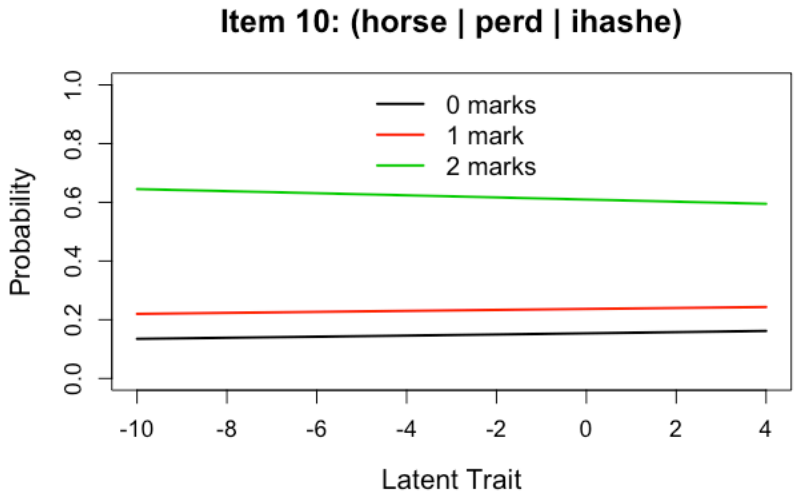

Item 5: (effort | poging | umzamo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-10
-8
-6
-4
-2
0
2
4
Latent Trait

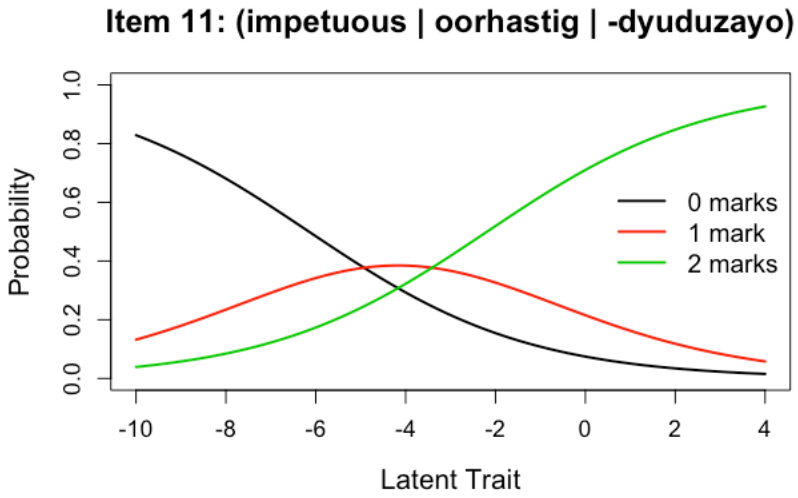

Item 6: (value | waarde | ixabiso)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-10
-8
-6
-4
-2
0
2
4
Latent Trait

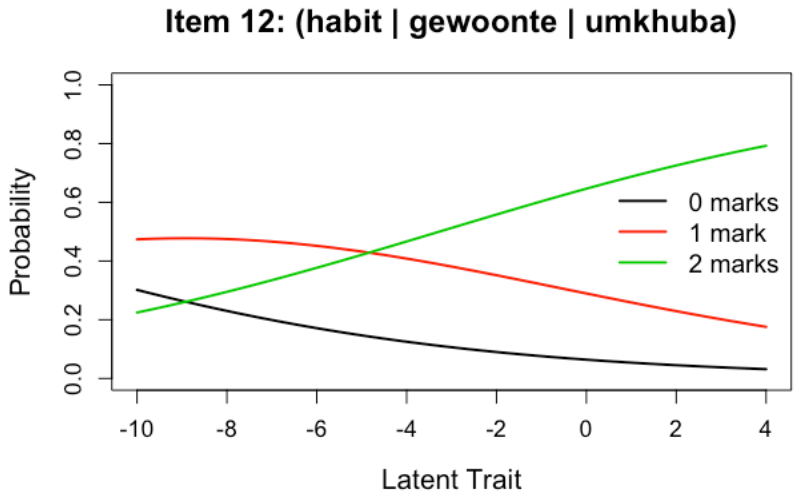

Item 7: (excellence | uitnemendheid | ukugqwesa)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-10
-8
-6
-4
-2
0
2
4
Latent Trait

