MULTIPLE-CHOICE QUESTIONS: A LINGUISTIC INVESTIGATION OF DIFFICULTY FOR FIRST-LANGUAGE AND SECOND-LANGUAGE STUDENTS

by

Penelope Jane Sanderson

submitted in accordance with the requirements for the degree of

DOCTOR OF LITERATURE AND PHILOSOPHY

in the subject

LINGUISTICS

at the

UNIVERSITY OF SOUTH AFRICA

PROMOTER: Prof EH Hubbard

NOVEMBER 2010

I declare that *Multiple-choice questions: A linguistic investigation of difficulty for first-language and second-language students* is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.

Abstract

Multiple-choice questions are acknowledged to be difficult for both English mother-tongue and second-language university students to interpret and answer. In a context in which university tuition policies are demanding explicitly that assessments need to be designed and administered in such a way that no students are disadvantaged by the assessment process, the thesis explores the fairness of multiple-choice questions as a way of testing second-language students in South Africa. It explores the extent to which two multiple-choice Linguistics examinations at Unisa are in fact 'generally accessible' to second-language students, focusing on what kinds of multiple-choice questions present particular problems for second-language speakers and what contribution linguistic factors make to these difficulties.

Statistical analysis of the examination results of two classes of students writing multiple-choice exams in first-year Linguistics is coupled with a linguistic analysis of the examination papers to establish the readability level of each question and whether the questions adhered to eight item-writing guidelines relating to maximising readability and avoiding negatives, long items, incomplete sentence stems, similar answer choices, grammatically non-parallel answer choices, 'All-of-the-above' and 'None-of-the-above' items. Correlations are sought between question difficulty and aspects of the language of these questions and an attempt is made to investigate the respective contributions of cognitive difficulty and linguistic difficulty on student performance.

To complement the quantitative portion of the study, a think-aloud protocol was conducted with 13 students in an attempt to gain insight into the problems experienced by individual students in reading, understanding and answering multiple-choice questions. The consolidated quantitative and qualitative findings indicate that among the linguistic aspects of questions that contributed to question difficulty for second language speakers was a high density of academic words, long items and negative stems. These sources of difficulty should be addressed as far as possible during item-writing and editorial review of questions.

Key terms: multiple-choice questions, multiple-choice question difficulty, multiple-choice question fairness, first-language university students, second-language university students, assessment validity, item-writing guidelines, readability, cognitive complexity, item analysis

Acknowledgements

Thanks to Prof Hilton Hubbard for unflinchingly whiling away some hours and some years with me on this. And to Prof AP Hendrikse for getting me started and keeping me going, and to all my other colleagues in the Linguistics department at Unisa, for shouldering extra workload along the way. Thanks to Harry Marx for his assistance with obtaining the raw data from the Unisa examination database, to Dr James Ridley for his assistance with the spreadsheets and data analysis, to Chantelle van Heerden and Prof Gherda Ferreira for rating the Bloom levels of the questions, to Oscar Kilpert for assisting with interpreting the Unisa statistics and to Alice Goodwin-Davey, who never runs out of enthusiasm and advice about the art and craft of setting multiple-choice questions. Special thanks to the thirteen interesting and interested students who participated in the think-aloud protocol and to James, Jenny, Trevor, Daniel and Robyn for giving me other interesting things to do with my time.

Table of Contents

Page

Chapter 3 Research design and research methods

Chapter 4 Quantitative results

Chapter 5 Qualitative results

Chapter 6 Conclusions and recommendations

Appendices

List of Tables

List of Figures

Chapter 1
Multiple-choice assessment
for first-language and second-language students

1.1 Introduction

This chapter aims to identify and contextualise the issue that lies at the core of this study and to describe in broad terms the research methods that will be used to investigate it. After a broad overview of issues relating to assessment fairness and validity in South Africa's multilingual context in section 1.2.1, the chapter addresses multiple-choice assessment more specifically in section 1.2.2, focusing on the many interconnected factors that can make a multiple-choice question (MCQ) 'difficult' for both mother-tongue and second-language speakers, and on how university students cope with multiple-choice assessment in a language that is not their mother-tongue. The aims and methodological framework of the study are introduced in 1.3 and 1.4 and the chapter concludes with a structural overview of the thesis in 1.5.

1.2 The focus of enquiry

The focus of this research is to identify linguistic issues that make multiple-choice questions difficult for both English mother-tongue (L1) and second-language (L2) South African university students to interpret and answer. In so doing I hope to contribute to the debate about how the fairness of multiple-choice assessment for testing L2 students can be improved. My research was prompted by concern over how the language of multiple-choice tests might be negatively influencing the large number of L2 students in my own first-year Linguistics course at the University of South Africa (Unisa), and of how to mitigate this risk.

1.2.1 Assessment fairness and validity in the South African university context

Issues of accountability and fairness are a concern for all educators, especially where classes are diverse. Decisions about assessment methods need to be taken with a full understanding of the student profile and of possible implications for different groups of students. In South Africa,

decisions about assessment methods in individual university courses take place within a larger context where dysfunctional apartheid or post-apartheid schooling has failed to prepare many, mostly black, students adequately for higher education (Hay & Marais 2004, Nel, Troskie-de Bruin & Bitzer 2009). This has led to a skewed participation rate whereby black South Africans, who make up 79% of the South African population, are underrepresented in higher education (Statistics South Africa Census 2001). In 2005, for example, the percentage of 20-24-year-olds in higher education was 60% for whites, 51% for Indians, and only 12% for Coloureds and black South Africans (Scott, Yeld & Hendry 2007). Of even greater concern than the skewed participation rate is that student performance continues to be racially differentiated, with black students performing worse than white students in most disciplinary fields (Scott, Yeld & Hendry 2007).

In an attempt to build a more equitable student profile and compensate for declining rates of students who meet the standard entry requirements, South African universities have systematically relaxed their entrance requirements in recent years (Jansen 2003:10). This has resulted in a growing percentage of academically underprepared students. While many South African universities have instituted costly academic support and 'bridging' programmes to try and address the academic deficit (see Hay & Marais 2004), a 2005 Department of Education report showed that 50% of the 120 000 students who enrolled in higher education in South Africa in 2000 dropped out before completing their undergraduate degrees, most of these in the first year, and only 22% completed their undergraduate degrees in the minimum timeframes (*Independent Online* 2008).

An additional challenge for students at tertiary level is the fact that the medium of instruction in most universities is English. While the South African language policy recognises 11 official languages, English is the most dominant language in higher education. Black African students typically speak English as a second, third or fourth language. For example, at my own open distance learning institution, the University of South Africa (Unisa), 77% of the approximately 240 000 students in 2007 were black, and only 27% of the 2007 student body spoke English as a first language (DISA 2010). Three-quarters of Unisa students are therefore learning through the medium of a second language. Since open distance learning involves self-study of primarily

written study material, the reading ability of students is of critical importance, and has a particularly important bearing on the comprehension of long question papers such as MCQs.

Many South African studies show that poor reading ability is the norm rather than the exception (e.g. Pretorius 2000a, Nel, Dreyer & Kopper 2004, see also 2.2.2). Students often lack a sufficiently broad and deep academic vocabulary to cope with university-level reading (Cooper 1995, Hubbard 1996, Schmitt & Zimmermann 2002, Coetzee 2003), resulting in difficulties with reading fluency and comprehension of content-area text (Perkins 1991, Nel, Dreyer & Kopper 2004:98). For example, Pretorius (2000b) found using inferential questions as an indicator of reading ability that, on average, the comprehension level of Unisa first-year Psychology and Sociology students was 53%, coupled with an extremely slow reading speed of 96 words per minute (Pretorius 2002:174). As Pretorius (2002) points out, this places them at the lowest or 'frustration' reading level according to Lesiak and Bradley-Johnson's (1983) three-tiered scale of reading ability based on reading test scores. 'Frustrated' readers have less than 90% decoding accuracy and less than 60% comprehension, resulting in serious reading problems, including an inability to learn effectively from prescribed texts (Lesiak & Bradley-Johnson 1983). Large numbers of South African tertiary students fall into this category.

The reading deficit obviously has its source at the lowest levels of schooling, with many South African primary and high school children reading well below their age levels. For example, the Department of Education's 2003 evaluation of a random 5% sample of Grade 3 learners showed learners achieving a worrying national average of 39% for reading and writing (Read Educational Trust n.d.). In her research on the reading comprehension of black South African Grade 8 L2 readers at two high schools, Strauss (1995) observed that vocabulary was a major stumbling block when reading textbook passages in English and that most of these high school learners could not infer the meaning of unknown words from context, guessing incorrectly more often than not (Strauss 1995:157). Learners experienced grave difficulties comprehending expository text and had extremely poor recall of what they had just read, including an inability to differentiate between unimportant details and main ideas (Strauss 1995). The inevitable result of these reading problems are that students are unable to learn effectively from texts. Students who are unable to use text-based cues to make inferences, work out the meaning of unknown words

and follow arguments are unable to 'read to learn' (Pretorius 2000a, 2002). As Pretorius (2002:190) notes, '[…] this handicap accompanies them through an uncertain scholastic career in primary and secondary school, and even up to tertiary level.'

1.2.1.1 What makes a test 'fair'?

Mindful of the context described above, national education policies and university tuition policies in South Africa are demanding explicitly that assessments need to be designed and administered in such a way that no students are disadvantaged by the assessment process. For example, some of the key criteria that have provided impetus for my study are the following 'quality criteria for distance education in South Africa', developed in 1998 by NADEOSA (the National Association of Distance Education Organisations of South Africa) and spelt out in detail in Welch and Reed (2005:20-43). A glance through these criteria reinforces the importance that current educational policy affords to assessment fairness. Criteria that relate to assessment fairness include the following excerpts:

> **Policy and planning**
> Equal opportunities are ensured for all learners, staff and other clients. (Welch & Reed 2005:21)
> **Learners**
> Research into learners and their needs is a high priority and is used to inform all aspects of policy. (Welch & Reed 2005:22)
> Special needs (for example, physical disability) are considered in the design of course materials, assessment arrangements, and communication with tutors. (Welch & Reed 2005:22)
> **Assessment**
> The level of challenge of the assessment is appropriate for the level of the qualification to which it leads. (Welch & Reed 2005:30)
> In distance education delivery between countries, care is taken that the assessment activities are designed and administered in ways that do not disadvantage learners in a range of contexts. (Welch & Reed 2005:30)
> **Management of the curriculum**
> The examination system, where it is necessary, is reliable and valid. (Welch & Reed 2005:38)

These criteria make it clear that South African higher education institutions need innovative assessments that provide fair opportunities for our diverse learner population. But what makes a test 'fair'? According to McCoubrie (2004:710), a fair test implies a defensible test that has been

compiled with due diligence by the test setter and is viewed as authentic by an examination stakeholder. Fair assessments will include a spread of questions from recall to deep learning, upfront provision of assessment criteria to students, and appropriate pass/fail standard setting.

Carstens (2000) takes a document-design perspective on examination papers as texts, suggesting that fair examination papers need to be designed in such a way that they meet the needs of readers who are reading in a highly formal, time-constrained context where no disambiguating tools are available (Carstens 2000:139). In her opinion, this makes it important that examination papers contain no irrelevant or unnecessarily complicated material, and are at the appropriate level of readability, where readability implies multiple aspects of a text from legibility to logical organisation and clarity of meaning (Mobley 1985, see 2.3 below). Fair tests also require fair questions. Mitchell and Roman (2006:12) believe that '(a)n examination question is fair when it is worded in an understandable way, requires students to respond from a position of competence, guides the students in what is required of them and does not try to trick students into giving the wrong answer'. Research into students' perceptions indicates that students perceive fair assessment as assessment which makes reasonable demands, relates to authentic tasks and realistic contexts, develops a range of skills and has long-term benefits (Sambell, McDowell & Brown 1997).

The many issues mentioned above suggest that test fairness is a multifaceted notion, requiring test-setters to provide clear instructions and make reasonable demands of their students while being constantly aware of the target audience and their reading abilities. The U.S. National Research Council makes the point that

> fairness, like validity, cannot be properly addressed as an afterthought after the test has been developed, administered and used. It must be confronted throughout the interconnected phases of the testing process, from test design and development to administration, scoring and interpretation.
>
> (National Research Council 1999:81, cited in Thompson, Johnstone & Thurlow 2002:6).

Although fairness is not a technical term, it draws on the technical notions of reliability and validity as used in test theory. These two concepts are contrasted below:

Reliability implies that a test gives dependable or consistent scores and that the set of test results is capable of being reproduced under differing conditions or situations (Lyman 1998). Reliability can be affected by a number of issues, including cheating, individual student motivation and health, examination conditions such as examiners giving help or extra time, subjectivity in scoring, and administrative errors in recording scores (Lyman 1998:9). Some of the statistical procedures for estimating reliability are discussed in the next chapter in section 2.5.1.

Validity, on the other hand, means that a test gives us information that is useful for our purposes and is equally able to measure the performance of all students, whatever their gender, race, language background or handicapping condition (Lyman 1998, McCoubrie 2004). 'Construct validity' is the term used to refer to the entire body of accumulated validity evidence about a test. It involves 'an effort to understand the psychological meaningfulness of both the test and the rationale that lies behind the test' (Lyman 1998:13). This requires logical arguments or statistics relating to the face validity of the items (the extent to which the test items appear to be appropriate, for example to lecturers and students), content validity (the relevance of the test's content in relation to the study material) or criterion-related validity (how closely the test results correlate with some standard of performance that is external to the test) (Lyman 1998:9). According to Rupp, Garcia and Jamieson (2001:186):

> current testing theory places construct-related evidence for validity as central to any test design […] To support inferences made on the basis of test scores, a wide array of theoretical rationales and empirical evidence can be gathered, such as examining the test's content, the response patterns of individuals, the relations to other variables, and the test's consequences.

To improve validity, test developers need to review and revise study material and assessment questions to avoid potentially insensitive content or language. According to McCoubrie (2004:711), 'Even carefully designed items benefit from rigorous evaluation and 30% to 60% warrant revision or deletion'. Among the reviews that Haladyna (1994:128) recommends before the test is administered are review with respect to accepted item-writing guidelines (such as those discussed in section 1.2.2.2 below), an editorial review for spelling and grammatical errors, a key check to see that the right answer has been identified and that there is only one right answer, and a content check to see that there is balanced coverage of the course content at an appropriate

educational level. According to Haladyna (1994), a review for bias should also be done to look for any aspect of a test item that treats a subgroup of test takers stereotypically or pejoratively (Haladyna 1994:135). Finally, he suggests that a test-taker review should ideally be done to pilot the test items on a representative sample of students, for example by offering students opportunities to comment orally or in writing on the items or provide rationales for their answers. Test-taker review can be used to detect items that fail to perform as intended, exposing ambiguous, misleading or unintentionally difficult items (Haladyna 1994:136). According to Haladyna (1994), these reviews are not merely an empirical exercise but to a large extent involve expert judgment by competent test developers. He believes that this kind of validation is vital because

> if items are misclassified by content, the key is incorrect, item-writing rules have not been followed, the editorial process has failed, or the items contain some kind of bias, then the response to that item may not be as expected, and inferences we make from test results are weakened. This is the very fabric of validity.
>
> (Haladyna 1994:16)

Validity also implies that test developers also need to investigate intergroup differences in test performance. Ghorpade and Lackritz (1998) urge educators to be sensitive to diversity issues, to study the impact of their examining practices on different students and to take action if their own classes produce different success rates for different social groups:

> Individual faculty members typically have little control over the abilities of their students. They have even less control over the ultimate success of their students. But faculty do have considerable control over the structuring of the learning environment because they typically select the texts and other reading materials, make the assignments, and design and administer the exams. Faculty who are sensitive to equal opportunity concerns should be willing to use this power to ensure that their appraisal methods provide an accurate and fair assessment of student achievement. This concern is particularly relevant in instances in which faculty have found, after self-study, that their methods do in fact result in adverse impact.
>
> (Ghorpade & Lackritz 1998:465)

The current study aims to investigate the impact of two MCQ examinations on L1 and L2 students to find out whether there is adverse impact on L2 students and whether this can be attributed to aspects of the language and readability of MCQs. For assessment to be reliable,

valid and fair, the same high standards need to be met by all. However, Fairbairn and Fox (2009) contend that 'there is evidence that many of the tests that are currently being used are not entirely appropriate for the task of measuring the academic knowledge, skills and abilities of ELLs [English language learners] due to the confounding of language ability and content knowledge' (2009:11). In some circumstances, special arrangements or accommodations may have to be made if the standard test is unsuitable or there are characteristics of learners that will significantly interfere with test performance (Goh 2004:8). The terms 'accommodation', 'modification', 'adaptation' and 'non-standard test administration' all refer to the notion of 'levelling the playing fields' for all students being assessed. Accommodation can involve changing the testing environment, the format, the content or the language of the test, for example by providing Braille question papers for blind students, extra time for dyslexic students or simplifying the language used in tests for English-language learners (see further discussion in section 2.4.4).

While accommodations such as those described above are a way of addressing the special needs of a minority of students, accommodations are not common in contexts where the majority of students are L2 speakers of English who are expected to have or develop functional competence in the language of tuition. One-size-fits-all assessments are the norm at South African schools and universities at present. When accommodations of various kinds are undesirable or impossible, Thompson, Johnstone and Thurlow (2002) advise that testers should apply the principles of 'universal design'. Universal design is a concept from architecture, meaning a design that works for most people without the need for adaptation. Thompson, Johnstone and Thurlow (2002:2) believe this idea applies equally to assessment. Rather than having to adapt existing assessments to suit minority students, students with disabilities, students with limited English proficiency and so on, they advise that new assessments should be developed in a way that allow for the widest participation and enable valid inferences about performance for the greatest number of students. This can be achieved, for example, by ensuring that the items are not biased, that the test instructions are simple, clear and intuitive, and that the test is as readable as possible (see sections 2.3, 2.4 and 2.5).

However it is defined, the discussion above indicates that fairness is a multifaceted notion, drawing on traditional notions of reliability and validity but also touching on social, ethical and linguistic questions. Issues of fairness in relation to the selection, design and development of multiple-choice assessment are taken up in the following section.

1.2.2 Multiple-choice assessment

Multiple-choice assessment is a common assessment option in very large-scale tests such as the U.S. Scholastic Achievement Tests (SATs) and some papers in the South African matriculation (school–leaving) examination, which have considerable influence over student certification and future placement. It is also commonly used at higher education institutions, especially at lower undergraduate levels.

Considerable research efforts have also been invested in exploring the comparative validity of multiple-choice and written assessment (e.g. Curtis, De Villiers & Polonsky 1989, Bennett, Rock & Wang 1991, Bennett & Ward 1993, Hancock 1994, Kniveton 1996, Miller, Bradbury & Lemmon 2000). While this issue is not addressed directly in my study, the literature indicates that the cognitive knowledge assessed by MCQs appears to predict and correlate well with overall competence and performance (McCoubrie 2004:709), although students generally perform slightly better on MCQ tests than on written tests (Miller, Bradbury & Lemmon 2000:168, Struyvens, Dochy & Janssens 2005). For example, Kniveton's (1996) research showed that UK students scored an average of 11% higher for MCQs than for essay tests, which he attributes to the fact that essay marks tend to cluster in the range 30-75% while MCQ marks can range from 0-100%, and that some MCQs can be right due to guessing.

Some of the other questions about MCQs that have been extensively researched over the last 60 years include: What can and can't be assessed using MCQs? How does the performance of boys and girls on MCQs compare? What is the effect of guessing on reliability and validity? More recent questions receiving research attention include: How does MCQ assessment affect the way teachers teach and students learn? How can MCQs be used to diagnose learning difficulties, e.g. by analysing patterns of responses? How can computers be used to offer better, more individualised MCQ tests? None of these issues will be the focus of the current study. However,

the study will contribute to the growing body of literature that attempts to answer questions such as: How should MCQs be written? What makes an MCQ difficult? How do L2 students fare with multiple-choice as opposed to L1 students? What strategies do students use when answering MCQs? Are particular question types such as 'Which of the following are false?' harder than others? and Does the readability of a MCQ affect its difficulty?

While multiple-choice assessment is used at most universities, it is particularly necessary at the University of South Africa (Unisa), which has approximately 240 000 students studying mostly part-time through open distance learning. Class sizes of several thousand students are not uncommon at Unisa and reliance on slow postal systems for submitting and returning handwritten assignments necessitates rapid marking turnaround times. Multiple-choice questions are a pragmatic choice for both assignments and examinations in this context as they allow computer administration (with or without immediate feedback), cost-effective, efficient and accurate marking by computer or optical reader, and, as a result, assessment at more frequent intervals than other forms of assessment.

Other recognised advantages of MCQ assessment are that it enables a test to cover a wider spectrum of the course content, than, for example, an essay on a single topic. This enables a more comprehensive evaluation of students' knowledge and thereby improves the reliability of the test (Brown, Bull & Pendlebury 1997:84). Objective scoring of answers also eliminates the possibility of teacher bias. Brown, Bull and Pendlebury (1997) suggest that computer analyses of MCQ scores can be undertaken, for example to identify which areas of the curriculum particular students are struggling with. Another important advantage of MCQs is that automated statistics can be generated and scrutinised before the results are finalised (see further discussion in section 2.5).

Disadvantages of multiple-choice assessment that have been identified include time-consuming question design that requires a considerable degree of expertise on the part of the test-setter and the failure of this format to allow for learner-centred perspectives, including other possible right answers and rationales (Fellenz 2004). Stiggins (2005) points out that multiple-choice assessment is not really objective except in the scoring, which is either right or wrong. In his view, it is highly subject to the professional expertise of the test setter's decisions about what to

test, how many questions to set, what the correct answers are, what incorrect answers are offered as choices, and so on (Stiggins 2005:85). Importantly, multiple-choice assessment is only appropriate when students have the reading proficiency to understand the test items.

MCQ assessment has also come in for a considerable amount of criticism from detractors who claim that it is predominantly used to test the simplest cognitive abilities, such as recall and comprehension (Bowman & Peng 1972, Frederiksen 1984, Messick 1987) and that it can lead to superficial learning (e.g. Balch 1964, Kress 1985, Messick 1987, Struyven, Dochy & Janssens 2005:337). Paxton (2000) for example, claims that MCQ assessment does not encourage interpretation and thinking, and that '…in fact the emphasis on fast response in limited time tests militates against reflection' (2000:111).

While there is much truth in these criticisms, these problems can be mitigated to some extent by ensuring that MCQs form only a part of the overall assessment strategy for a course, and by creativity and due diligence on the part of test-setters. Fellenz (2004:705) believes that many of the disadvantages associated with MCQs can be overcome 'through conscientious use of the format and by employing it alongside other examination methods as part of a carefully designed assessment protocol'. 'Conscientious use of the format' would presumably involve scrupulous pre-test and post-test checking, avoiding overuse of MCQs in inappropriate situations, and fully exploiting all the levels of knowledge, skills and abilities that can be tested using this format. Martinez (1999:208) acknowledges that while MCQs are often simply recognition items, they can also be written so as to 'elicit complex cognitions, including understanding, prediction, evaluation and problem solving'. Well-informed and conscientious assessors are therefore crucial to ensuring fairness throughout the process of designing multiple-choice assessments, compiling questions, marking tests and analysing and acting on test results.

1.2.2.1 Setting MCQs

As far as the formats of MCQs are concerned, the most commonly encountered multiple-choice format is an initial question or statement (the stem) followed by a set of possible options, one of which is the correct answer (the key) and the rest of which are incorrect (the distractors). Students are required to choose the best answer choice from the range of options provided. An

example from the 2007 Linguistics examination in Appendix B is given below, with an asterisk to identify the key:

1.	The unconscious, informal process of 'picking up' a language in the pre-adolescent years is known as	STEM
	[1] developmental psycholinguistics	DISTRACTOR
	*[2] language acquisition	KEY
	[3] language learning	DISTRACTOR
	[4] the critical period	DISTRACTOR
	[5] additive bilingualism.	DISTRACTOR

Other possible common multiple-choice formats include 2-option alternatives (True/False, Yes/No), matching terms from two columns, multiple-response items where more than one right answer is allowed, and complex 'Type K' items where a preliminary set of data is followed by various combinations of options (Brown, Bull & Pendlebury 1997, Haladyna 2004:41-96, Stiggins 2005). An example of a Type K multiple-choice item is provided here:

Which of the following are fruits?

A Tomatoes B Avocados C Potatoes

[1] A and B
[2] A and C
[3] B and C.

An increasingly popular format is the context-dependent item set, which consists of a scenario or text followed by a number of MCQs relating specifically to the text or more generally to the topic addressed in the text. This format, well-known to most people from reading comprehension tests, allows authentic texts (newspaper cuttings, case studies, research data, extracts from academic articles, etc.) to be included in the assessment so that students can be tested on whether they can apply their knowledge and skills to new contexts. An example of an item set from the 2007 Linguistics MCQ examination in Appendix B is provided below:

Read the following case study and then answer Questions 6 to 8.

Mr Dlamini is a businessman who speaks Zulu as L1 and English as L2. He attended a school in Soweto where Zulu was the medium of instruction. Now he is learning German through Unisa in order to conduct international business transactions.

6. Mr Dlamini is learning German primarily for

[1] instrumental reasons
[2] integrative reasons
[3] interference reasons
[4] internal reasons
[5] idiosyncratic reasons.

7. In Mr Dlamini's case, learning German involves

[1] first language acquisition
[2] second language learning
[3] spontaneous language learning
[4] foreign language learning
[5] third language acquisition.

8. Mr Dlamini's schooling can be described as

[1] mother-tongue education
[2] an immersion programme
[3] a submersion programme
[4] a transitional programme.

One issue that has been researched in detail (more than 27 studies since 1925) is the issue of how many options an MCQ should have (cf overview in Rodriguez 2005). While most tests seem to use four or five options as the norm, Haladyna, Downing and Rodriguez (2002:312) suggest that test developers should write as many effective options as they can. Stiggins (2005:101) suggests that the number of options can vary in the same test. However, research shows that it is very unlikely that item writers can write more than three plausible functioning distractors, where a 'functioning distractor' is defined as one that is selected by at least 5% of students (Haladyna, Downing & Rodriguez 2002). This point is illustrated by a look at the frequency of responses to the various answer choices in Question 8 above:

8. Mr Dlamini's schooling can be described as

65,9%	*[1]	mother-tongue education
14,6%	[2]	an immersion programme
4,3%	[3]	a submersion programme
12,9%	[4]	a transitional programme.

Option [1] is the key and is selected by 65,9% of students. The functioning distractors here include only answer choices [2] and [4]. Rodriguez (2005:10-11) concludes that three-option MCQs are optimal in most settings because they are easier for item writers to set, take less time to administer, allow for more items in a set time period and therefore result in better content coverage and test score reliability.

There is a wealth of literature advising educators on how to write good multiple-choice items that can improve accessibility and limit ambiguity and avoidable misinterpretation (e.g. Haladyna 1994, 1997, 2004, Brown, Bull & Pendlebury 1997, Osterlind 1998, Haladyna, Downing & Rodriguez 2002, McCoubrie 2004, Stiggins 2005, Fairbairn & Fox 2009). Most of this advice is based on accumulated experience, common sense and on general principles of good writing.

Thomas Haladyna (e.g. 1994, 1997, 2004) is the acknowledged expert in multiple-choice design and research. He and his colleagues have proposed, collated and researched a wide range of item-setting guidelines (e.g. Haladyna & Downing 1989, Crehan & Haladyna 1991, Crehan, Haladyna & Brewer 1993, Haladyna 1994:61-86, Haladyna, Downing & Rodriguez 2002:312). These guidelines relate either to the format of items (e.g. 'Avoid complex type-K items'), the content of items (e.g. 'Focus on a single problem') or to the language of MCQs. In this study the focus is primarily on the language of MCQs, and content and format will be discussed only in so far as they relate to language. The following eight MCQ guidelines from Haladyna (2004) relating specifically to language will be investigated in the present study with respect to their impact on question difficulty for L1 and L2 students:

(a) Avoid negative words such as *not* or *except* in the stem
(b) State the stem in a question format instead of an incomplete sentence format

(c) Simplify vocabulary and aim for maximum readability
(d) Keep items as brief as possible
(e) Avoid very similar answer choices
(f) Try and keep items grammatically parallel
(g) Avoid All of the above (AOTA)
(h) Keep None of the above (NOTA) to a minimum.

The question of what makes an MCQ difficult is addressed in 1.2.2.2 below, followed by an overview of multiple-choice assessment as it affects L2 students in section 1.2.2.3.

1.2.2.2 What makes an MCQ difficult?

Investigating the difficulty of an MCQ is itself a complex enterprise – it covers everything from how well students did on the question (what percentage answered correctly) to more subtle issues such as the cognitive level at which the question is pitched, the degree to which the language of the question is comprehensible and whether or not the lecturer's intention is clear. As Steiner (1978) points out,

> even a cursory reflection suggests that when we say 'this text is difficult' we mean, or intend to mean, a number of very different things. The rubric 'difficulty' covers a considerable diversity of material and methods.
>
> (Steiner 1978:19)

In an educational context, a 'difficult' question may have a positive connotation as a 'higher-order' question that challenges and stretches the more able students, requiring analysis or application to new contexts. These are a necessary part of a well-designed test. A wealth of recent literature (see review in section 2.6) attests to the fact that MCQs can be written and designed in an academically responsible way that tests both lower-order knowledge, content and attitudinal outcomes as well as more complex cognitions such as the ability to classify, compare, draw conclusions, identify main ideas and use existing knowledge to answer new items (cf Haladyna 1997, Case & Swanson 2001, Fellenz 2004, McCoubrie 2004, Stiggins 2005, Williams 2006).

Whatever their level of cognitive difficulty, MCQs are often perceived as ‘difficult’, ‘tricky’ or ‘unfair’ (Roberts 1993, Paxton 2000). A ‘difficult’ question in this sense has a negative connotation, implying that students come up against ‘impenetrability and undecidabilities of sense’ (Steiner 1978:18) or even of deliberate attempts to trick them into a wrong answer (cf Roberts 1993). (Trick questions are discussed in more detail in section 2.4.2 in the next chapter).

Part of the problem is that MCQs require both a great deal of reading (Kilfoil 2008:129) and quite sophisticated reading comprehension in addition to knowledge of the content subject they are aiming to test. According to Haladyna (2004:241), ‘testing policies seldom recognize that reading comprehension introduces bias in test scores and leads to faulty interpretations of student knowledge or ability’. As regards reading comprehension, the difficulty of an MCQ can lie at several levels, including word choice, syntax and the effort required to process the sentence. Fellbaum’s (1987) research on multiple-choice cloze questions in the U.S. Scholastic Aptitude Tests (SATs) showed that students are faced with the following tasks when completing a multiple-choice cloze item:

(a) comprehension of the meaning of all the words in the item as well as in the options
(b) comprehension of the syntactic relations among the sentence constituents and the syntactic relations across clauses, in particular as expressed by conjunctions such as *even, though*
(c) recognition and retention during the sentence processing of lexically or morphologically expressed antonymic relations.

(Fellbaum 1987:206)

In order for these processes to take place optimally, it is important that MCQs are fair in the sense of being well-structured, free of errors and written in such a way that as many students as possible will understand what is being asked. Carstens (2000:147) suggests that the linguistic complexity of examination papers should be minimised by following Plain Language guidelines (see e.g. Flesch 1962, PlainTrain Plain Language Online Training Program). Reader-orientated plain language guidelines such as the following can help to make ideas as clear as possible:

(a) Use short, familiar words and expressions where possible, e.g. *do* instead of *accomplish*, *although* rather than *notwithstanding the fact that.*
(b) Don't change verbs like *require* or *produce* into more complex nouns like *requirement* or *production.*
(c) Don't use strings of nouns like *world food production.*
(d) Keep sentences short, simple, active and unambiguous.
(e) Use linking words such as *although*, and *on the other hand* to make the relationships between sentences clear.
(f) Avoid double negatives.
(g) Organise information logically, e.g. step by step or from general to specific.

(extracts summarised from PlainTrain Plain Language Online Training Program)

In a comparative study, Brown (1999) reported that for students who understood the content of a science test, a plain language version was better able to accurately assess their knowledge than a linguistically more complex original version. (For students who did not understand the content, the version made little if any difference in performance.) Linguistic simplification of tests is explored in more detail in section 2.4.4.

According to Thompson, Johnstone and Thurlow (2002), aiming for maximum comprehensibility does not imply a relaxing of standards or a change in what should be measured (Thompson, Johnstone & Thurlow 2002:8). Graff (2003:129) agrees, arguing that although there are those who continue to insist that 'writing more accessibly and reaching a wider audience means dumbing yourself down and compromising your intellectual standards […], general accessibility is fully compatible with intellectual integrity'. In content-subject tests, where adequate comprehension of questions by testees is essential to the test's ability to do what it sets out to do, I would argue that general accessibility is in fact a prerequisite of intellectual integrity.

1.2.2.3 Multiple-choice assessment and L2 students

My own research explores the extent to which two MCQ Linguistics examinations at the University of South Africa are in fact 'generally accessible' to L2 students. If even mother-

tongue English speakers find MCQs 'tricky', how valid are MCQs as a way of testing L2 speakers? Paxton (2000) was among the first to raise the question of what L2 students in South Africa find difficult about the language of MCQs and to provide some initial answers. She calls for more systematic research because 'we know very little about the ways in which assessment instruments impact on the development of literacy skills or about how and why particular forms of assessment advantage some and disadvantage others' (Paxton 2000:111). She concludes that

> The issue of language medium for assessment is an area that is relatively unexplored in South Africa and it seems crucial that we know more about the ways in which particular assessment procedures impact on second language speakers of English […]
> (Paxton 2000:114).

There is ample evidence to suggest that students tend to do better in MCQs than in other kinds of tests (e.g. Kniveton 1996, Struyven, Dochy & Janssens 2005), and two South African studies, Curtis, de Villiers and Polonsky (1989) and Miller, Bradbury and Wessels (1997:78) indicate that both L1 and L2 students typically achieve higher scores for MCQ examinations than for essay examinations. Kilfoil (2008:129) suggests that this could be because L2 students' reading skills are generally better developed than their writing skills. In recent years, however, many educational researchers have begun raising concerns about language as a source of bias that can compromise the test performance of non-native speakers of English when they write MCQ examinations (e.g. Ghorpade & Lackritz 1998, Paxton 2000, Abedi et al. 2000, Goh 2004, McCoubrie 2004, Dempster & Reddy 2007, Fairbairn & Fox 2009). Bosher and Bowles (2008) believe that for non-native English speakers taking exams in content subjects, 'every test becomes a test of language proficiency' (Bosher & Bowles 2008:166). Lampe and Tsaouse (2010) cite Bosher's (2009:263) sobering claim that 'Both [U.S. nursing] students and faculty rated taking multiple-choice tests as the most difficult task out of 74 language and culture-related skills and tasks that students must be able to perform successfully in the nursing program.'

In an attempt to address this problem, Fairbairn and Fox's (2009:14) guidelines for test developers listed below are intended to help make tests comprehensible for students who are not mother-tongue English speakers:

(a) Use short, clear sentences or stems
(b) Use consistent paragraph structure
(c) Use present tense and active voice where possible
(d) Minimise rephrasing or rewording ideas
(e) Use pronouns carefully
(f) Use high-frequency words
(g) Avoid or explain colloquialisms or words with more than one meaning
(h) Write items with a reading level below grade level
(i) Identify non-content vocabulary that may be difficult as well as cognates and false cognates (similar-looking words with different meanings in the two languages) through empirical work with second-language test takers.

Several of these overlap with Haladyna's MCQ guidelines outlined earlier, for example (a) mirrors Haladyna's guideline to keep items as brief as possible, while (f) – (i) mirror Haladyna's advice to simplify vocabulary and aim for maximum readability. According to Haladyna, Downing and Rodriguez (2002:315) and Stiggins (2005:101), test setters should aim to make their tests readable by the weakest readers in the group to avoid putting these students at risk of failure. This implies that L2 students' reading levels should be taken into consideration in the design of questions. Kilfoil (2008:130) agrees that the language of MCQs 'should receive attention so that it does not become an unnecessary barrier to some students and an advantage to others.' Carstens (2000:146) argues that while subject-field terminology is essential to academic precision, in an examination situation, difficult general academic words could and in fact should be substituted by more familiar synonyms or paraphrases for the sake of L2 students:

> The comprehension speed and ease of a non-mother-tongue speaker will certainly be enhanced by using words such as *use* instead of *utilize*, *find out* instead of *ascertain*, *speed up* instead of *expedite*, etc.
>
> (Carstens 2000:146).

Bosher and Bowles (2008:166) and Lampe and Tsaouse (2010) also note that one source of difficulty for L2 candidates writing MCQ items in undergraduate nursing programmes was the greater processing time needed to complete the test. Haladyna (1994:27) and Kniveton (1996)

suggest, in an English L1 context, that generally one can administer about one MCQ item per minute. Additional time needs to be built in for reading scenarios or case studies according to Haladyna, Downing and Rodriguez (2002:324) and also if many of the students have English as a second language (Prins & Ulijn 1998:149).

1.3 Aims of the study

Against the background sketched in section 1.2 above, the aim of my research is to answer the following questions:

(a) Which kinds of multiple-choice questions are 'difficult'?
(b) What kinds of multiple-choice items present particular problems for second-language speakers of English?
(c) What contribution do linguistic factors make to these difficulties?

The aims of the study are primarily descriptive-analytical, but also theoretical-methodological and applied. These are explained in more detail below:

At a **theoretical-methodological** level the aim of the study is to contribute to the debate on MCQ validity and fairness, drawing together methodology from classical test theory, readability research and qualitative test-taker feedback to investigate the notion of MCQ difficulty in a triangulated way. By identifying difficult questions statistically, analysing them linguistically and hearing what the students themselves say about why particular questions are difficult, the study aims to make a contribution to the debate around the effect of readability on MCQ difficulty and to the empirical investigation of item-writing guidelines relating to the language of MCQs. The study also attempts to shed some light on the interaction between the cognitive difficulty and linguistic difficulty of questions. Although it is not one of the primary aims of the study, some contribution will also be made to the empirical verification of Bloom's cognitive levels.