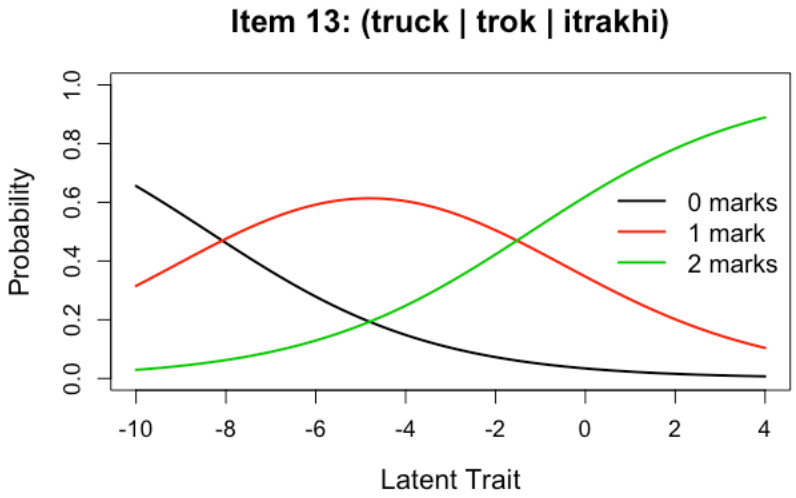

Item 8: (probability | waarskynlikheid | ithuba)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-10
-8
-6
-4
-2
0
2
4
Latent Trait

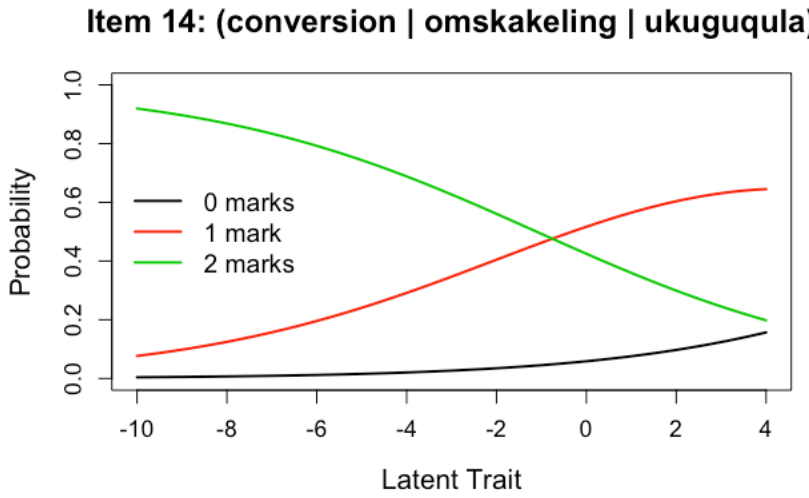

Item 9: (recurrent | terugkerend | -phindaphindayo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-10
-8
-6
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

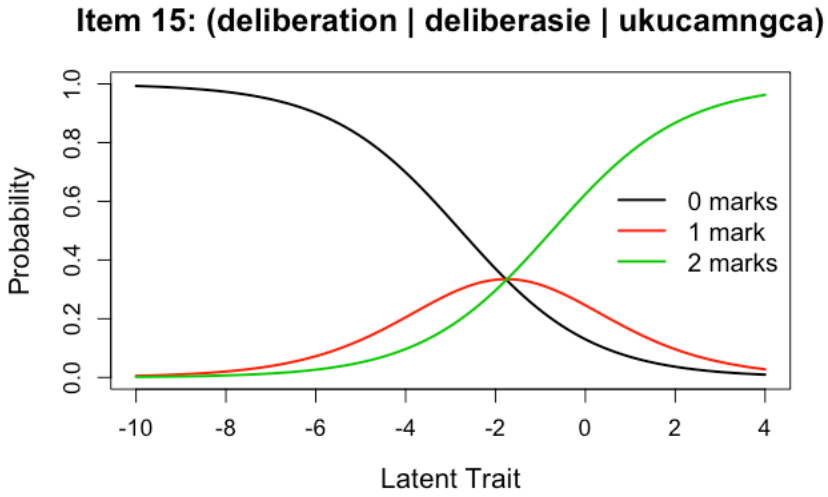

Item 10: (horse | perd | ihashe)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-10
-8
-6
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