At a **descriptive-analytical** level, the study aims to explore linguistic factors that contribute to MCQ item difficulty and that increase the gap between L1 and L2 performance in MCQs in a

first-year content subject at the University of South Africa. In particular, the study investigates to what extent violations of eight item-writing guidelines relating to the language of the question (see 1.2.2.1 above) result in more difficult questions or a negative impact on L2 students. The item-writing guidelines in question deal with the length of items, incomplete sentence stems, negative words such as *not* and *false*, readability and vocabulary density, similar answer choices, parallel grammatical structure of answer choices, AOTA items and NOTA items. The study thus attempts to identify particular linguistic aspects of MCQs that make items more difficult to comprehend and answer.

At an **applied** level it is hoped that the findings of this study will inform guidelines for item design and result in tests that are fairer and better able to measure understanding of the discipline concerned rather than of English language. This in turn will assist L2 students and improve the validity of MCQ assessment.

Section 1.4 below offers an overview of the larger methodological framework into which the research fits.

1.4 Overview of methodological framework

Prompted by concern over how the language of MCQ tests might be negatively influencing the L2 students in my own first-year Linguistics course, my research falls within the growing area of study known as educational linguistics. Educational linguistics, according to Hult (2008:10), ‘integrates the research tools of linguistics and other related disciplines of the social sciences in order to investigate holistically the broad range of issues related to language and education’. The term ‘educational linguistics’ was first used in the 1970s by Bernard Spolsky (Spolsky 1978), who saw it as a special focus area within applied linguistics, concentrating on ‘those parts of linguistics directly relevant to educational matters as well as those parts of education concerned with language’ (Spolsky 2008:2). Covering a range of themes from language of instruction to literacy to assessment, the purpose of educational linguistics is always ‘to inform or to be informed by educational practice, either directly or indirectly’ (Hult 2008:20). The nature of the inquiry in educational linguistics is very often ‘transdisciplinary’, in that the researcher begins with an issue, concern, problem, etc. and then draws on relevant methodological and analytical

resources to investigate it in innovative ways (Hult 2008:13, see also Halliday 1990, Hornberger 2001).

One of the ways in which the problem of language as a source of 'unfair' performance variance for L2 speakers of English when they take MCQ examinations has been researched is to statistically contrast the responses of minority language students with those of academically comparable English L1 students (e.g. Angoff 1993, Dorans & Holland 1993, O' Neill & McPeek 1993, Gierl 2005 (see section 2.5.3)). A second, more linguistically-orientated methodology is to attempt to relate item difficulty to item readability (e.g. Prins & Ulijn 1998, Homan, Hewitt & Linder 1994, Hewitt & Homan 2004, Dempster & Reddy 2007 (see section 2.3.3)) or alternatively to relate item difficulty to the occurrence of various linguistic features such as the length of the stem, syntactic difficulty of the stem, negation in the stem or the degree of similarity within answer choice sets (see section 2.4.1). Researchers in the latter category include Bejar, Stabler and Camp (1987), Sireci, Wiley and Keller (1998), Rupp, Garcia and Jamieson (2001), Dempster and Reddy (2007) and Bosher and Bowles (2008).

We have seen that there are increasing calls for MCQs to be piloted on and scrutinised by a sample of students similar to those on which the MCQs are to be used (see for example Haladyna 2004, Fairbairn & Fox 2009). Methodologies such as focus groups and questionnaires that allow students a voice in the evaluation of MCQs are becoming more prominent, both in empirical research on MCQs and in test validation. For example, Haladyna (1994:136) recommends a test-taker review as a way of detecting ambiguous or misleading MCQ items. This typically takes the form of a think-aloud protocol – a technique where participants are asked to vocalise their thoughts, feelings and opinions as they perform a task such as reading, writing, translating, using a product or, as in this case, writing a test. The resulting verbalisations can include paraphrases, elaborations, explanations, inferences or misrepresentations of the text itself (Taylor & Taylor 1990:77-78) and are useful in understanding students' answering strategies, mistakes that are made and how the test questions could be improved to avoid these problems (cf. Connolly & Wantman 1964, Haney & Scott 1987, Norris 1990, Farr, Pritchard & Smitten 1990, Strauss 1995, Prins & Ulijn 1998, Paxton 2000, Haladyna, Downing & Rodriguez 2002). Paxton (2000:118) suggests that '[t]hink aloud protocols on multiple-choice questions might give us a sense of

where the stumbling blocks come and whether language medium does in fact put some students at a disadvantage.' Fairbairn and Fox (2009) are unequivocal in their criticism of tests that fail to do this:

> In spite of agreement in principle within the broader testing community that the active elicitation and use of test taker feedback and response is central to ethical testing practice […], there is little more than lip-service paid at present to the systematic incorporation of 'bottom-up' test taker feedback in testing processes and a continued reliance on 'top-down' psychometric approaches in arguments for validity, reliability, and fairness.
>
> (Fairbairn & Fox 2009:16)

The current study attempts to address Fairbairn and Fox's concern and take a multipronged view of MCQ 'difficulty', using all three of the methods described above (statistical analysis of student performance on each item, linguistic analysis of the test questions and test-taker feedback) on the same set of test items. By investigating MCQ texts and students' MCQ performance from several angles, my intention was to obtain rich data on the reasons why particular items are difficult to understand and to answer, for both L1 and L2 speakers of English. My hope was that this would enable me to identify trends and make recommendations that could improve both the comprehensibility and the validity of multiple-choice test items.

1.4.1 Quantitative aspects of the research design

The quantitative research will include statistical analysis of the examination results of two classes of students writing MCQ examinations in first-year Linguistics at Unisa in the normal course of their studies. The measures that will be used (see section 2.5) are the difficulty or p-value of each item (the percentage of students who answer correctly), the discrimination of each test item (the extent to which each item distinguishes between stronger and weaker candidates) and a measure that compares average item difficulty for L1 and L2 students. This will enable me to identify MCQs that are difficult in that they are answered incorrectly by many students or cause particular problems for L2 students.

The statistical data will be coupled with a linguistic analysis of the texts of the same two MCQ examinations in first-year Linguistics to quantify the readability level of each MCQ and its density of academic words. Correlations will be sought between difficulty and various indicators

of readability. The research will quantify whether and in what ways the MCQs adhere to the eight item-writing guidelines relating to readability, negatives, long items, incomplete sentence stems, similar answer choices, grammatically non-parallel answer choices and AOTA and NOTA items and investigate how these impact on question difficulty, discrimination and fairness as measured by the difference in average item difficulty for L1 and L2 students. These findings will shed light on whether these eight guidelines are empirically supported or not. Finally, by categorising each item according to the Bloom level at which the question is posed, an attempt will also be made to tease apart the respective contributions of cognitive difficulty and linguistic difficulty on student MCQ performance.

1.4.2 Qualitative aspects of the research design

To complement the quantitative portion of the study, semi-structured interviews will probe L1 and L2 students' opinions of multiple-choice assessment, followed by a think-aloud protocol in which students talk me through an 80-item Linguistics MCQ examination. The purpose of the qualitative investigation in this case is both descriptive, attempting to provide insights into the problems experienced by individual students in reading, understanding and answering MCQs, and interpretive, attempting to relate these difficulties to aspects of the language of the questions. No prior hypothesis will be tested, but there is a general expectation that difficult questions (those that students leave out, get wrong or have difficulty answering) might be associated with linguistic features such as negation, ambiguity, academic words and long or complex sentences.

1.4.3 Participants

The participants in the quantitative portion of the study included the entire student body of a first-year course in Linguistics at Unisa. The course was entitled LIN103Y (Multilingualism: the role of language in the South African context) and the results that were analysed came from the second semester of 2006 (136 students) and the second semester of 2007 (117 students). Unisa students are adults doing either part-time or full-time open distance education studies and coming from a wide spectrum of educational backgrounds. Their ages range from 18 to post-retirement age and they live mostly in Southern Africa but can be based anywhere in the world.

Soon after the final examination, 13 of these students from the Johannesburg and Pretoria area accepted my request to participate in a think-aloud protocol of the MCQ examination they had just written. Three of these were English mother-tongue students and ten spoke English as a second language, percentages which were approximately in proportion to the L1/L2 language profile of the entire class. Of the ten L2 students selected, eight different mother tongues were represented (see section 3.5.3).

1.4.4 Ethical considerations

Consideration was given to ensuring the rights of participants to dignity, privacy, non-participation, anonymity and confidentiality (Seliger & Shohamy 1989:195, Tuckman 1999). In the quantitative portion of the study, average MCQs results were calculated for L1 and L2 students and anonymity and confidentiality of individual student results was therefore not an issue. In the qualitative portion of the study, students participated on a voluntary basis and real names were not used in the study. Individual students gave their written consent and were informed orally by the researcher on the nature of the study (see Appendix C).

Although the interviews with the 13 students were time-consuming and often tiring for students, lasting approximately two hours each, they took place in the holiday period after the semester had finished and before the start of the following semester and therefore didn't take students away from their studies or affect their results in any way. Students were offered opportunities to cut the interviews short if there were signs of tiredness, but none of the students took me up on this offer. Students made no objections to the nature of the questioning and the interviews were conducted in good spirit, allowing plenty of time for discussion of student-initiated topics. Unisa students seldom interact face-to-face with their lecturer and those interviewed took the opportunity to ask me content-related questions, discuss their own linguistic backgrounds and experiences and make their voices heard as regards the academic and other challenges they face as distance education students.

From the point of view of research ethics, it is also important that feedback regarding the research findings is given in an attempt to improve future teaching and assessment. Interim findings and results were therefore presented at a seminar on multiple-choice assessment hosted

by Unisa's Directorate of Curriculum and Learning Development in 2007, at the Association for the Study of Evaluation and Assessment in Southern Africa (ASEASA) conference in 2008, and in a presentation to participants in Unisa's 'young academic' staff development programme in 2009.

1.5 Structure of the thesis

Chapter 1 introduced the research questions, context and methodology of the study, describing MCQs and their use in higher education in South Africa with a focus on ensuring fairness and validity for both L1 and L2 students.

Chapter 2 is a literature review that provides a deeper understanding of the research context outlined in chapter 1. It begins by exploring linguistic aspects of academic writing and its comprehension by L2 students. The focus then moves more specifically to the language of multiple-choice assessment and to ways of ensuring test fairness and comprehensibility by following accepted and empirically researched item-writing guidelines. Readability research and readability formulas are discussed in regard to how aspects of text difficulty can be quantified, followed by an overview of how statistical item analysis can contribute to MCQ fairness and validity. The chapter concludes with an overview of research on the cognitive levels of difficulty that MCQs can probe.

Chapter 3 describes the research design, data, methods and procedures selected for the study. While the various quantitative and qualitative methodologies for understanding MCQ difficulty are introduced and discussed in Chapters 1 and 2, they are described and justified here in greater detail in relation to the current study. These include statistical item analysis of two 80-item MCQ examination papers, readability analysis and classification of the linguistic characteristics of the two papers, semi-structured interviews and a think-aloud protocol.

Chapter 4 presents the findings of the statistical analysis of student results on individual MCQs. Difficulty, discrimination and differences between average L1 and L2 performance are calculated for incomplete statement stems, negative questions, long questions, similar answer choices, grammatically non-parallel answer choices, AOTA and NOTA items. The effect of item

readability, density of academic words and cognitive level of the question on the performance of L1 and L2 students is also described and an attempt is made to identify the linguistic characteristics of the most difficult questions and those which result in the widest gap between L1 and L2 performance.

Chapter 5 describes the think-aloud interviews and reports on the findings of the qualitative portion of the research. Profiles are provided of the 13 participants, together with a comparison of the scores they achieved for the actual MCQ examination and the think-aloud protocol covering the same test material. Students' opinions of multiple-choice assessment are discussed as are the MCQ-answering strategies they employed. The difficulties experienced by both L1 and L2 students in comprehending and answering the MCQs in the think-aloud protocol are classified into difficulties relating to readability, difficulties relating to violations of the other item-writing guidelines and other (e.g. content and layout-related) difficulties.

Chapter 6 offers conclusions and recommendations for users of MCQs based on a synthesis of the qualitative and quantitative findings. It revisits the research questions, highlights the contribution and limitations of the study and offers suggestions for future research.

Chapter 2
Literature review

2.1 Introduction

This study attempts to find out what kinds of MCQs are difficult for Linguistics students at the University of South Africa (Unisa), focusing on the issue of how the language of test questions impacts on readability, difficulty and student performance for L2 English students in comparison to L1 students. Much if not all the literature reviewed in this chapter is underpinned by an interest in the difficulty of test questions, particularly in cases where students are linguistically diverse.

The literature that relates to the notion of MCQ difficulty is interdisciplinary – coming amongst others from education and specific disciplines within education, from applied linguistics, document design, psychology and psychometrics. By way of contextualisation, section 2.2 offers a discussion of academic discourse and reading in a second language, with a focus on identified lexical and grammatical features that contribute to text difficulty. The issue of whether lexical and grammatical aspects of text can be used to predict text difficulty informs the literature reviewed in section 2.3 on readability formulas. Section 2.4 focuses more narrowly on the language of multiple-choice questions. This section includes a discussion of empirical research on item-writing guidelines, trick questions and Grice's cooperative principle in relation to MCQ item-writing, as all of these offer principles that can be followed to help minimise the linguistic difficulty of MCQs. Section 2.5 provides a discussion of relevant literature on the statistical analysis of MCQ results, including the difficulty of items for students and subgroups of students. Section 2.6 addresses the issue of MCQ difficulty from the point of view of research on the cognitive levels that can be tested using MCQs.

2.2 Academic texts and L2 readers

In order to provide some background context to the linguistic characteristics of MCQs, a brief description is given here of some research findings relating to the language of academic

discourse and of how L2 students cope with reading and making sense of MCQs. Particular attention will be paid to vocabulary and syntactic constructions that have been shown by other research to be 'difficult' and which may therefore contribute to the difficulty of questions in my own study.

2.2.1 Linguistic characteristics of academic text

Analysis of the linguistic characteristics of academic text shows that there are particular constellations of lexical and grammatical features that typify academic language, whether it takes the form of a research article, a student essay or the earliest preschool efforts to 'show-and-tell' (Schleppegrell 2001:432). The prevalence of these linguistic features, Schleppegrell argues, is due to the overarching purpose of academic discourse, which is to define, describe, explain and argue in an explicit and efficient way, presenting oneself as a knowledgeable expert providing objective information, justification, concrete evidence and examples (Schleppegrell 2001:441). Biber (1992) agrees that groups of features that co-occur frequently in texts of a certain type can be assumed to reflect a shared discourse function.

At the lexical level, this discourse purpose of academic writing is achieved by using a diverse, precise and sophisticated vocabulary. This includes the subject-specific terminology that characterises different disciplines as well as general academic terms like *analyse, approach, contrast, method* and *phenomenon*, which may not be particularly common in everyday language (Xue & Nation 1984). The diversity and precision of academic vocabulary is reflected in relatively longer words and a higher type-token ratio than for other text types (Biber 1992). (Type-token ratio reflects the number of different words in a text, so a 100-word text has 100 tokens, but a lot of these words will be repeated, and there may be only say 40 different words ('types') in the text. The ratio between types and tokens in this example would then be 40%.) Academic text is also often lexically dense compared to spoken language, with a high ratio of lexical items to grammatical items (Halliday 1989, Harrison & Bakker 1998). This implies that it has a high level of information content for a given number of words (Biber 1988, Martin 1989, Swales 1990, Halliday 1994). Academic language also tends to use nouns, compound nouns and long subject phrases (like *analysis of the linguistic characteristics of academic text*) rather than subject pronouns (Schleppegrell 2001:441). Nominalisations (nouns derived from verbs and other parts of speech) are common in academic discourse, e.g. *interpretation* from *interpret* or

terminology from *term*. These allow writers to encapsulate already presented information in a compact form (Martin 1991, Terblanche 2009). For example, the phrase *This interpretation* can be used to refer to an entire paragraph of preceding explanation.

Coxhead (2000) analysed a 3,5 million-word corpus of written academic texts in arts, commerce, law and science in an attempt to identify the core academic vocabulary that appeared in all four of these disciplines. After excluding the 2000 most common word families on the General Service List (West 1953) which made up about three-quarters of the academic corpus, Coxhead (2000) identified terms that occurred at least 100 times in the corpus as a whole. This enabled her to compile a 570-item Academic Word List (AWL) listing recurring academic lexemes ('word families') that are essential to comprehension at university level (Xue & Nation 1984, Cooper 1995, Hubbard 1996). For example, the AWL includes the word *legislation* and the rest of its family – *legislated, legislates, legislating, legislative, legislator, legislators* and *legislature* (Coxhead 2000:218). This division of the list into word families is supported by evidence suggesting that word families are an important unit in the mental lexicon (Nagy et al. 1989) and that comprehending regularly inflected or derived members of a word family does not require much more effort by learners if they know the base word and if they have control of basic word-building processes (Bauer & Nation 1993:253). The words in the AWL are divided into ten sublists according to frequency, ranging from the most frequent academic words (e.g. *area, environment, research* and *vary*) in Sublist 1, to less frequent words (e.g. *adjacent, notwithstanding, forthcoming* and *integrity*) in Sublist 10. Each level includes all the previous levels.

The AWL has been used primarily as a vocabulary teaching tool, with students (particularly students learning English as a foreign language) being encouraged to consciously learn these terms to improve their understanding and use of essential academic vocabulary (e.g. Schmitt & Schmitt 2005, Wells 2007). Researchers have also used the AWL to count the percentage of AWL words in various discipline-specific corpora, for example in corpora of medical (Chen & Ge 2007), engineering (Mudraya 2006) or applied linguistics texts (Vongpumivitch, Huang & Chang 2009). These studies show that AWL words tend to make up approximately 10% of running words in academic text regardless of the discipline. Other high frequency non-everyday

words in discipline-specific corpora can then be identified as being specific to the discipline, rather than as general academic words. However, there is some debate (see e.g. Hyland & Tse 2007) as to whether it is useful for L2 students to spend time familiarising themselves with the entire AWL given that some of these words are restricted to particular disciplines and unlikely to be encountered in others, and may be used in completely different ways in different disciplines (Hyland & Tse 2007:236). For example, the word *analysis* is used in very discipline-specific ways in chemistry and English literature (Hyland & Tse 2007).

Notwithstanding this debate, it is incontrovertible that the AWL words occur more frequently in academic discourse than in other domains such as fiction (Hyland & Tse 2007:236). While AWL words make up approximately 10% of running words in academic text, they made up only 1,4% of running words in Coxhead's (2000) comparative fiction corpus. They are therefore infrequent in non-academic texts and unlikely to be well understood by first-year students. The AWL is important for my own study in that the density (percentage) of AWL words in university-level MCQs might be a factor influencing question difficulty and will therefore be measured in order to compare the lexical density of MCQs.

At the grammatical level, academic text is known to be complex in its use of embedded clauses and of a varied set of conjunctions that mark logical connections between parts of the text (Halliday & Hasan 1976, Schleppegrell 2001). Evidence of this is provided by Biber's (1988, 1992) corpus-based approach to measuring text complexity. Biber (1992) analysed a million-word oral and written corpus of British English to establish patterns of co-occurrence of 33 linguistic features known to be associated with discourse complexity (for example long words, passives and nominalisations). He then used statistical methods to identify the occurrence of these linguistic features over a range of registers, including academic text. While all the features are surface structure (directly observable) markers of complexity, Biber (1992:141) acknowledges that they are 'not fully representative of the linguistic correlates of discourse complexity'. This implies that features representing cohesion, informational packaging and rhetorical organisation, which cannot be easily identified or counted by a computer, also contribute to text complexity (see also section 2.3.2 below.)

As regards surface linguistic features, Biber (1992:181) argues that different texts and different text types are not simply more or less complex than one another, but more or less complex in terms of five separate dimensions, the first of which reduces complexity (see (a) below) and the other four of which increase the complexity of a text (see (b) below):

(a) Reduced discourse complexity, according to Biber (1992), is achieved by structural reduction (signalled by contractions and *that*-deletions after verbs like *believe*), by less specified reference (signalled, for example, by the pronoun *it*, the pro-verb *do* and demonstrative pronouns) and by coordinating rather than embedding clauses to create a more fragmented and less dense structure. This reduced structure dimension is strongly associated with the difference between speech and writing. Spoken language tends to have more structural reduction, while carefully planned, written registers have a below-average frequency of reduced forms.

(b) Greater discourse complexity, according to Biber (1992), results firstly from the passive constructions common in expository text, and secondly from the elaboration of reference signalled by *wh*-relatives, *-ing* relatives and *that*-relatives. These frequent relative clauses are typical of informational genres. The third dimension, integrated structure, is signalled by longer words, a high type-token ratio and a high frequency of nouns and noun-noun compounds, prepositions, attributive adjectives, nominalisations and phrasal coordination. A highly integrated structure makes a text denser, more precise, and reflects a high informational focus. Finally, *wh*-clauses and conditional clauses, adverbial subordination, *that* complement clauses and infinitives are frequent in registers that express and justify personal attitudes and feelings rather than presenting factual information. These ‘frame’ the text by explaining and justifying the writer’s stance and are referred to by Biber as ‘framing elaboration’.

Academic writing, like all forms of expository text, tends to use many passives (Biber 1992:151). Biber (1992) argues that academic writing tends not to display reduced structure

because it is written rather than spoken, and carefully produced and edited rather than produced without prior planning. It would typically have high referential elaboration (many relative clauses) as it is informational rather than non-informational, and would have high integrated structure as it is an informational written genre (all other genres score low). Academic writing would typically score low on framing elaboration as it is impersonal and factual rather than expressing personal attitudes and feelings.

Academic writing, including multiple-choice questions, is therefore likely to be located towards the 'difficult' end of the spectrum on all five dimensions except framing elaboration, and would include many of the grammatical constructions that Biber (1988) describes as complex, including nominalisations, passives and subordinate clauses of various kinds as well as a high type-token ratio and longer-than-average words. The example Biber gives of an official document shares the features of academic text in the prevalence of passives (underlined), referential elaboration features (in bold) and integrated structure features (in italics) and in the absence of framing elaboration and reduced structure features:

> *Questions about marriage and children* were again included, as they had been *at the 1911 and 1951 censuses*. The *1961 census questions* related *to* all *women* **who** were or had been married, and so repeated the *enquiry* made *fifteen years* earlier *by* the *1946 family census* conducted *on behalf of* the *royal commission on population*. The *questions about children* … extended *to* all *women* **who** were or had been married.
>
> (Biber 1992:153)

The following example of an MCQ from the examination in Appendix A illustrates a similar prevalence of passives (underlined), referential elaboration features (in bold) and integrated structure features (in italics):

47. Which *of* the *following* offers the *best definition of* the *sociolinguistic term dialect*?
    - [1] *Dialects* are *mutually intelligible forms of different languages*.
    - [2] A *dialect* is a *substandard, low status, often rustic form of language*.
    - [3] *Dialects* are *language varieties* associated *with particular geographical areas*.
    - [4] *Dialects* are *language varieties* associated *with particular social classes.*
    - [5] The *term* '*dialect'* refers *to languages* **that** have no *written form.*

Biber's findings are relevant for my own research in that the surface-structure syntactic features he associates with discourse complexity, particularly the ones that typify academic writing, may be expected to function as potential sources of difficulty for students. These linguistic features may result in difficult MCQs with low p-values and high differential between L1 and L2 scores, and may also be self-identified by students as 'difficult' in the think-aloud protocol.

2.2.2 L2 students' comprehension of academic text

Given its specialised vocabulary and complex syntax described above, it is not surprising that academic writing is particularly difficult for second language readers. In the fields of education, psychology, applied linguistics and document design there has been a wealth of research into aspects of L2 students' comprehension of academic writing (e.g. Alderson & Urquhart 1984, Berman 1984, Devine, Carrell & Eskey 1987, and in South Africa, Blacquiere 1989, Pretorius 2000 and Coetzee 2003). These illustrate the extent and nature of the difficulties L2 students have with reading comprehension. Lexical difficulties are addressed first in the discussion below, followed by syntactic causes of difficulty.

At the lexical level, we saw in section 1.2.1 that insufficient specialised academic vocabulary among L2 speakers is a hindrance to comprehension of academic text (e.g. Cooper 1995, Hubbard 1996). Vocabulary levels are acknowledged to be a good predictor of reading comprehension as the two have a reciprocal relationship whereby more exposure to texts improves vocabulary knowledge and good vocabulary is a necessary (but not sufficient) condition for reading comprehension (Thorndike 1973, Daneman, MacKinnon & Gary Waller 1988, Pretorius 2000a:151). Vocabulary levels are also correlated with academic performance, as shown, for example, by Cooper (1995) who noted low academic and basic vocabulary levels among low-performing L2 students at Unisa and Vista universities. Pretorius (2000a) found that borderline and failing L2 undergraduate students also tended to have poor vocabulary inferencing ability (the ability to work out the meaning of unfamiliar words from clues provided in the text) while pass and distinction students had stronger abilities in this component of reading. McNaught (1994) illustrates the problems that African-language speakers in South Africa have in understanding the full import of logical connectives such as *consequently, therefore, although* in science texts. Unfamiliar words are therefore an important cause of

difficulty in texts. My own research tracks this aspect of MCQs by calculating AWL density and the number of unfamiliar words (those not among the 3000 most-common everyday words identified by Dale and O'Rourke (1981)) in each question as well as by think-aloud protocols in which students identify particular words or phrases that they find difficult.

In a paper on the nature and role of syntactic problems encountered by university students in reading foreign language texts, Berman (1984) showed that her final-year Hebrew L1 students studying TEFL at Tel Aviv university experienced difficulty understanding aspects of their study material despite their proficiency in English. Both the findings and methodology of Berman's study are of relevance to my own research.

In her classroom, Berman (1984) observed that homonyms were a particular source of difficulty for L2 students. For example, the homonym *since* tended to be interpreted in its most common sense (time) even if the intended sense was cause. Coetzee's interviews with students in the biological sciences at Vista University confirmed that L2 students in South Africa also experienced problems when the difference between the everyday meaning and the specific technical meaning of homonyms like *tissue* and *coat* was not made explicit (Coetzee 2003:290, cf also Dempster & Reddy 2007). Students also had trouble comprehending academic texts when writers used lexical substitution, using different words or phrases to refer to the same concept (Coetzee 2003:290).

Berman (1984) also indicated that her L2 students struggled to comprehend nominalisations (Berman 1984) as these tend to be abstract academic terms that are informationally dense (see, for example, the italicised nominalisations in the example below (from Appendix A). Nominalisations also alter the canonical NVN structure of sentences, for example, the nominalisation *failure* in option [1] below means that the sentence does not follow the NVN pattern of the verb-based alternative *Speakers fail to pass their language on to their children*. This can impede comprehension for some students and can cause ambiguity if sentence subjects are unclear:

66. Which of the following does **not** contribute to language shift?

[1] A *failure* to pass a language on to one's children
[2] *Industrialisation* and *urbanisation*
[3] *Isolation* from other language groups
[4] When speakers cease to evaluate their own language positively.

Poor knowledge of syntax apparently causes even more problems for L2 students than difficult vocabulary (Berman 1984:147, Dempster & Reddy 2007). Berman (1984) asked her class of Hebrew L1 third-year students to discuss specific difficulties in a text that over half the class considered to be difficult. In a time-consuming but illuminating discussion, students described their difficulty making sense of textbook passages containing substitution of *do* for verbs, or pronouns for nouns (Berman 1984:152). Her students struggled to correctly identify the antecedents of pronouns, thinking for example that *it* in the following example referred to the eye.

> *The eye is exposed to much more information than the brain can possibly use, and the brain selects and processes only as much as it can handle.*
>
> (Berman 1984:150)

Pretorius (2000a) showed that L2 undergraduate students who are weak readers have particular problems in perceiving relations between parts of sentences. McNaught (1994) makes the same point, illustrating the problems that African-language speakers in South Africa have in understanding the full import of logical connectives such as *consequently, therefore, although* and prepositions such as *in, on, at, across* in their science texts. Coetzee (2003:298) also observed that L2 students in South Africa experienced problems with resolving referents and interpreting cause-effect relationships. Pretorius' research with South African L2 undergraduate students showed that while pronouns were usually linked correctly to their antecedents, students were less successful at identifying the antecedents of other kinds of anaphors, including anaphoric synonymy, paraphrase and determiners (Pretorius 2000a). Pretorius (2000a) showed a clear link between academic performance and the ability to resolve anaphors (Pretorius 2000a:139) and recommends explicit teaching of reading strategies (including anaphoric resolution) to low-achieving students.

For Berman's (1984) students, discontinuous phrases, embedded clauses and deletion (e.g. of relatives) proved particularly problematic. For example:

> *The fact that the eye is open and exposed to light is no indication that visual information is being received and processed by the brain.*
>
> (Berman 1984:150)

In this sentence, students failed to understand what was exposed to light due to the ellipsis in the sentence, glossed over the crucial negative *no*, and had difficulty sorting out the subject from the predicate due to the subordinate clauses in both. Coetzee (2003) showed that L2 undergraduate students in South Africa experienced similar problems understanding sentences containing embedded clauses and often failed to retrieve the deleted element in cases of ellipsis. For example, students could not work out the object of the verb *chew* (tobacco? Anything?) in the following example:

> *The workers must neither chew nor smoke tobacco while handling plants or working in the fields.*
>
> (Coetzee 2003:298)

According to Berman (1984), discontinuous constructions, nominalisations, anaphors and embedded sentences make it more difficult for L2 readers to identify the subject, verb and object of a sentence and the agent-action-patient semantic relations which are crucial to understanding:

> […] non-apprehension of SVO relations and basic constituent structure made it impossible for students to get to the propositional core of these rather complicated sentences; truncation or deletion of grammatical markers rather than overt lexical repetition, subordination or pronominalization made it hard for students to perceive relations between parts of sentences; and discontinuous, logically dependent elements were unravelled only after some prompting. (Berman 1984:152)

Many of these 'difficult' constructions (nominalisations, ellipsis, compound nouns, passives) have an indirect link between form and meaning where the more 'usual' semantic category is replaced by a different semantic category (Lassen 2003:25). For example, English typically uses verb phrases for describing processes, noun phrases for participants, adverbial phrases for place, manner and time and co-ordinators to express logical-semantic relations. Less common

constructions, such as as nominalisations to describe processes or passives without agents, have an indirect link between syntactic form and meaning (Lassen 2003) that 'is neither better or worse in itself; but it is more sophisticated, and so has to be learnt' (Halliday 1985:321). This kind of incongruence is problematic for L2 readers but common in academic writing:

> It follows that where there is a conflict between the more basic ordering of semantic and syntactic relations and the surface form of sentences, as in passives, or where material is preposed before the surface subject, or where adverbial clauses precede rather than follow the main clause, readers might be expected to encounter difficulty.
>
> (Berman 1984:140)

As we saw in the 'eye' example earlier, Berman's (1984) study also showed that students reading English textbook passages tended to ignore some negatives, for example … *is no indication that* …, or … *are unaware of.* She concludes that the processing of negatives was an important source of difficulty for her L2 students. This finding is echoed by Kostin (2004), who identified two or more negatives as a particular feature that related to item difficulty of MCQs in TOEFL listening tests. Fellbaum (1987:201) explains why negative MCQs are difficult to understand:

> By far the most common – and more subtle – antonymy is the lexicalised kind expressed either by morphological prefixes (*in-, un-, dis-* etc.) or in words like *lack, lose, contrast with, deny, reject, different from, rare, little* and so on. They build into the sentence an opposition or contradiction that the student must catch, for it is almost always precisely this opposition that is crucial for the choice of the correct option.

This section has focused on some of the vocabulary and syntactic constructions that L2 students struggle to understand in their academic texts. Efforts to understand what makes texts readable have 'catalysed an enormous body of ongoing research into what comprehension means and about the processes of reading and writing in general' (Schriver 2000:140). This includes a large body of work done over the last 70 years in an attempt to predict the success that groups of readers will have with particular texts. This issue is taken up in section 2.3 below.

2.3 Readability and readability formulas

Chall and Dale (1995:80) define readability as 'the sum total (including the interactions) of all those elements within a given piece of printed material that affect the success a group of readers

have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting'. Readability therefore encompasses multiple aspects of a text including its legibility, organisation, content and conceptual difficulty, clarity of meaning and the prior knowledge it assumes from readers (Mobley 1985:44). Chall's preface to *Readability revisited: The new Dale-Chall readability formula* (Chall & Dale 1995:i) explains that 'To him [Edgar Dale, who died in 1985], readability was much more than a research tool. It was a way to extend education and understanding to ever larger numbers of people, and in so doing, to make us all more knowledgeable and more humane.' Harrison and Bakker (1998) explain that definitions of readability are of two kinds. The first is a very broad definition that takes into account 'all the elements which might affect the ease with which readers can comprehend a text' (as in Chall and Dale's definition above) while the other 'is a more practical need to give writers or educators a quick approximate guide to the level of difficulty of texts' and usually involves a formula (Harrison & Bakker 1998:122).

To some extent, readability has become an unfashionable term, replaced more often these days by terms like 'accessibility', 'comprehensibility', 'usability' and 'considerate text' (Chall & Dale 1995:88), but the concept remains a highly topical one, particularly in education, and in recent years there has been renewed interest in devising new readability formulas (Homan, Hewitt & Linder 1994, Harrison & Bakker 1998). Readability formulas are discussed and critiqued in sections 2.3.1 and 2.3.2 below.

2.3.1 Readability formulas

A readability formula is an equation which combines those text features that best predict text difficulty. Based on analysis of a handful of short samples scattered throughout the text, readability formulas allow an estimate to be made of the reading level of the text. For example, the Flesch Reading Ease score (Flesch 1948) rates text on a 100-point scale; the higher the score, the easier it is to understand the document.

Readability formulas are developed by studying the relationship between text features (e.g. length of words, length of sentences) and text difficulty (as measured, for example, by reading comprehension, reading rate, or expert judgment of difficulty) (Chall & Dale 1995:79-80).

Harrison and Bakker (1998) explain that most readability measures follow the same basic pattern: a constant plus a weighted word factor (such as the number of long words per sentence) plus a weighted sentence factor (such as the number of words per sentence) (1998:123). For example, the formula for the Flesch Reading Ease score is:

> 206.835 – (1.015 x average sentence length in words) – (84.6 x average number of syllables per word)

More than a hundred linguistic factors related to difficulty have been identified by readability researchers (Chall & Dale 1995:81). For example, at the lexical level, Klare (2000:19) states that short words, frequent words, Anglo-Saxon words, non-technical words, words used in their common meaning and concrete words all contribute to more readable text. Among the syntactic factors contributing to readability are short sentences, short phrases, sentences with few prepositional phrases and sentences with few compound or complex constructions (Klare 2000:19). Despite the multiple factors that have a bearing on readability, readability can nevertheless be predicted fairly successfully using simple measures such as word frequency and sentence length. This is because text complexity tends to manifest itself in the simpler classic measures of word frequency and sentence length, and so looking at 'surface aspects' of difficulty can apparently be used successfully to predict subtler aspects of comprehensibility (Chall & Dale 1995:99).

More than 50 readability formulas have been published since 1920 in an attempt to provide a quick approximate guide to the level of difficulty of texts, but only a few have been widely taken up. These include Dale-Chall (Dale & Chall 1948, Chall & Dale 1995), Flesch (1948, 1950), Fry (1968, 1977) and Bormuth (1969). Three of the more recent formulas, the new Dale-Chall readability formula (1995), Harrison and Bakker's (1998) suggested readability measures and the Homan-Hewitt formula (Homan, Hewitt & Linder 1994) which is intended specifically for MCQs are discussed in more detail in sections 2.3.1.1, 2.3.1.2 and 2.4.5 below.

2.3.1.1 The new Dale-Chall readability formula (1995)

The new Dale-Chall readability formula (Chall & Dale 1995), like the original Dale-Chall Readability Formula (Dale & Chall 1948), is based on only two factors: word familiarity and

average sentence length. Together these two measures correlate strongly (0.92) with reading comprehension as determined by cloze comprehension scores (Chall & Dale 1995:6), making the Dale-Chall formula consistently the most accurate of the readability formulas (Klare 2000:22-23).

The issue of word familiarity in the Dale-Chall Formula is calculated by recording the number of unfamiliar words in a 100-word sample. Unfamiliar words are defined as those which are not on the 3000-word list known to 80% of American 4th graders (Chall & Dale 1995:16-29). This list is based on familiarity scores from the empirically validated *The Living Word Vocabulary* (Dale & O'Rourke 1981). Apart from a few American English terms (e.g. *billfold, boxcar, burro, bushel, catsup, crosswalk, jack-o-lantern* etc.), these are simple and generic enough to be applicable in other English-speaking contexts, especially as the (unfamiliar) American English words would simply not occur in texts from other countries. Regular morphological variants such as plurals, possessives or past tense forms of the listed words are counted as familiar, as are proper names after their first (unfamiliar) occurrence.

Sentence length in the Dale-Chall readability formula is calculated by counting the number of complete sentences in a 100-word sample. Hyphenated nouns like *lady-in-waiting,* contractions and acronyms are counted as single words, but hyphenated adjectives like *anxiety-provoking* are counted as two words. (Note that word-processing programs such as Microsoft Word can be used to facilitate readability analyses, e.g. by counting words automatically or measuring word frequency, average sentence length, average word length, etc. or by calculating readability scores based on the Flesch Reading Ease score.)

Once the number of unfamiliar words and number of complete sentences has been calculated, these two values are simply read off tables, resulting in a readability score between 1 and 16. These reflect the reading ability expected at American grade levels ranging from 1 (first grade, rudimentary reading) to 16 (College graduate level, advanced, specialised reading). These scores can be used in various practical ways, for example to select appropriate readers or comprehension texts for school children of various ages and abilities. In my own study the Dale-Chall reading levels of individual questions will be used to compare their readability.