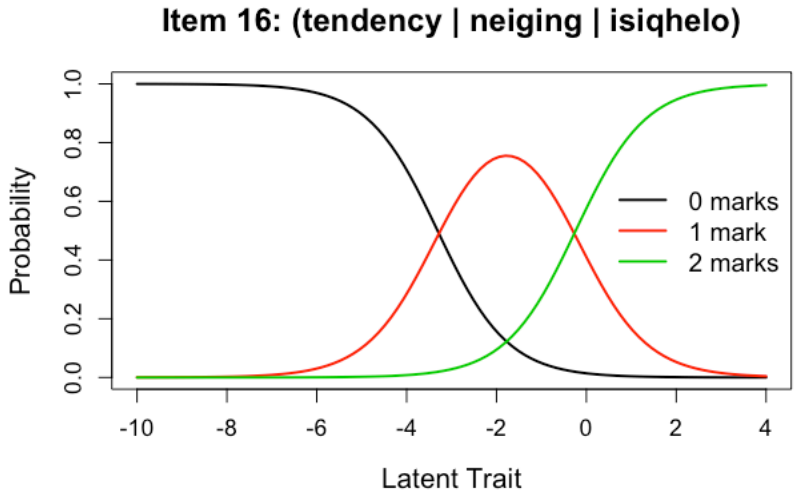

Item 11: (impetuous | oorhastig | -dyuduzayo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-10
-8
-6
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

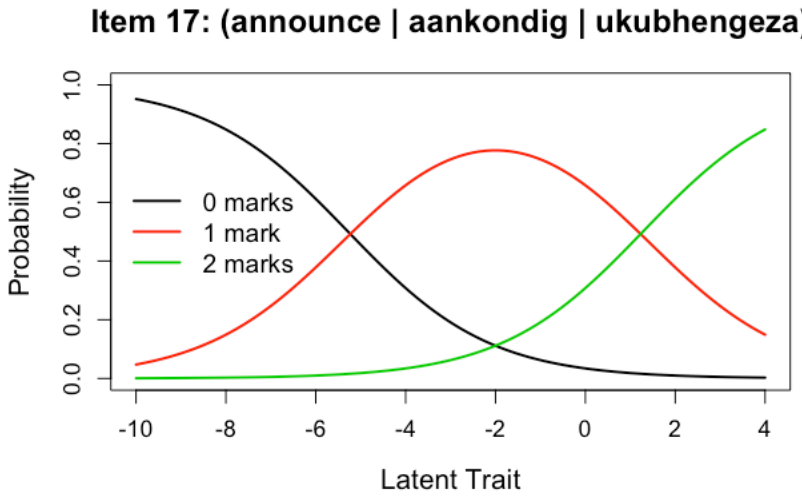

Item 12: (habit | gewoonte | umkhuba)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-10
-8
-6
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

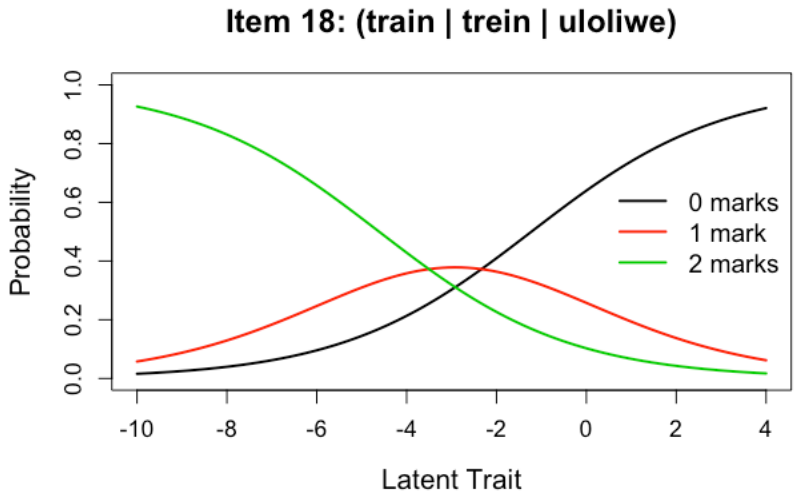

Item 13: (truck | trok | itrakhi)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-10
-8
-6
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