2.3.1.2 Harrison and Bakker's readability formula (1998)

Harrison and Bakker (1998) have developed readability software that runs within Microsoft Word and calculates both traditional and their own suggested readability measures. Harrison and Bakker (1998:132) showed that text containing long sentences that are broken up into 'packets' using one of eight punctuation marks, namely . , : ; (– ! ? continues to be readable. They conclude that sentence length and packet length taken together might be a good predictor of readability, particularly for the evaluation of English passages intended for L2 readers (Harrison & Bakker 1998:137).

Harrison and Bakker (1998) suggest that the syntactic measures above should be supplemented with a measure of lexical density. Lexical density affects the accessibility of text as it reflects the level of information content for a given number of words. Following Halliday (1989), Harrison and Bakker (1998:125) indicate that more readable texts tend to be more similar to oral language in their lexical density, containing a lower ratio of lexical items to grammatical items. Using lexical items per clause as their measure of lexical density (Halliday 1989), Harrison and Bakker (1998:131) found that university students perceive lexically less dense texts as easier to read even when this was at variance with predictions made by readability indicators like the Flesch Reading Ease score. Harrison and Bakker (1998:136) conclude that lexical density is a powerful readability concept which can give writers and editors valuable feedback on prose passages. While I do not use the Harrison and Bakker readability formula, the prominence they afford to lexical density will be followed up in my study by calculating the AWL density (number of academic words as a percentage of total words) in each MCQ.

2.3.2 Criticisms of readability formulas

Although readability formulas continue to be used by educators and publishers, and are experiencing a surge of renewed interest, readability research lost impetus in the 1980s. This was partly due to concern over misuse of readability scores and partly to increasing interest in less quantifiable, cognitive aspects of readability in the 1980s and 1990s. Schriver (2000:138) sums up many of the points of criticism against readability formulas. These include the contention that sentence length and word frequency are not the best predictors of comprehension, that formulas do not measure complex issues such as idea density and that formulas ignore reader-specific

factors such as motivation, interest and the reader's intended purpose in reading the text. All of these issues are discussed in more detail below. These criticisms were typical of the new paradigm in readability, which drew on ideas from linguistics and cognitive psychology and stressed the role of cohesion, idea density and relationships between ideas as important components of readability (Chall & Dale 1995:93). Kintsch, Kemper and Armbruster are some of the key figures in the 'new' readability paradigm (see e.g. Kintsch & Miller 1981, Huckin 1983, Armbruster 1984, Kemper 1988). The collected articles in Davison and Green (1988) *Linguistic complexity and text comprehension: Readability issues reconsidered* reflect this new paradigm in their attempt to link readability research more closely to recent research on reading and language processing. The articles in Davison and Green (1988) are highly critical of traditional readability formulas for a number of reasons, some of which are summarised in (a) to (d) below:

(a) Readability formulas oversimplify the issue of readability by focusing exclusively on sentence length and word difficulty

Anderson and Davison (1988:28) suggest that sentence length and word difficulty may not in fact be the best indicators of readable text. They point out that long, rare words like *unladylike* and *helplessness* are usually semantically transparent derivatives of frequent words (in this case, *lady* and *help*). As we saw in section 2.2.1, regular members of a word family tend to be understood without difficulty (Bauer & Nation 1993) and cause problems only for weak decoders who have difficulty segmenting words morphologically (Anderson & Davison 1988). Anderson and Davison (1988:30-31) also cite research by Freebody and Anderson (1983) which showed that 6th grade texts were only 4% less comprehensible when up to a third of the content words were replaced by more difficult synonyms e.g. *descending* for *falling*, *pulverise* for *grind* etc. Freebody and Anderson concluded that 'it takes a surprisingly high proportion of difficult vocabulary items to create reliable decrements in performance' (Freebody & Anderson 1983:36).

Similarly, sentence length alone accounts for a very small percentage of the variance in comprehension, with long sentences often understood just as well as short sentences, except by very weak readers. For example, the presence of connectives like *so, or,* and *because* in long sentences actually facilitate comprehension by making the meaning relations between clauses

explicit and requiring readers to make fewer inferences (Anderson & Davison 1988:32). Harrison and Bakker (1998:135) point out that average sentence length (as used for readability calculations) is often quite different from most frequently occurring sentence lengths since there are often a few very long or short sentences that affect the average.

Bruce and Rubin (1988:7) point out that

> because most readability formulas include only sentence length and word difficulty as factors, they can account only indirectly for factors which make a particular text difficult, such as syntactic complexity, discourse cohesion characteristics, the number of inferences required, the number of items to remember, the complexity of ideas, rhetorical structure, and dialect.

Chall et al. (1996:67-69) suggest that as the difficulty of texts increases on the linguistic dimensions of vocabulary and sentence length and complexity, they also tend to become more difficult with regard to the density and complexity of ideas and the prior knowledge needed by the reader to understand ideas not specifically explained in the text. More advanced comprehension strategies such as reasoning, critical thinking, evaluating and predicting are required to understand more difficult writing.

To address the limitation of traditional readability formulas, Chall et al. (1996) offer a qualitative method for assessing text difficulty more holistically and impressionistically, as teachers, writers and librarians have always done. They provide exemplars of narrative and expository texts at different levels from first grade to postgraduate that can be used as benchmarks for estimating the comprehension difficulty of other texts. These graded extracts also sensitise teachers and parents to what readers need to be able to cope with as they advance through different levels of education. It should be pointed out that the reading levels assume mother-tongue accuracy and fluency and are therefore not unproblematic when transferred to other, more diverse contexts, where both writers and readers of texts may be L2 speakers. For L1 speakers, however, Chall et al. (1996) note that adult readers' estimates of text difficulty tend to correlate very well with quantitative readability scores and with judgments of text difficulty based on cloze scores and oral reading errors. Qualitative procedures for assessing text difficulty have the advantage that they are simple and quick to apply, and 'can be more sensitive to the great variety of text

variables that differentiate texts, including vocabulary, syntax, conceptual load, text structure and cohesion' (Chall et al. 1996:2). The main differences between text difficulty at different levels described in Chall et al. (1996) are summarised below:

> Levels 1 and 2 use familiar, high-frequency vocabulary and simple language that tends to be descriptive, informal and repetitive, for example short sentences joined with *and.*
>
> At Levels 3-4 more content-related vocabulary is introduced and explained or can be inferred from surrounding text. The sentences are longer and the description is more detailed.
>
> Levels 5-6 and 7-8 display a more formal writing style. Vocabulary tends to be less familiar and more abstract and fewer words are defined. Ideas as well as facts and events are presented, requiring more advanced reading comprehension, prior knowledge and critical thinking.
>
> Levels 9-10 and 11-12 also present more than facts, with ideas being compared and interpreted. Knowledge of unfamiliar words is expected, there is an increase in technical vocabulary and sentences are often long and complex.
>
> Levels 13-15 and 16+ demand even more in terms of vocabulary, syntax and particularly interpretation. Both general and content-related vocabulary is difficult, words have multiple meaning and connotations, idea density is high, sentences are long and complicated. Readers have to be active, interested and bring a broad and deep understanding of general knowledge to their reading.

(b) Reader-specific factors are not taken into account by readability formulas

As Chall et al. (1996) acknowledge, the prior knowledge of readers is an important factor contributing to text accessibility. Anderson and Davison (1988:44-45) show that comprehension will vary depending on the match between readers' prior knowledge and the knowledge presupposed by a text. The concept of 'schema theory' from cognitive psychology postulated the

idea that humans construct abstract cognitive templates of semantic context on the basis of experience and store these in long-term memory. For example, our seaside schema (or 'frame' or 'script') will be activated to help us to make sense of a biology text on tides and tidal pools. Schemas that are shared by readers and writers are used to make inferences that help us to understand text even when the writer's assumptions are not made explicit (Huckin 1983:92).

In addition to prior knowledge, it has been shown that a text that is perceived as having uninteresting content and boring presentation is less well understood than a text which falls within a particular reader's interests. At a stylistic level, texts that are high in 'human interest', containing many names, first and second person pronouns, spoken sentences and questions, are also more interactive and engaging and therefore more readable (Flesch 1962). Klare (2000:18) reminds us that readability formulas can measure stylistic aspects of difficulty but not content, which is equally important to overall readability.

(c) Working memory and text-processing constraints are an important element of difficult text that readability formulas do not take into consideration

Anderson and Davison (1988:40) make the important point that sentences that are difficult to understand are often a result of the interaction of several factors all being processed at once in some limited space in working memory. This point is particularly relevant to MCQs, which require students to hold a stem in working memory while reading through several possible answer choices.

For example, Anderson and Davison (1988:35-37) show that phrases with left-branching structures, i.e. grammatical elements that precede the head of the phrase, are harder to process, slower to read and recalled less well than phrases that follow the normal head-first order. An example of a left-branching structure is *Because Mexico allowed slavery, many Americans and their slaves moved to Mexico* (as opposed to the more usual *Many Americans and their slaves moved to Mexico because Mexico allowed slavery*). Preposed adverbial clauses and other left branches must be held in working memory until the main clause constituents are found. These can overload temporary processing capacity, especially if there are other processing complications like pronouns for which antecedents must be identified.

Kemper (1988:152) suggests that texts are difficult to comprehend when the reader is required to make too many causal inferences, e.g. when not all the actions, physical states and mental states in the underlying schema are explicitly stated in the text. Although readers attempt to fill in the gaps, reading speed and comprehension are reduced. Kemper (1988) found that simplifying the words and sentences of a text has little effect on the speed or accuracy of subjects' answers to questions about a text. Performance is similar for texts written at an 11th grade reading level and ones revised to a 7th grade reading level as measured by the Flesch readability formula. Flesch readability affects text comprehension only when texts are already difficult to understand because they require high levels of inferencing (Kemper 1988:162).

Urquhart (1984) reports on experimental research into the effect of rhetorical ordering on readability that confirmed that time-ordered and spatially-ordered texts were quicker to read and easier to recall than non-time-ordered texts (Urquhart 1984:167). Time order should rhetorically follow the same sequence as the events described, while spatial order should be described in a unilinear way, e.g. left to right or front to back. These 'natural' principles of organisation are also mentioned in plain language guidelines and guidelines for MCQs, which recommend listing options in numerical order or in some other logical way, such as from specific to general (Haladyna 2004:113, cf also section 1.2.2.1).

(d) Misuse of readability formulas and scores

As can be seen from (b) and (c) above, language processing constraints and reader-specific factors have a major bearing on text comprehension which is not factored into readability formulas. Another major problem with readability formulas according to Schriver (2000:138-139) is that they can be and have been misused by writing-to-the-formula. For example, simply splitting up sentences and adding more full stops gives you a better readability score but almost certainly a less comprehensible text. Readability formulas are often misused as guidelines for revision, even though most were intended not as writing guidelines but as predictors of text difficulty *after* a text has been written (Harrison & Bakker 1998:124). Klare (2000:4) also reminds us that readability formulas cannot measure difficulty perfectly because they are merely estimates, and finally that they can never be seen as measures of *good* style (Klare 2000:24-25).

Despite these criticisms and caveats, however, the classic readability measures continue to be used because of their predictive power and their ease of application. Classic readability measures can usefully be combined with measures of idea density and cohesion to get a broader picture of text accessibility. For example, the later version of the Dale-Chall formula (Chall & Dale 1995) includes greater provision for reevaluating readability scores in the light of qualitative judgements of reader characteristics such as prior knowledge, interest in the topic and the way it is presented, and the effect of text organisation, conceptual difficulty and density on passage difficulty (Chall & Dale 1995:10-11).

2.4 The language of multiple-choice questions

While MCQs have many of the characteristics of academic texts (see section 2.2 above), they also display particular characteristics that make them a genre of their own. These include incomplete sentence stems, several possible sentence-endings for a stem, resulting in long sentences, and anaphoric options like All of the above (AOTA) and None of the above (NOTA) which require students to consider all the previous answer choices in their deliberations. Item-writing guidelines such as those listed in 1.2.2.1 address the unique structural properties of MCQs and assist test-setters to make them as clear as possible. Empirical research on these guidelines is reviewed in 2.4.1 below, followed by a discussion on trick questions in 2.4.2, and on how L2 students cope with MCQs in 2.4.3. Section 2.4.4 considers how Grice's conversational maxims can inform MCQ design and 2.4.5 looks at readability formulas for MCQs.

2.4.1 Empirical research on item-writing guidelines

Haladyna and Downing (1989:72) observed in 1989 that that few item-writing 'rules' had received adequate empirical study, but a growing body of empirical research since then has focused on whether or not contraventions of particular guidelines do cause problems for students and affect test validity and reliability (e.g Rich & Johanson 1990, Kolstad & Kolstad 1991, Tamir 1993, Varughese & Glencross 1997, Harasym et al. 1998). It should be noted that some of the empirical research on language difficulties in MCQs focuses on language testing and reading comprehension MCQs (e.g. Bejar 1985, Fellbaum 1987, Rupp, Garcia & Jamieson 2001, Kostin

2004) rather than on content subject MCQ testing. These situations are different in that the former are testing language proficiency as the construct, while the latter are usually not.

One empirical study that is relevant for my study is Sireci, Wiley and Keller's (1998) investigation of the item quality of 285 Accountancy MCQ items from the CPA professional accountants examination in the United States. They compared the item quality (as measured by difficulty and discrimination) of MCQs that violated none, one or more than one of the guidelines proposed by Haladyna and Downing (1989). A similar methodology will be used in my own study to identify which of the MCQ language guidelines have empirical support. Sireci, Wiley and Keller (1998) do point out however that

> (t)he use of item difficulty and discrimination statistics to evaluate item quality also has limitations. Items that violate item-writing guidelines may undermine test validity in ways that do not show up in item statistics. For example, some items may facilitate test anxiety. Others may take longer to answer.
>
> (Sireci, Wiley & Keller 1998:8)

They suggest that item-writing guidelines could also be researched using timed response studies, experimental studies that compare parallel versions of items that do and do not violate guidelines, or, as in my own research, think-aloud interviews with students.

Because there were no AOTA or NOTA items in the CPA test, and insufficient negative items, these guidelines could not be tested. The guidelines that Sireci, Wiley and Keller (1998) did investigate statistically were therefore 'Avoid complex Type-K items' (see section 1.2.2.1 above) and 'Avoid incomplete statement stems'. The 22 type-K items appeared more difficult and less discriminating, suggesting that the guideline to avoid this format has some validity and that these items might be confusing for test takers. More relevant to my study is their finding that although two-thirds of the items used the incomplete stem format, this did not have any substantive effect on either the difficulty or the discrimination of the items and that there was therefore no evidence to support the incomplete statement guideline. There was also some preliminary evidence that when an item violates more than one guideline, item quality may diminish, leading to their suggestion that 'there may be a cumulative effect when an item violates more than one guideline' (Sireci, Wiley & Keller 1998:6). This latter suggestion will also be followed up in my discussion.

Further justification and some empirical research findings relating to the eight MCQ guidelines that form the focus of the study are described in (a) to (h) below:

(a) Avoid negative words such as *not* or *except* in the stem

The rationale for the guideline is that the negative word or prefix may be overlooked by students, resulting in a wrong answer. While typographic emphasis using capitals or boldface can mitigate this risk to some extent, students may also have difficulty understanding the meaning of negatively phrased items because they are more difficult to process (Harasym et al. 1993, Tamir 1993, Haladyna 1994:73, Haladyna 2004:111). Haladyna, Downing and Rodriguez' (2002) survey of the literature indicated that 63% of authors supported this guideline, with 19% viewing negatives as acceptable and 19% not mentioning this guideline at all (Haladyna, Downing & Rodriguez 2002:316).

Although it is generally believed that negative questions lead to worse performance by students (Cassels & Johnstone 1980, Haladyna, Downing & Rodriguez 2002, Haladyna 2004:111), the empirical findings are somewhat mixed. For example, Harasym et al. (1992) found some indication that negatively worded stems were *less* difficult than positive stems. In a small-scale South African study, Varughese & Glencross (1997:179) found no significant difference between student performance on Biology MCQs in the positive mode (e.g. *Which of the following is correct?*) and the negative mode (*Which of the following is false/incorrect?*) and noted that in fact some questions were easier when asked in the negative mode. Tamir (1993) found that lower-level questions are answered equally well in the positive and negative mode but that higher-level questions with more processing steps are significantly more difficult in the negative mode. Tamir (1993:312) and Varughese and Glencross (1997:179) also make the point that negative questions such as *Which of the following is false?* have the advantage of exposing students to more right options than wrong options. Haladyna (2004) also recommends avoiding negative words in the answer choices for the same reasons described above.

(b) State the stem in a question format instead of an incomplete sentence format

Haladyna (1994:70) argues that questions are a more direct way to elicit a response from students than incomplete statements. Incomplete statements may be more difficult to process as

they need to be kept in short-term memory while reading the various possible options, or requiring the student to skip back-and-forth between the stem and the options. Cloze formats with a missing word in the middle of the stem also suffer from this drawback and sometimes make it unclear to students what the question is actually about. Question stems are complete sentences and therefore provide a fuller context for activating relevant cognitive schemas for answering the item. Haladyna (2004:108) replaces this guideline with a guideline in favour of 'focused' stems that contain the main idea versus unfocused stems that fail to indicate to students what the question is about. For example, the focused question stem in (a) is preferable to the unfocused incomplete statement stem in (b) below:

(a) What is corporal punishment?

A. A psychologically unsound form of school discipline
B. A useful disciplinary technique if used sparingly.
C. An illegal practice in our nation's schools.

(b) Corporal punishment

A. has been outlawed in many states
B. is psychologically unsound for school discipline
C. has many benefits to recommend its use.

(Haladyna 2004:109)

(c) Simplify vocabulary and aim for maximum readability

A test is intended to test subject ability, not reading ability, and unnecessarily difficult or unfamiliar vocabulary will affect test validity by disadvantaging weaker readers and L2 students (Haladyna 1994:66, 2004:106). Fellbaum (1987:203) states that frequency of occurrence has a direct bearing on the familiarity of a lexical item and hence on the degree of difficulty in comprehending or completing a sentence containing it. We will see in section 2.4.4 that L2 performance on maths and science questions improved significantly when the wording of questions was simplified (Bird & Welford 1995, Prins & Ulijn 1998), while Abedi et al. (2000) showed that simplifying vocabulary or providing glossaries improved the test performance of both L1 and L2 students.

(d) Keep items as brief as possible

Wordy items that require extended reading time decrease content coverage and reliability (Haladyna 1994:64, 2004:106-110). Fellbaum (1987) showed that there was no link between difficulty of the item and the overall number of words/clauses but that MCQs containing the complementiser *that* tended to be difficult. Bejar (1985:291) comments of MCQ comprehension items that '[…] it seems reasonable to suggest that if comprehensibility of a sentence is affected by some measure of syntactic and semantic complexity then psychometric difficulty of an item based on that sentence will to some extent also depend on the syntactic and semantic complexity of the sentences'.

(e) Avoid very similar answer choices

Paxton (2000:112) interviewed University of Cape Town students about their views on multiple-choice assessment, finding that some L2 students reported difficulties in distinguishing between options that were very similar in meaning. Roberts (1993) reported that student and lecturer perceive items with very fine discrimination between items as 'tricky' (see section 2.4.2 below).

(f) Try and keep items grammatically parallel

While there doesn't seem to be empirical research on this issue, Haladyna recommends keeping options homogeneous in grammatical structure to avoid accidentally alerting students to the odd-one-out, which is often the right answer. For example, the single-word noun phrase in option [4] contrasts with all the other options and suggests that [4] is the correct answer:

What reason best explains the phenomenon of levitation?

[1] Principles of physics
[2] Principles of biology
[3] Principles of chemistry
[4] Metaphysics

(g) Avoid All of the above (AOTA)

Testwise test-takers know that when AOTA is offered as an option, it is usually the right answer (Harasym et al. 1998). Secondly, AOTA questions can reward students for partial knowledge, in that if the student recognises that options 1 and 2 are correct, but is not sure about option 3, he or

she can still deduce that the answer must be AOTA (Haladyna 1994:78, Haladyna 2004:117). Sireci, Wiley and Keller (1998:8) cite Gross (1994), who argues that item-writing guidelines like avoid AOTA and NOTA are defensible logically and are not in need of empirical support because by design they diminish an item's ability to distinguish between candidates with full versus partial information.

(h) Keep None of the above (NOTA) to a minimum

According to Haladyna (1994:78), testwise test-takers know that when NOTA is offered as an option, it is usually not a real option but has just been added due to a failure of creativity on the part of the test-setter. He suggests that distractors that are implausible or ignored by most students should rather be omitted as they serve only to increase the amount of reading in the test (Haladyna 1994:78). As we saw in (g), Gross (1994) argues that NOTA should be avoided because it rewards students for partial information. NOTA items tend to be more difficult than average but not more discriminating (Crehan & Haladyna 1991, Frary 1991, Crehan, Haladyna & Brewer 1993). However, Rich and Johanson (1990), Frary (1991), Kolstad and Kolstad (1991), Knowles and Welch (1992) and Dochy et al. (2001) suggest that NOTA questions can be useful when used in moderation. By the third edition of *Developing and validating multiple-choice test items* in 2004, Haladyna had adapted his guideline to suggest that NOTA should be kept to a minimum rather than avoided entirely (2004:116-117).

Research findings relating to the eight guidelines in the study have been presented above. While some of the research findings relating to particular guidelines are mixed rather than conclusive, most test writers would not want to be accused of setting 'unfair' or trick questions. McCoubrie (2004) explains that:

> Subtle grammatical chicanery, particularly the use of negatives and imprecise terms (e.g. *frequently*) may cause confusion amongst examinees. Any confusion over grammar or question structure invalidates the test as (1) this extra grammatical variable does not relate to knowledge of the subject, (2) it discriminates against examinees for whom English is not their first language and (3) benefits the wily and experienced examinee.
>
> (McCoubrie 2004:710).

The next section takes up this issue of what constitutes an 'unfair' question.

2.4.2 Trick questions

Roberts (1993) asked students and lecturers to try to define the concept of a 'trick question'. In their responses, students indicated that multiple-choice and true/false test items tended to be trickier than other types of test questions (Roberts 1993:332) and that the following contributed to 'trickiness' in questions:

(a) trivial content focusing on unimportant detail
(b) testing content in the opposite way to which it was taught
(c) answers that require more precision than was required in the course material
(d) 'window dressing', i.e. irrelevant or excess information in the stem, particularly when the topic on which the information is provided is not the focus of the question
(e) an intent to mislead or confuse (e.g. using a trivial word that may be easily overlooked or a confusing use of the negative that actually turns out to be crucial)
(f) answer choices that look a great deal alike and have only slight differences
(g) highly ambiguous items that everyone has to guess.

(Roberts 1993:334)

While the first four of these deal mainly with content, the last three points deal directly with the wording of the question. Roberts (1993:343) concludes that trickiness is multifaceted: deliberate intention is important, but so is ambiguity, which is usually unintended. In fact, the examples that Roberts gives of intentional trickery, using a trivial word that may be easily overlooked or a confusing use of the negative that actually turns out to be crucial, may just as easily be unintended. Although the current study does not focus on deliberate attempts to mislead students, Roberts' (1993) provides as interesting typology of reasons why questions can turn out to be more difficult than intended. In particular, Roberts' categories of easily overlooked words/negatives, very similar options and highly ambiguous items are all linguistic issues that may be associated with 'difficult' items that most students get wrong or that are associated with a large average score difference between L1 and L2 candidates. Roberts (1993:344) concludes that 'Future work should focus on trying to clarify the definition, if possible, and then see if trickiness is a function of factors like item difficulty, length of items, or complexity.'

2.4.3 MCQs and L2 students

Specific studies on linguistic aspects of MCQs that affect comprehension and performance by L2 speakers are fairly limited but include attempts to relate linguistic characteristics of test items to students' comprehension (e.g. Bosher & Bowles 2008, Rupp, Garcia & Jamieson 2001, Lampe & Tsaouse 2010). There is also a body of research on the effects of text simplification on L2 test comprehension and performance (cf. Berman 1984, Bird & Welford 1995, Prins & Ulijn 1998, Brown 1999, Abedi et al. 2000). Key findings of this research are discussed below.

In their research with South African high school pupils, Prins and Ulijn (1998:145) demonstrated that aspects of the ordinary language of mathematics texts often cause unnecessary comprehension difficulties to students, which in turn influence achievement, particularly for L2 readers. Their findings showed that high-achieving mathematics students all did equally well on purely numerical versions of mathematical problems, but that African language students did much worse than English or Afrikaans students when these were turned into English word problems (Prins & Ulijn 1998). Using a taped think-aloud protocol, readability problems were identified by disturbances in the normal reading rhythm, including rereading, stumbling over words, heavy breathing or overt complaints (Prins & Ulijn 1998:147). This method of identifying readability problems is also used in my own study. Prins and Ulijn (1998) noted that L2 students experiencing comprehension difficulties tend to make use of strategies like rereading, reformulating, adopting a slower reading rate or translating into the mother tongue. African L2 students experienced more readability problems, for both linguistic and cultural reasons, than either of the other two groups. All students experienced difficulty understanding unfamiliar terms like *optimal search line* and *percentage daily capacity*, as well as with the structural organisation of questions where crucial information was mentioned only right at the end of a long question. L2 students had particular problems with information that was too abstract, too condensed or about which they had insufficient prior knowledge.

Prins and Ulijn (1998) conducted an experimental follow-up that involved simplifying the language of the questions on the basis of suggestions obtained from the think-aloud protocols to ascertain whether improved readability would improve test scores. A group of 300 students from 12 different schools with a scholastic achievement level of between 55% and 75% wrote tests

containing a mix of original, linguistically simplified and numerical versions of mathematical problems. The linguistically simplified version improved test scores by an average of 14% for African language students, an average of 19% for Afrikaans students and an average of 12% for English L1 students (Prins & Ulijn 1998:148). Everyone therefore benefited from the increased readability, with L2 speakers benefiting more than L1 speakers.

In a similar study, Bird and Welford (1995) also showed that performance on science questions answered in English by Botswana learners was significantly improved when the wording of questions was simplified. At the syntactic level, Berman (1984) reports on a small-scale experiment involving syntactic simplification of a 300-word text while leaving vocabulary items intact. The simplified version deliberately reduced instances of ellipsis, pronominalisation and substitution. Hebrew L1 third-year students who read the simplified version did consistently better on MCQs and written questions relating to understanding the facts, gist and specific anaphors in the text.

Abedi et al. (2000) compared the effects of four different accommodation strategies – simplified English, a glossary, extra time, and a glossary plus extra time – on the performance of L2 students in maths word problems in the U.S.A. The students' scores improved with all these accommodations except the glossary only, possibly because consulting the glossary took up valuable test time. The glossary and extra time gave the greatest score improvements to both first and second-language speakers, but the only accommodation that narrowed the gap between L1 and L2 speakers was simplified English.

This issue was also addressed in a study by Bosher and Bowles (2008) that investigated difficulties experienced by L2 students answering MCQs in an undergraduate nursing course in the USA. Two of their research questions are of direct relevance to my research, namely 'What is the relationship between the linguistic complexity of test items and students' comprehension of those test items?' and 'From an L2 student's perspective, what makes a test item difficult to understand?'

Bosher and Bowles' (2008) research involved five students engaging in a think-aloud protocol about items they had answered *incorrectly* and found that 35% of all flaws in their MCQ test items were due to 'unnecessary linguistic complexity in the stem or options, grammatical errors, and lack of clarity or consistency in the wording' (Bosher & Bowles 2008:165). In Bosher and Bowles' view, a number of factors contribute to linguistic complexity for both L1 and L2 speakers, including word frequency, word length, and morphological complexity (2008:168) as well as syntactic factors such as 'passive voice constructions, long noun phrases, long question phrases, comparative structures, prepositional phrases, conditional clauses, relative clauses, subordinate clauses, complement clauses, modal verbs, negation, and abstract or impersonal constructions' (Bosher & Bowles 2008:165).

A group of experts on course content, language, L2 reading comprehension and testing then collaborated to modify the MCQs that the students had answered incorrectly to improve comprehensibility. Their modifications included glosses for low-frequency nonessential technical words and decreasing average sentence length from 15,3 to 11,8 words. Bosher and Bowles (2008:169) showed that the five L2 students identified the revised items as more comprehensible 84% of the time. Students commented that the modifications that assisted their understanding of the questions were using shorter, simpler sentences rather than longer, more complex sentences, stating information directly rather than hiding it in the sentence, using questions rather than the incomplete sentence format and highlighting key words such as *most*, *best* and *first* (Bosher & Bowles 2008:168).

In a study that followed on from that of Bosher and Bowles (2008), Lampe and Tsaouse (2010) showed that nursing students answering textbook MCQs had difficulty comprehending abstract or unfamiliar terms like *index of suspicion* or *measurable loss of height*, sometimes missed important terms like *most*, *first* or *best* in MCQ stems and struggled to identify the focus of incomplete statement stems like

> *A patient with respiratory disease has a shift to the left in the oxygen-hemoglobin dissociation curve. The nurse recognizes that this finding indicates that:*
>
> (Lampe & Tsaouse 2010:64)

In their view, the incomplete statement format means that it takes extra time to work out what the question is, further pressurising L2 students who are already at a time disadvantage (Lampe & Tsaouse 2010:64). The relative clauses in the last sentence above are presumably also part of the problem, as they are in the following example stem:

> *An inappropriate reaction by the immune system [in which antibodies form against self-antigens] [mistaken as foreign] best describes:*
>
> (Lampe & Tsaouse 2010:65)

In a larger-scale quantitative study, Rupp, Garcia and Jamieson (2001) looked at 214 reading and listening comprehension multiple-choice items completed by 87 L2 speakers to try to account for which characteristics of the comprehension text, the MCQ item, and the interaction between text and item were associated with item difficulty. Conceding that much remains unknown in what aspects of the MCQ item create difficulty, Rupp, Garcia and Jamieson showed, using two different statistical methods, that the most likely candidates associated with item difficulty were the word length of the stem, the language difficulty of the stem, negation in the stem, the degree of similarity within answer choice sets, the number of words in incorrect options, negation in incorrect options and the number of plausible distractors (Rupp, Garcia & Jamieson 2001:187). These are exactly the kinds of issues that the MCQ guidelines attempt to minimise. They observed that most difficult items combined the features of longer sentences, high information density, complex processing, high type-token ratio (number of unique words divided by total number of words) and greater lexical overlap between distractors and correct answer (Rupp, Garcia & Jamieson 2001:208). The latter was measured in terms of the number of distractors with at least one content word in common with correct answer. These presumably require close reading and fine discrimination to identify the correct answer (Rupp, Garcia & Jamieson 2001:196). Rupp, Garcia and Jamieson (2001) suggest that a combination of difficult factors can lead to processing overload and to students misunderstanding or giving up on the question.

As this overview has shown, MCQs are difficult for L1 and particularly for L2 readers because of academic language features such as specialised vocabulary, long sentences, high density of information, extensive use of anaphor and its often complex and incongruent syntactic structure. All of these features, particularly in combination, can make MCQs harder for students. Studies

such as Prins and Ulijn (1998), Abedi et al. (2000) and Bosher and Bowles (2008) reaffirm that linguistic simplification assists all students but particularly L2 students, and that all test writers should make a conscious effort to maximise the clarity and readability of their papers (see also Ghorpade & Lackritz 1998, Carstens 2000, Fairbairn & Fox 2009).

Section 2.4.4 provides a discussion of the relationship between MCQ guidelines and Grice's conversational maxims. If these are followed, both can improve clarity and communicative effectiveness by tailoring discourse to the needs of the person being addressed.

2.4.4 MCQ guidelines and Grice's cooperative principle

The question and answer format of MCQ assessment is rather like a conversation between lecturer and student because of the turn-taking involved in reading individual questions (the examiner's 'turn') and then internally deliberating and selecting the answer (the student's 'turn'). As a result, I believe it is interesting to relate the guidelines about multiple-choice item-writing to Grice's (1975) cooperative principle (quoted in Sperber & Wilson 1986:33), which also takes the form of guidelines, in this case about how speakers can best make their conversation relevant and appropriate to hearers. Although the literature on MCQ writing does not make the connection between Grice's co-operative principle and the many MCQ guidelines in existence, the similarities are striking in my opinion and worthy of further comment.

Grice's co-operative principle states that a speaker should 'Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged'. The co-operative principle is enacted in conversation by adhering to the following nine maxims:

Maxims of quantity

1. Make your contribution as informative as is required.
2. Don't make it more informative than is required.

Maxims of quality

3. Don't say what you believe to be false
4. Don't say that for which you lack adequate evidence.

Maxim of relation

5. Be relevant

Maxims of manner

6. Avoid obscurity of expression
7. Avoid ambiguity
8. Be brief
9. Be orderly.

(Grice 1975)

In some ways MCQs are by their very nature in conflict with many of these conversational maxims and textual expectations and therefore constitute rather 'uncooperative' text. At the most obvious level, the maxim of quality 'Don't say what you believe to be false' is deliberately flouted in multiple-choice questions, where both right and wrong answers to each question are listed for students to choose from. The fourth maxim, 'Don't say that for which you lack adequate evidence', is also often flouted when writing distractors. Adherence to the other maxims, however, is an important step towards fair multiple-choice and many of these maxims echo the guidelines for setting good questions as discussed in section 2.4.1 above.

As regards the maxims of quantity, the first maxim is 'Make your contribution as informative as is required'. For an MCQ to be sufficiently informative there needs to be at least one right answer choice and a focused stem that makes it clear to students what the question is about. Haladyna (1994:71) explains that the test taker should always know what is being asked in the item, preferably by the time he or she has finished reading the stem. This is one of the reasons why he recommends question stems rather than incomplete statements. The second maxim tells us that our conversational contribution, in this case the MCQ item, should not be 'more informative than is required'. This picks up on issues like the recommendation to avoid excessive verbiage in the stem (Haladyna 1994:72) or 'window dressing' as described by Roberts (1993), which inveigles test-takers into evoking 'garden path' cognitive schemas that are then not followed up in the question and answer choices. Overly long questions or questions with too many options would also be inconsiderate in providing more information than required.

As mentioned above, the maxims of quality are flouted in MCQs. However it is possible to make an effort to avoid saying things 'for which you lack adequate evidence'. For example, overgeneralisations and opinion-based questions like 'Which is the best comedy film ever made?' (Haladyna 2004:103) are indefensible as they lack objective supportive evidence.

The maxim of relation tells us to 'Be relevant'. Inconsiderate MCQs would flout this maxim by providing 'humorous' or ridiculous distractors, window dressing in the stem or superfluous distractors like NOTA.

The maxims of manner include 'Avoid obscurity of expression'. This would include following plain language principles, avoiding unfamiliar non-technical words and complex sentences and ensuring that questions are readable. Maxim seven, 'Avoid ambiguity' is contravened by MCQs containing pronouns without clear antecedents and by lack of clarity generally, for instance questions containing alternatives that are only just wrong, technical terms that are not the same as those used in the study material or unnecessary complexities such as later answers that are 'more correct' than earlier correct answers. Maxim eight, 'Be brief' is not met when questions are long-winded with long sentences, many options or unnecessary repetition of shared material in each option instead of relocating it to the stem as Haladyna (1994:73) suggests. The final maxim of manner, 'Be orderly', is contravened when the ordering of the options is illogical. Haladyna (1994:75) recommends that options should be ordered in logical or numerical order to prevent students from hunting unnecessarily for the answer. Urquhart (1994) agrees, indicating that ordered texts are quicker to read and easier to recall.

Intentional trick questions would be extreme examples of inconsiderate text (cf Armbruster 1984) and violation of Gricean maxims while unintended trick questions would violate Gricean maxims and MCQ guidelines to a lesser but still unacceptable extent. Adherence to all of these maxims (except the maxims of quality) may therefore help in setting fair multiple-choice questions.

2.4.5 Readability formulas for MCQs

In recent years, readability as a specific issue in MCQ validity is beginning to be acknowledged. Allan, McGhee and van Krieken (2005:2) point out that readability, and factors connected to it, can form a significant illegitimate source of difficulty in examination questions. As far as measuring readability of MCQ test items is concerned, Homan, Hewitt and Linder (1994), Hewitt and Homan (2004), Allan, McGhee and van Krieken (2005) and Dempster and Reddy (2007) are among the few researchers to have addressed this question explicitly, although as we shall see below, there is disagreement as to whether this serves any useful purpose.

As illustrated in section 2.3.1.1 above, application of traditional readability formulas has required passages of at least 100 words. However, MCQs are usually considerably shorter than 100 words. According to Chall and Dale (1995:7-8), the Dale-Chall readability formula can be used for items shorter than 100 words (including test items) simply by prorating up to 100 words, e.g. an MCQ of 60 words and 3 sentences converts to 5 sentences per 100 words and an MCQ of 60 words with 5 unfamiliar words converts to 8 unfamiliar words per 100 words. These can then be converted to readability scores using the tables in the normal way. Homan, Hewitt and Linder (1994:350) disagree, arguing that traditional readability formulas are not suitable for measuring the readability of individual MCQs because they are intended for multiple 100-word samples. Allan, McGhee and van Krieken (2005:6) note that readability formula results will be significantly skewed when applied to short texts and that the grade level score would not be valid. Even if Allan, McGhee and van Krieken (2005:6) are correct, however, I would argue that Dale-Chall readability scores or any other readability scores would still provide a way of comparing the readability of MCQ items *with each other*.

To address this perceived problem, Homan, Hewitt and Linder (1994) developed and validated the Homan-Hewitt (HH) readability formula, which is specifically designed for single MCQ items. Their underlying assumption is that differences in readability level will affect item difficulty. This was supported by empirical research which showed class mean performance scores dropping progressively as readability levels of MCQs increased (Homan, Hewitt & Linder 1994:355). The Homan-Hewitt formula predictor variables are

(a) the number of unfamiliar words (WUNF). Like the Dale-Chall (1995) readability formula, the HH formula uses the *Living Word Vocabulary* (Dale & O'Rourke 1981) which gives empirically measured familiarity scores for American students from grade 4 to college level. The formula requires at least 80% of fourth graders to know the word in order for it to qualify as familiar. HH classifies unfamiliar words as words which have more than two letters different from the headword listed in the *Living Word Vocabulary*, e.g. if *follow* was on the list, *following* would be considered unfamiliar.