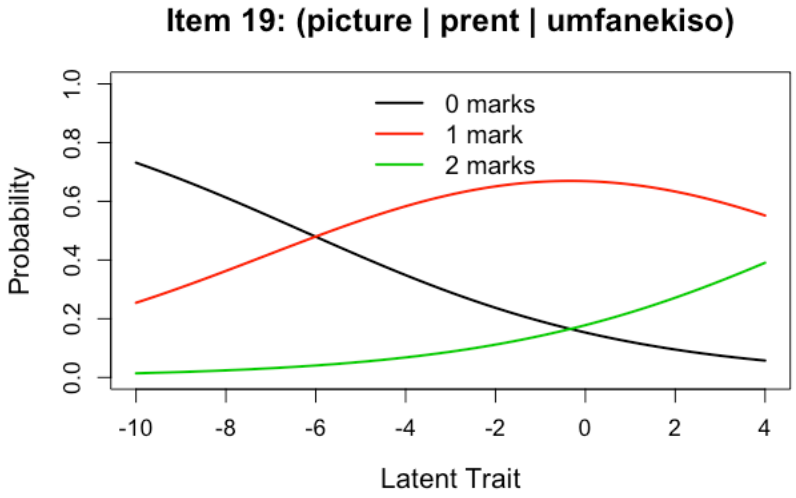

Item 14: (conversion | omskakeling | ukuguqula)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-10
-8
-6
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

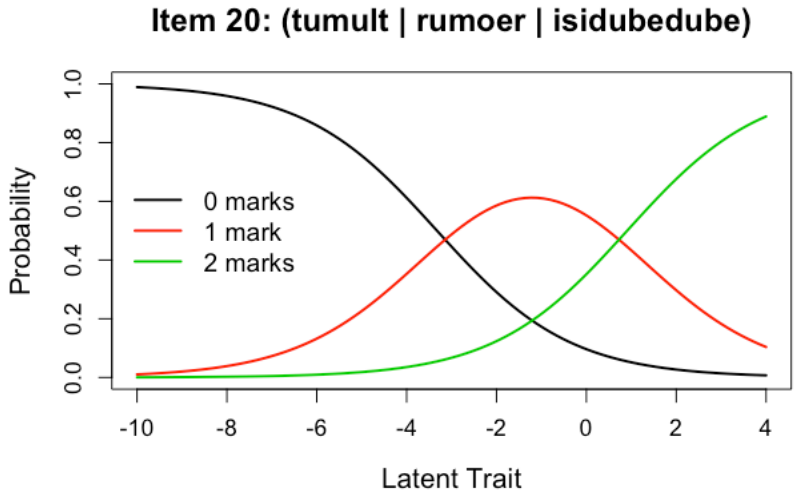

Item 15: (deliberation | deliberasie | ukucamngca)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-10
-8
-6
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

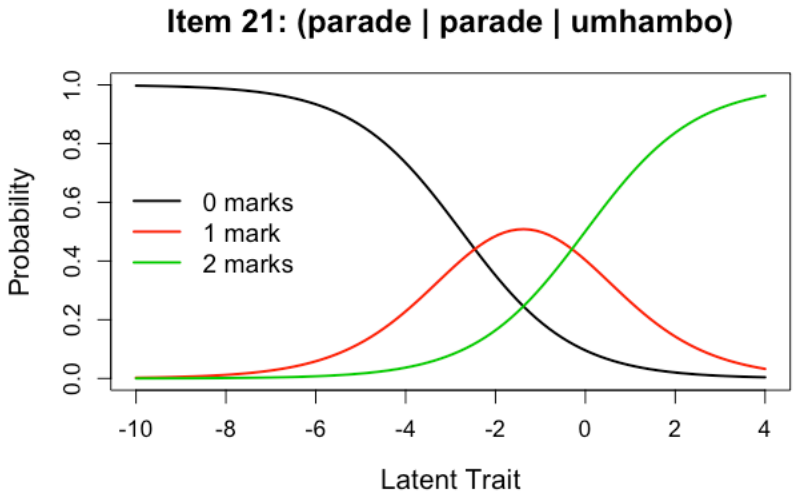

Item 16: (tendency | neiging | isiqhelo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
-10
-8
-6
-4
-2
0
2
4
Latent Trait
0 marks
1 mark
2 marks