(b) the number of long words (words with 7 letters or more) (WLON).

(c) sentence complexity (WNUM) based on average number of words per clause (or more precisely per 'minimally terminable grammatical unit' as described in Hunt (1965) cited in Homan, Hewitt & Linder 1994:355).

The HH prediction equation is Y = 1.76 + (.15 x WNUM) + (.69 x WUNF) – (.51 x WLON).

Using the HH readability formula, Hewitt and Homan (2004) examined the correlations between the item readability and item difficulty of 3^{rd}-5^{th} grade social science MCQs. Their findings showed a positive correlation – on average, the higher the readability level, the more students got that item wrong. However, some items with high readability levels were easy, while some items with low readability levels were difficult. Hewitt and Homan (2004) therefore acknowledge that readability is clearly not the only factor that affects item difficulty.

Dempster and Reddy (2007) attempted to use the Homan-Hewitt readability formula in the South African context to investigate performance on 73 science MCQs by learners from African schools as compared to racially mixed schools. To address the problem of the culturally inappropriate American word familiarity lists, they used a local school dictionary instead of the extensively empirically validated *Living Word Vocabulary*. Having changed one of the variables they were not able to apply the Homan-Hewitt formula and come to a single readability score, but rather just listed all three factors (WNUM, WUNF and WLON) for each MCQ. This loses the predictive power of a readability formula which weights the various factors in order to come to a single score.

Dempster and Reddy (2007:914) found that the average difference in MCQ scores between learners at African schools (almost all L2 speakers of English) and racially mixed schools (both L1 and L2 speakers) was around 10%. They characterised 13 questions as 'difficult to comprehend' in that there was a difference of over 22% between the two groups. These questions were further analysed for 'difficult' constructions such as passives, logical connectives, qualifiers including adverbs, adjectives, prepositional phrases and embedded clauses, nominalisations, words that have dual meanings and questions phrased in the negative. Dempster and Reddy (2007:917) found that of the difficult items for L2 speakers, seven contained passives, three used *when* as a logical connective, five contained qualifiers like prepositions and embedded clauses, two contained nominalisations and two contained words with dual meanings or specific terminology. Of the five easy questions (with above-average number of learners answering correctly and only a small difference between African and racially mixed schools), there was one passive, no logical connectives, no nominalisations, one qualifying phrase and two potential vocabulary problems.

Though they found the sentence complexity measures more useful than vocabulary load in predicting learner response patterns, overall readability measures were not found to be reliable predictors of the percentage of South African learners choosing the correct answer (Dempster & Reddy 2007:919-920). Dempster and Reddy (2007:920) conclude that problems with the readability of items overlie a lack of knowledge, skills and reasoning ability in science and that the interactions between readability of items, language proficiency of the learners, and their achievement scores is not straightforward. It is for this reason that I intend in my study to attempt to rate question readability as well as cognitive demand of each MCQ so that an attempt can be made to disentangle the effects of these two aspects of difficulty.

As far as the practicability of using readability measures for MCQs is concerned, Allan, McGhee and van Krieken (2005:11) note that while many of the language boards (e.g. New Zealand, Australia and Hong Kong) screen their papers to ensure that questions can be read without difficulty, none of them use readability formulas to check the reading level of questions. They echo Homan and Hewitt's (2004) point that readability is not the only factor that affects item difficulty: 'Success in predicting difficulty of questions using a readability formula does not, of

itself, imply that this difficulty is caused directly and only by the increased level of readability' (Allan, McGhee & van Krieken 2005:7). They conclude that there are significant elements in the theoretical framework which make readability formulas inappropriate for MCQs. These include questions that are too short to obtain reliable estimates, graphs and diagrams in questions, subject-specific terminology which candidates would be expected to know and yet which would have a significant impact on overall readability score whatever formula is used, and the fact that word familiarity lists and reading grade levels are not easily transferable to other contexts. Hewitt and Homan (2004:2) disagree, arguing that the readability level of tests is 'the forgotten validity variable' and that individual item readability should be measured for standardised tests, if not for all items, then at least for items which few students answer correctly.

2.5 Post-test statistical analysis of MCQs

While calculating readability scores for individual MCQs is probably not necessary or practical, another possible way of improving the validity and fairness of an MCQ test is to scrutinise the preliminary test statistics after the test has been administered to identify items that need to be discarded or revised. One of the benefits of MCQ assessment is that answer sheets can be marked by computer and automatic reports can be generated and scrutinised before the results are finalised. Psychometrics is an active and specialised field of research that analyses the response patterns of MCQs in one of two different paradigms, Classical Test Theory (CTT) and Item Response Theory (IRT), or sometimes in a theory-free way. Both paradigms are used to evaluate the reliability, validity and difficulty of MCQ tests and items. This section reviews some of the literature within test theory, focusing on studies which look at issues of difficulty and fairness for different groups of students.

Classical test theory is based on the work of Spearman (1904), Gulliksen (1950) and Kuder and Richardson (1937) (cited in Mislevy 1993), and provides a measure of the difficulty and discrimination of each MCQ item and the frequency with which each option is selected, as well as the notion of reliability – the accuracy with which a test ranks a group of examinees.

Item Response Theory (IRT) is a more recent test theory, pioneered by research by Lord (1952) and Rasch (1960) (cited in Mislevy 1993:23). Item Response Theory became more widely used

in the 1970s and 1980s as computers became able to handle the large numbers of computations required (Gleason 2008:9). IRT differs from classical test theory in that the item response, rather than the test score, is the fundamental unit of observation. Mislevy (1993:23-4) explains that IRT concerns examinees' overall proficiency in a domain of tasks, providing a mathematical model for the probability that a given person will respond correctly to a given item, a function of that person's proficiency parameter and one or more parameters for the item. The various IRT models can measure test reliability, item difficulty and item discrimination (see Richichi 1996). They also make use of graphs known as item characteristic curves, which illustrate the probability that someone with a certain ability would answer that item correctly. In the case of multiple-choice questions, there would be separate curves indicated for the key and for each of the distractors. Like frequency of response tables, these curves enable test users to identify items that are not functioning as intended, for example where a particular distractor is chosen more frequently as student ability levels increase. Gleason (2008:15) suggests that Item Reponse Theory can assist test users to improve tests by identifying items that provide little information about testees and that could be removed to improve test validity while simultaneously shortening the test. IRT also has the potential to create computer-adaptive tests by identifying items that could provide information for a particular examinee's estimated ability level.

Debate is ongoing between adherents of the two theories, but both are still in use (Hambleton & Jones 1993). Classical test theory measures have high face validity and continue to be widely used by teachers, test developers and in empirical research (see for example Varughese & Glencross 1997, Ghorpade & Lackritz 1998, Sireci, Wiley & Keller 1998, Dempster & Reddy 2007). According to Burton (2005:71), IRT in the analysis of academic test items is sophisticated and modern but may not give better results than classical test theory. Haladyna (1994:19) also makes the point that IRT, being more statistically sophisticated, tends not to be as well understood by test practitioners. Neither approach takes account of the content of the items or of the decision-making processes involved in answering them, issues which are of increasing interest to researchers (Mislevy 1993:22).

There are a number of statistical formulas for quantitatively estimating the reliability of an MCQ test after it has been administered. For example, reliability coefficients can be used to correlate

two halves of the test score for the same group of individuals (e.g. the odd items versus the even items). Alternatively the Kuder-Richardson formula (KR-20) calculates a reliability coefficient for the test as a whole based on the number of test items, the proportion of correct answers to an item, the proportion of incorrect responses, and the variance (Bodner 1980:90). KR-20 values can range from 0,00 to 1,00 but should preferably be above 0,5 (or ideally 0,8) if there are more than 50 items. Higher reliability coefficients will be found for longer tests and for groups that vary more in ability (Bodner 1980, Haladyna 2004).

After the test has been administered, item analysis allows judgments to be made about the goodness of individual questions. Haladyna (1994:18) explains that the purpose of item analysis is to improve test items. Results should therefore be used to clean up the test and drop items that fail to perform as intended. Two CTT parameters that are useful in analysing the quality of an individual test item are the proportion of the students who choose a particular answer to the question (facility) and the correlation between the probability of a student choosing this answer and the student's total score on the examination (discrimination).

2.5.1.1 Facility (p-value)

The percentage of students who get an item correct is known as the facility (or p-value) of the item (calculated by R/N where R is the number of correct answers to that item and N is the number of candidates). P-values range from 0 to 1. For example a p-value of 0,573 indicates that 57,3% of the students answered the item correctly. For ease of reference, this measure will also be referred to as ‘difficulty’ in this thesis, although strictly speaking item difficulty refers to the percentage of *wrong* answers (calculated by 1- R/N where R is the number of correct answers to that item and N is the number of candidates) (Brown, Bull & Pendlebury 1997:94). P-values are the traditional measure of how difficult an MCQ item is and continue to be used in empirical research (e.g. Sireci, Wiley & Keller 1998, Dempster & Reddy 2007, Hewitt & Homan 2004).

Good assessment needs to include a range of questions of varying difficulty, which should manifest itself in a range of p-values. There is some variation in the recommended range of p-values, with Kehoe (1995b) suggesting that ‘(o)n a good test, most items will be answered correctly by 30% to 80% of the examinees’ and Carneson, Delpierre and Masters (n.d.) advising

that p-values should preferably be between 30% and 70%. Bodner (1980:189) suggests that questions that are answered correctly by more than 85% or less than 25% of the students are of questionable validity. Despite the slight variation in the recommended range of p-values, there is broad consensus that very easy and very difficult questions are not desirable in a test. This probably relates to the discrimination of very easy and very difficult questions as discussed below.

2.5.1.2 Discrimination

Discrimination is an important measure that tells an assessor how well a multiple-choice item differentiates between better and weaker candidates. One way to measure discrimination is (H-L)/N where H = the number of correct responses to this item by the top-third of test scorers (or top quarter, etc) and L = the number of correct responses by the bottom-third of test scorers (Brown, Bull & Pendlebury 1997:95, Benvenuti 2010). Used in this way, discrimination scores range from -1 to +1. Items with negative or near-zero discrimination should be discarded because this means that the weak students do just as well as the better students or, in the case of negative discrimination, even better. This can happen, for example, when ‘a badly worded item is taken at face-value by less-able examinees but found confusing by more able persons’ (Osterlind 1998:53).

Discrimination of an item can also be measured by the correlation coefficient reflecting the tendency of the students selecting the right answer to have high scores on the test as a whole (Kehoe 1995b). As Bodner explains:

> In theory, the student who answers a given question correctly should have a tendency to perform better on the total examination than a student who answers the same question incorrectly. We therefore expect a positive correlation between the probability of a student getting a question right and the student's score on the exam. When the correlation coefficient for a correct answer is negative, something is drastically wrong with the question. Either the wrong answer has been entered into the grading key, or the question is grossly misleading. Conversely, we should expect a negative correlation between the probability of selecting a wrong answer and the total score on the exam.
>
> (Bodner 1980:189-90)

The p-value or proportion of students answering an item correctly also affects its discrimination power, as very easy items that almost everyone gets right will obviously not discriminate much

between better and weaker candidates. Similarly, difficult items that hardly anyone gets right will have low discrimination.

It can be shown statistically and empirically that test score reliability depends upon item discrimination (Haladyna 1994:146). To improve discrimination, items that correlate less than 0,125 (or ideally 0,15) with the total test score should be restructured or discarded. Discarding items with very low or negative discrimination before the final run of a test is completely justified and will improve overall reliability of the student rankings (Kehoe 1995b).

It is also useful to look in detail at the percentage of students who choose each of the options for a particular multiple-choice item. Frequency of response tables can alert the assessor to problematic questions where high proportions of students choose one or more of the distractors rather than the correct response. Frequency of response tables enable the assessor to see which distractors are performing as intended and to identify non-functioning distractors (chosen by hardly anyone) which can then be replaced or eliminated (Haladyna 1994:154).

2.5.1.3 Statistical measures of test fairness

Of increasing interest within both theoretical paradigms is the issue of test fairness. Light can be shed on this issue by analysing the relative performance of subgroups of examinees on individual test items (e.g. Angoff 1993, Dorans & Holland 1993, O' Neill & McPeek 1993, Gierl 2005). Studies of what was known as 'item bias' began in the 1960s in an attempt to identify and remove test items that were biased against minority students, for example by containing culturally unfamiliar content. Item bias occurs when a test score is less valid for one group of test-takers than for another group, and therefore contaminates test score interpretations and uses. In a 1980 Illinois court case (Golden Rule insurance company versus Mathias), for example, a test was claimed to be biased against candidates from ethnic minorities. In the court's deliberations, a black-white difference in item difficulty (p-value) that was greater than the observed test score difference for the two racial groups was taken as evidence of item bias (Haladyna 1994:162). This basic methodology will be used in my study to contrast the performance of L1 and L2 students. Angoff (1993:14) points out that a raw difference in subgroup scores is insufficient for claiming item bias as the subgroups may differ with regard to

the nature and quality of prior education. This is certainly the case in the South African context, where second-language English speakers (except some Afrikaans speakers) typically experienced disadvantaged schooling under apartheid and post-apartheid education systems. The score differences between the two groups on a question therefore need to be compared with the average score differences between these groups on the test as a whole.

Nowadays the more neutral term ‘differential item functioning’ (or DIF) is preferred over the term ‘item bias’, because ‘bias’ has a specific technical meaning for statisticians and is apt to be misconstrued by the general public (Haladyna 1994:162). However, item bias is not strictly equivalent to DIF as DIF has the specific sense that an item displays different statistical properties for different groups after controlling for differences in the abilities of the groups (Angoff 1993).

There are several statistical methods and software programs for calculating DIF (for example, Mantel & Haenszel 1959, Lord & Novick 1968 cited in Angoff 1993). Many of these involve ranking the responses of one group of students (e.g. L1 English speakers) and matching these against the responses of another ranked group of students (e.g. L2 English speakers) with similar scores. The DIF methods of choice, such as the Mantel-Haenszel method, emphasise the importance of comparing comparable students from the two groups (Dorans & Holland 1993:37). Most methods use equal ability as measured by total test score as a measure of comparability.

DIF studies typically include a statistical analysis followed by a substantive analysis, where various expert reviewers attempt to identify the reasons for the differential performance or to identify a common deficiency among the items displaying DIF (Gierl 2005:4). While considerable progress has been made in the development of statistical methods for identifying DIF, there has been little progress in identifying the *causes* of DIF. Angoff (1993:19) points out that DIF results are sometimes difficult to understand, as seemingly reasonable items sometimes have large DIF scores. Conversely, even if an item does display DIF, the item may still be fair and useful, for example, girls may simply do worse than boys in geometry. Angoff (1993:17) explains that ‘[i]n and of itself, a high DIF value does not indicate an unfair question, just an intergroup difference between matched examinees’.

This section has reviewed some of the statistical techniques for evaluating the difficulty, discrimination and fairness of MCQs. However, since issues of difficulty and fairness cannot be analysed using statistics alone, Section 2.6 turns its focus to cognitive aspects of question difficulty, reviewing the literature on the issue of which cognitive abilities can be assessed with MCQs.

2.6 MCQs and levels of cognitive difficulty

Discussion of the difficulty of MCQs would be incomplete without an understanding of the various levels of cognitive difficulty that an item can probe. One of the aims of the study is to separate intended cognitive difficulty from the kinds of unintended linguistic difficulty that arises from ignoring the guidelines discussed in section 2.4.1 above.

The classic taxonomy of the cognitive demand of particular assessment questions is Bloom (1956). The *Taxonomy of educational objectives. Book 1. Cognitive Domain* was created collaboratively by a group of 34 American Psychological Association university examiners attending annual meetings between 1949 and 1953. Their aim was to facilitate communication among educators by drawing up a hierarchical classification of the range of possible educational goals or outcomes in the cognitive area, including activities such as remembering and recalling knowledge, thinking, problem solving and creating (1956:2). A thousand pre-publication copies were distributed widely for additional input, so the taxonomy is very much a tested, group product that was and continues to be used as a well-known standard tool among educators.

Written at the height of behaviourism, Bloom's taxonomy clearly relates to the describable intended behaviour of individual students. The taxonomy includes six major categories (1 knowledge, 2 comprehension, 3 application, 4 analysis, 5 synthesis, 6 evaluation) and also detailed subcategories. The condensed version of Bloom's taxonomy of educational objectives (Bloom 1956:201-207) includes the following categories:

1 Knowledge

1.10 knowledge of specifics

1.11 knowledge of terminology

1.12 knowledge of specific facts

1.20 knowledge of ways and means of dealing with specifics 1.21 knowledge of conventions

1.22 knowledge of trends and sequences

1.23 knowledge of classifications and categories

1.24 knowledge of criteria

1.25 knowledge of methodology

1.30 knowledge of the universals and abstractions in a field 1.31 knowledge of principles and generalizations

1.32 knowledge of theories and structures

2 Comprehension

2.10 translation

2.20 interpretation

2.30 extrapolation

3 Application

4 Analysis

4.10 Analysis of elements

4.20 Analyses of relationships

4.30 Analysis of organizational principles

5 Synthesis

5.10 Production of a unique communication

5.20 Production of a plan, or proposed set of operations

5.30 Derivation of a set of abstract relations

6 Evaluation

6.10 Judgments in terms of internal evidence

6.20 Judgments in terms of external criteria

One of the predictions made by Bloom is that the six categories are hierarchical and that higher-order cognitive skills (such as application or analysis) logically include the lower-order skills (such as knowledge and comprehension). Comprehension (level 2) would not be possible, for example, without knowledge of the relevant terms, concepts and facts (level 1). This also implies that questions at lower levels on the hierarchy should be answered correctly more frequently than questions at higher cognitive levels on the hierarchy (Bloom 1956:18).

Extensive research efforts in the 1960s and 1970s were invested in attempting to validate Bloom's taxonomy by testing the predictions made by the model. For example, Seddon (1978) reviews the empirical research relating to the hierarchical ordering of the six categories, which indicated that though knowledge, comprehension, application and analysis are generally in the theoretically expected order, synthesis and evaluation are often not (Seddon 1978:310). Applying the taxonomy to MCQs (and written items) in an Educational Psychology course, Hancock (1994) observed that the test items he investigated did not necessarily follow a general trend of increasing difficulty as Bloom's level of cognitive complexity increased. He points out that the notion of cognitive complexity is often confused with item difficulty, whereas the two are in fact distinct attributes that are often not even correlated (Hancock 1994). My own view is that cognitive complexity and item difficulty are not likely to correlate strongly because of the additional influence of linguistic difficulty such as ambiguity or syntactic complexity. This issue will be explored further by attempting to compare the readability of items at the same Bloom level.

As pointed out already, a well-constructed assessment needs to vary the cognitive difficulty from simple recall questions to more subtle reasoning and application. Stiggins (2005) believes that MCQs work well for assessing knowledge and content outcomes, and for some types of reasoning, including analytical reasoning, classifying and drawing conclusions. Complex performance (like playing the cello), novel interpretations and critical thinking are unlikely to be assessable using multiple-choice (Martinez 1999:210). In this respect, there will always be some degree of construct underrepresentation and therefore of test validity when multiple-choice is the only assessment method in a course (Martinez 1999:210). Multiple-choice assessment should therefore form part of an overall assessment strategy that includes a range of assessment types.