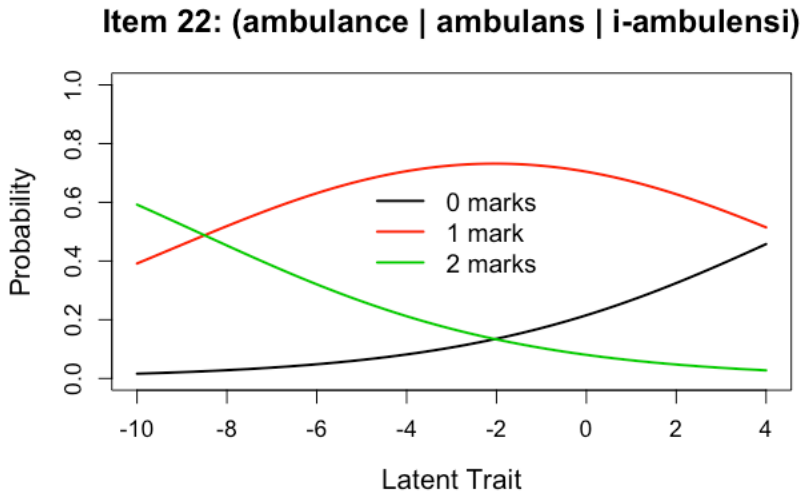

Item 17: (announce | aankondig | ukubhengeza)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-10
-8
-6
-4
-2
0
2
4
Latent Trait

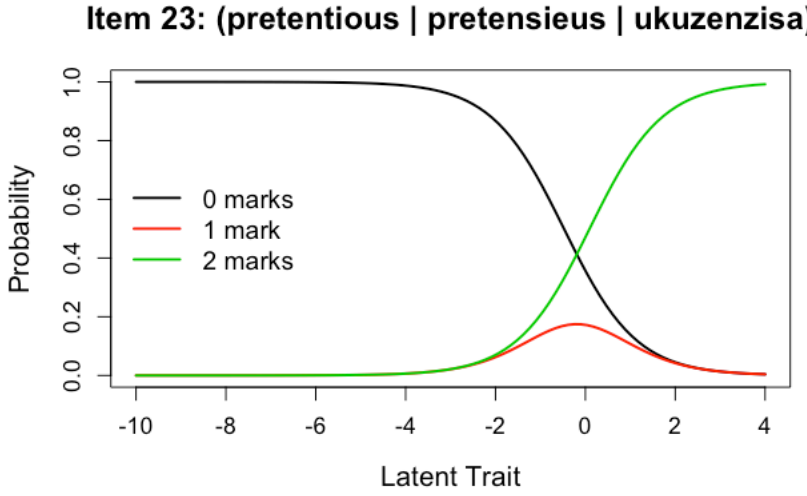

Item 18: (train | trein | uloliwe)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-10
-8
-6
-4
-2
0
2
4
Latent Trait

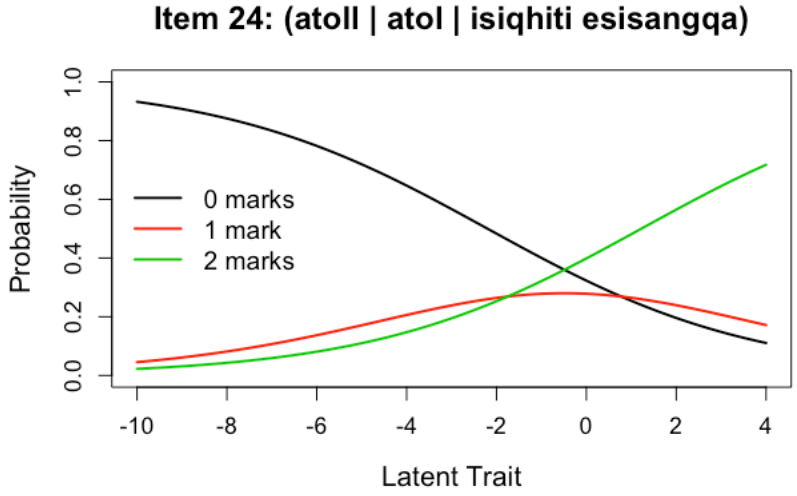

Item 19: (picture | prent | umfanekiso)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-10
-8
-6
-4
-2
0
2
4
Latent Trait


Item 20: (tumult | rumoer | isidubedube)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-10
-8
-6
-4
-2
0
2
4
Latent Trait


Item 21: (parade | parade | umhambo)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-10
-8
-6
-4
-2
0
2
4
Latent Trait


Item 22: (ambulance | ambulans | i-ambulensi)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-10
-8
-6
-4
-2
0
2
4
Latent Trait


Item 23: (pretentious | pretensieus | ukuzenzisa)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-10
-8
-6
-4
-2
0
2
4
Latent Trait


Item 24: (atoll | atol | isiqhiti esisangqa)
Probability
0.0
0.2
0.4
0.6
0.8
1.0
0 marks
1 mark
2 marks
-10
-8
-6
-4
-2
0
2
4
Latent Trait

Appendix T
IICs for MVT items 1-24 in Study 3 (N = 494) for ability range -10 to 4

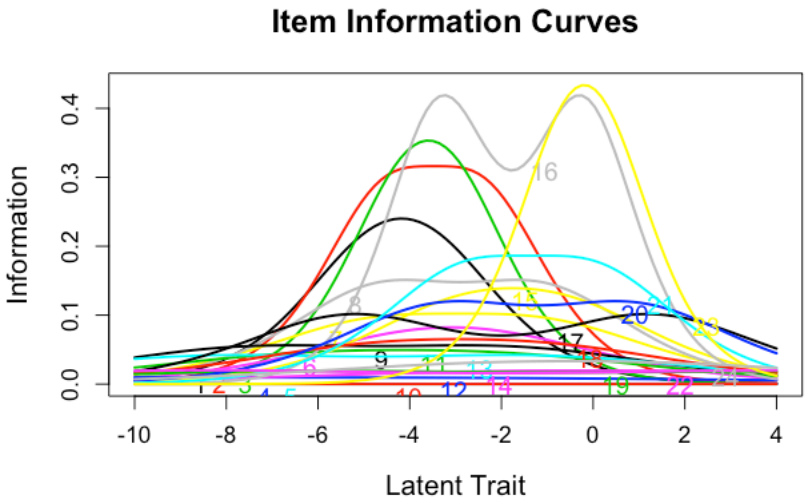


Figure T-1. IICs for MVT items 1-24 in Study 3 (*N* = 494) for ability range -10 to +4.

Appendix U
TIF for the 24-item MVT in Study 3 ($N = 494$) for ability range -10 to +4

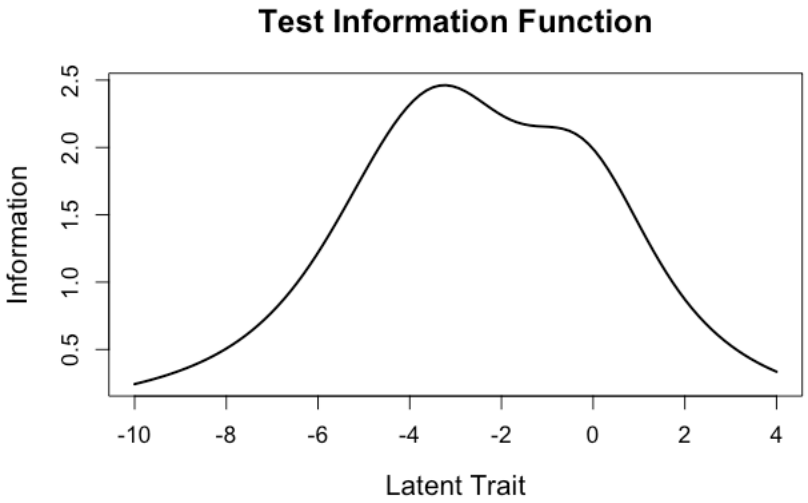


Figure U-1. TIF for the 24-item MVT in Study 3 ($N = 494$) for ability range -10 to +4.