Martinez (1999) offers an interesting exploration of the relationship between cognition and test item format. He points out that while multiple-choice items are perfectly capable of eliciting complex cognitions, including understanding, prediction, evaluation and problem solving, in practice they tend to be used more often for testing lower-order thinking (Martinez 1999:208). Paxton (2000)'s research on MCQs in first-year Economics at the University of Cape Town echoed this finding, noting that two-thirds of the questions were of a low-level definitional kind as opposed to application. Martinez (1999) cites research by Bowman and Peng (1972), who classified 800 Psychology examination questions into four categories based on their cognitive demand: memory, comprehension, analysis and evaluation. The first two categories accounted for over 80% of the items (Martinez 1999:209). Benvenuti (2010) found that 81% of the questions in first-year Informations Systems courses at several South African universities were recall questions, 12% were comprehension questions and only 6% required application or analysis. This lopsided distribution of cognitive requirements seems fairly common and may not be fully appreciated by test designers (Martinez 1999:209). Haladyna (1994:8) states this point even more strongly: 'Ample evidence exists to suggest a crisis in the measurement of higher level outcomes for achievement tests. Teachers at all levels seldom adequately define, teach and test higher level outcomes.'

There is some disagreement as to whether MCQs can test the most complex of Bloom's levels, namely synthesis and evaluation. Govender (2004:96) believes that MCQs are better for testing the first four levels of Bloom's taxonomy, and to a limited extent synthesis and evaluation. One caveat pointed out by Hancock (1994) is the tendency in the literature to rate items on the basis merely of the *name* of the Bloom category, as opposed to the detailed description provided in the original work. In my view, MCQs cannot test synthesis and evaluation in Bloom's sense, where synthesis is defined as the 'putting together of elements to form a new whole' and evaluation as 'quantitative and qualitative judgments about the value of material and methods for given purposes, using given or own criteria'. I would argue that choosing amongst lecturer-selected options does not allow either synthesis or evaluation. This view is supported by the fact that most empirical research has ended up classifying MCQ items into a maximum of four cognitive categories. For example Bowman and Peng (1972) used four levels – memory, comprehension, analysis and evaluation, while Paxton (2000:111) classified questions as either definition or

application. Homan, Hewitt and Linder (1994:353) used only knowledge and comprehension questions in their study of the relationship between mean scores and readability levels but did not distinguish between these two categories.

Fellenz (2004:709) comments that '… the literature suggest that the cognitive levels can be assessed with relative ease and high inter-rater reliability by experienced teachers'. For example, Hancock (1994) obtained 86% and 91% agreement between two independent raters, familiar with the course, in classifying MCQs and written questions in terms of the first four of Bloom's levels. He notes that perfect agreement cannot be expected since 'different students can meet the same objective in different ways depending on their prior learning experiences, making multiple categories potentially valid for a given objective or test item'.

A revision of Bloom's taxonomy was published by Anderson and Krathwohl (2001) and an overview given in Krathwohl (2002). While there are still six categories in the revised taxonomy, the names of the categories were changed to 'remember, understand, apply, analyze, evaluate and create', and the order of the two highest categories was interchanged. Another important change in the taxonomy was the two-dimensional rather than single-dimensional model, in which the cognitive process (remember, understand, apply, analyse, evaluate, create) intersects with a knowledge dimension detailing whether the knowledge is factual (knowledge of terminology and elements), conceptual (knowledge of interrelationships between elements), procedural (how and when to use skills, techniques and methods) or metacognitive (awareness and knowledge of one's own cognition). In this model, the knowledge dimension (factual/conceptual/ procedural/metacognitive) forms the vertical axis and the cognitive process (remember/understand/apply/analyse/evaluate/create) forms the horizontal axis (Krathwohl 2002:214). Individual assessment tasks can then be classified in the intersection of the relevant columns/rows, as done for example by Benvenuti (2010).

Haladyna (1994) points out that cognitive taxonomies like Bloom are mostly prescriptive (non-theoretical descriptions) and empirical research neither supports nor refutes them. He believes that we need to move in the direction of theoretically based taxonomies for defining and measuring higher-level thinking like Royer, Cisero and Carlo (1993) which, although still in

their infancy, can tell us more about cause-effect relationships, and principles governing how various types of higher-level thinking are developed (Haladyna 1994:88).

2.7 Conclusion

A wide and interdisciplinary selection of literature has been presented in this chapter that has a bearing on the notion of MCQ difficulty and fairness for both first and second language examinees. Chapter 3 will focus on the research methodology used in the present study to ascertain to what extent the language of multiple-choice questions is indeed a barrier for L2 students in comparison to L1 students.

Chapter 3
Research design and research methods

3.1 Introduction

This chapter describes the research aims (section 3.2), the research design (section 3.3), and the theoretical framework that informs the research (section 3.4). This is followed in section 3.5 by a description of the data used in the study and in section 3.6 by a justification and detailed explanation of the methods and procedures selected for the study.

3.2 The research aims

The aims of the study were described in section 1.3 and include descriptive-analytical, theoretical-methodological, and applied aims. The three research questions listed again below are focused primarily on addressing the descriptive-analytical aims, although the findings will also have both theoretical-methodological and applied implications:

1. Which kinds of multiple-choice questions (MCQs) are 'difficult'?
2. What kinds of multiple-choice items present particular problems for L2 speakers?
3. What contribution do linguistic factors make to these difficulties?

3.3 The research design

In order to answer the above questions, my research design will be both quantitative – aiming to answer questions about relationships between measured variables with the purpose of explaining, predicting and controlling phenomena, and qualitative – aiming to answer questions about the complex nature of phenomena with the purpose of describing and understanding phenomena from the participants' point of view (Leedy & Ormrod 2001:101). Several researchers have pointed out the need for multiple perspectives on the difficulty and validity of MCQ items – with triangulation of statistical and interview data enabling more sensitive interpretation (e.g. Norris 1990, Haladyna 1994, Sireci, Wiley & Keller 1998, Paxton 2000, Rupp, Garcia & Jamieson

2001). Confirmation of the need for both qualitative verbal reports and more quantitative methodology for validating MCQs is also provided by Norris (1990), who concludes his qualitative investigation as follows:

> The verbal reports of thinking collected for this study contained a wealth of information useful for rating the quality of subjects' thinking and for diagnosing specific problems with items, such as the presence of misleading expressions, implicit clues, unfamiliar vocabulary, and alternative justifiable answers to the answer keyed as correct. Given the results of this study, it is reasonable to trust this diagnostic information as an accurate representation of problems that would occur with the items taken in a paper-and-pencil format. Moreover the verbal reports provide more direct information on the exact nature of problems of these sorts than that provided by traditional item analysis statistics, such as difficulty levels and discrimination indices (Haney & Scott 1987). Thus there is the possibility of complementary roles for traditional item analyses and verbal reports of thinking.
>
> (Norris 1990:55)

By investigating an MCQ text and students' MCQ performance from several angles, using statistical item analysis, linguistic and cognitive analysis of individual items (see section 3.6.1 below) and students' verbal reports (see section 3.6.2 below), I hope to obtain rich data on the reasons (linguistic or otherwise) why particular items are difficult to understand and to answer, for both L1 and L2 speakers of English. This should enable me to identify trends and make recommendations that will improve both comprehensibility and validity of multiple-choice test items.

3.3.1 Quantitative aspects of the research design

A descriptive quantitative research design (Leedy & Ormrod 2001:114) has been chosen for this part of the study in order to identify the characteristics of MCQs that students find difficult and to explore possible interactions and correlations among phenomena, in this case the home language of the students, the readability and cognitive levels of the questions, and whether or not the questions adhere to various language-related item-writing guidelines. The dependent variable in this case is MCQ difficulty for L1 and L2 students, as measured by a relatively low percentage of students answering the item correctly. The explanatory (independent) variables that could potentially explain the dependent variable are the readability scores of the questions and the degree of adherence to a set of language-related MCQ guidelines. The controlled variable is the

cognitive complexity of the question in terms of Bloom's taxonomy, as this could potentially interfere with an investigation of the effect of language on MCQ difficulty.

The quantitative research included secondary data analysis of existing statistical examination data in order to answer descriptive or causal research questions (Mouton 2001:164) and readability and linguistic analysis of the text of the examination questions.

In this study, this data involved an entire class of students writing MCQ examinations in first-year Linguistics in the normal course of their studies. The primary data provides information on the difficulty and discrimination of each MCQ item, to which I added a measure that compares item difficulty for L1 and L2 students. This will enable me to identify MCQs that are difficult in that they are answered incorrectly by many students or cause particular problems for L2 English students.

The secondary analysis of statistical data will be coupled with a quantitative analysis of the texts of two MCQ examinations in first-year Linguistics. Aspects of the language of the questions will be quantified, namely their readability and density of academic words and correlations will be sought between question difficulty and these aspects of the language of the questions. An attempt will be made to elucidate the effect of eight item-writing guidelines on the difficulty of items for L1 and L2 students. Both these methods were non-reactive, involving no interaction between researcher and students.

3.3.2 Qualitative aspects of the research design

Qualitative approaches focus on phenomena that occur in natural settings, portraying as many dimensions of the issue as possible and acknowledging the multiple perspectives held by different individuals (Leedy & Ormrod 2001:147). The purpose of the qualitative investigation in this case was both descriptive and interpretive, attempting to provide insights into the problems experienced by first-year students in reading, understanding and answering MCQs. Scholfield (1994:32) notes that comprehension cannot be quantified in a non-reactive way, but requires obvious questioning or testing of readers. The primary data in this portion of the study was gathered in a reactive way, involving quasi-naturalistic interviews which were intended to

explore and describe why particular MCQs are experienced by students as 'difficult'. These interviews were quasi-naturalistic in that they mirrored the student's behaviour in an examination situation, reading each MCQ, going through a process of reasoning and then selecting the correct answer. No prior hypothesis was tested, but there was a general expectation that difficult questions (those that students left out, got wrong or had difficulty answering) might be associated with linguistic features such as negation, ambiguity, academic words and long or complex sentences.

The case study method was selected, using semi-structured interviews and a think-aloud protocol. The think-aloud protocol is a technique where participants are asked to vocalise their thoughts, feelings and opinions as they perform a task such as reading, writing, translating, using a product or, as in this case, writing a test. The resulting verbalisations could include paraphrases, elaborations, explanations, inferences, and/or misrepresentations of the text itself (Taylor & Taylor 1990:77-78) and are useful in understanding mistakes that are made and getting ideas for what the causes might be and how the text (or product) could be improved to avoid those problems.

3.3.3 Internal and external validity

In conclusion, this study brings together several different approaches to MCQ difficulty in its investigation of MCQs in a first-year university-level content subject, namely Linguistics. To improve internal validity, and to do justice to the multiple aspects of difficulty, this study attempts to triangulate both quantitative and qualitative data and provide a detailed description of the kinds of experiences students face in interpreting and answering MCQs. Replicating the procedures on two separate examination papers from 2006 and 2007 will also aid validity and indicate whether the findings of both years corroborate each other or are dissimilar.

As far as external validity is concerned, the findings relating to MCQs in Linguistics may be generalisable only to other content subjects in the humanities. In terms of their language backgrounds the students in the quantitative investigation are representative of South African students in general. All were multilingual and between them they spoke all 11 official languages as well as a few foreign languages and displayed a range of English proficiency. The case study

method has the inherent limitation that it can investigate only a small number of cases and may yield results that may not be widely generalisable to other contexts. However the level of detail possible in a think-aloud analysis of reasons may provide evidence of types of difficulties encountered by students with the language of MCQs that may well be useful and generalisable in other situations where MCQs are used to test L2 students.

3.4 The theoretical framework

The theoretical framework in which this study is situated is diverse due to the number of different disciplines that share an interest in text difficulty. These include test theory, readability studies, empirical investigations of guidelines for MCQ item writers, studies that focus on the cognitive complexity of items and investigations of how students actually answer MCQs. In fact, in an attempt to look at real problems relating to language in education in a new way, the emerging discipline of Educational Linguistics encourages an eclectic synthesis of methodologies (Hult 2008:13).

Much of the discussion, for example on the MCQ writing guidelines, is atheoretical, but the theories that are alluded to include classical test theory, interactionist theories of reading, cognitive approaches to readability and behaviourist models of cognitive difficulty. These were introduced in Chapter 2 but are briefly revisited here.

Classical Test Theory (CTT) is based on the work of Spearman (1904), Gulliksen (1950) and Kuder and Richardson (1937) (cited in Mislevy 2003:22), and provides a statistical measure of the difficulty and discrimination of each MCQ item and the frequency with which each option is selected, as well as the notion of reliability the accuracy with which a test ranks a group of examinees. These CTT measures are used in the study to quantify aspects of the difficulty and reliability of two MCQ examinations.

Reading comprehension research since the 1970s has shown that reading comprehension involves a complex, multifaceted collection of decoding and comprehension (Just & Carpenter 1987, Pretorius 2000). Interactive models of reading comprehension recognise the interaction between decoding and comprehension processes and suggest that decoding and comprehension

share a limited resource pool, with overall comprehension suffering if working memory is overloaded or if decoding processes take up too much attention (Perfetti 1988). Problems in constructing a coherent representation of the meaning of a text can arise either from text-based variables (such as how readable, ambiguous or considerate the text is) or reader-based variables (such as a reader who lacks the required decoding skills or background knowledge). The current study assumes an interactive approach to MCQ reading comprehension but focuses on text-based variables that influence reading comprehension.

While readability formulas tend to be atheoretical, more recent theories of readability draw on ideas from linguistics, applied linguistics and cognitive psychology and stress the importance of less measurable aspects of texts such as cohesion, idea density and reader-specific factors such as motivation, interest and prior knowledge (e.g. Kintsch & Miller 1981, Huckin 1983, Armbruster 1984, Kemper 1988, Davison and Green 1988). These aspects of readability are explored indirectly in the MCQ think-aloud protocol.

Written at the height of behaviourism, Bloom's taxonomy (Bloom 1956) (see section 2.6) describes the intended behaviour of students by identifying six hierarchically-ordered categories of educational objectives. This hierarchy is used to evaluate the cognitive difficulty of the MCQs in the study.

3.5 Research data

Section 3.5 below offers a description of the research data in this study, including the course, the MCQs and a profile of the students.

3.5.1 The course

The course I chose to base this research on is a first-year Linguistics course at the University of South Africa (Unisa), the country's largest tertiary institution, which teaches by means of distance education. The course, LIN103Y, *Multilingualism: The role of language in South Africa*, was an introduction to sociolinguistics and language acquisition. The course offers blended learning, with printed course material that is also available electronically, one optional face-to-face session, an asynchronous discussion forum and additional electronic resources. The

course runs for a semester (half-year) and the periods under investigation in this study were July to October 2006 and July to October 2007. The course had won an Excellence in Tuition award in 2006 and the quality of the study material and assessment was therefore assumed to be good. This course was selected because I am the course co-ordinator and have access to student records and to students themselves and a professional interest in seeing to what extent the assessment could be validated.

3.5.2 The MCQs

First-year courses tend to make the heaviest use of MCQ assessment owing to the large student numbers. The 160 MCQs investigated in the study came from two 2-hour 80-item multiple-choice examinations (counting 80% towards a student's final mark) written in October 2006 and October 2007 for LIN103Y. The examinations had a dichotomous (right or wrong) scoring key drawn up beforehand and were informal rather than standardised, with a fair to generous time limit of 1½ minutes per question. The examinations included short texts/case studies (ten texts in 2006 and nine texts in 2007) and sets of questions either based directly on the preceding text or more generally on the topic area related to the text (see Appendices A and B). KR-20 internal reliability measures for both examinations were very high at 0.912 (2006) and 0.928 (2007). As discussed in Chapter 2 (section 2.3.2.2), high reliability coefficients tend to be obtained for longer tests and when the testees vary significantly in ability. Both of these factors probably contributed to the high reliability in this case.

3.5.3 The students

Unisa students are adults doing either part-time or full-time distance education studies and come from a wide spectrum of educational backgrounds. Their ages range from 18 to retirement age and they live mostly in Southern Africa but can be based anywhere in the world.

The October 2006 multiple-choice examination in LIN103Y was written by 136 students. One-third (33%) of these indicated on the Unisa registration form that their home language was English. Nearly half (46%) of students indicated that they spoke one of South Africa's nine official African languages as a mother tongue, mostly Northern Sotho and Zulu, while 13%

spoke Afrikaans as their home language and 8% spoke another African or foreign language. The number (not percentage) of students speaking particular languages is indicated in Figure 3.1.

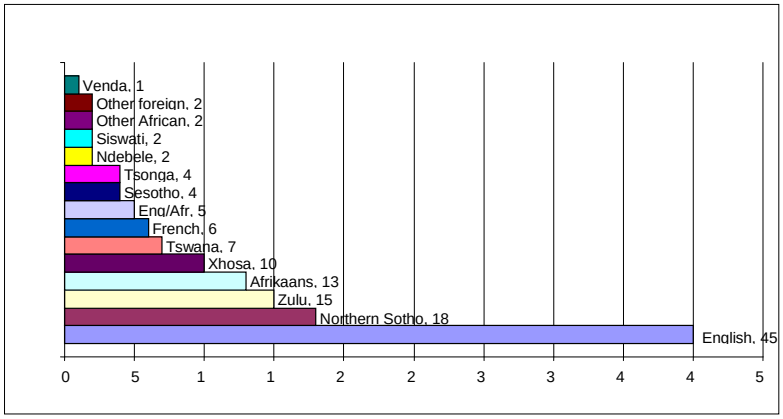

Figure 3.1 Home languages of LIN103Y students 2006

The October 2007 multiple-choice examination in LIN103Y was written by 117 students. Just over a quarter (26%) of these indicated on the Unisa registration form that their home language was English, 10% were Afrikaans, and 38% of students indicated that they spoke one of South Africa's nine official African languages as a mother tongue, mostly Siswati and Zulu. Just over a quarter (26%) spoke another African or foreign language, mostly Mauritian students indicating French as L1. The number (not percentage) of students speaking particular languages is indicated in Figure 3.2 below.

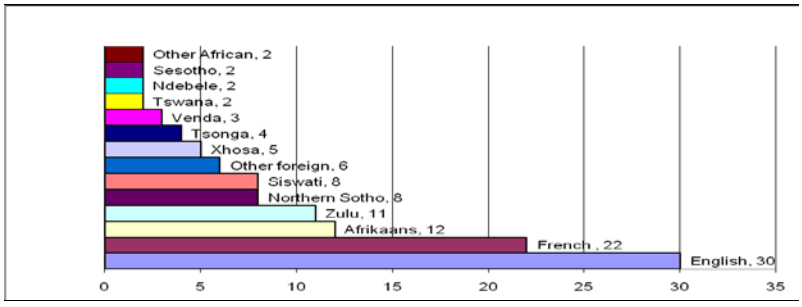

Figure 3.2 Home languages of LIN103Y students 2007

An assumption was made that the home language indicated by the student on his or her Unisa registration form was a true reflection of the student's mother tongue. However it should be noted that in the South African context, determination of mother tongue is often a complicated matter. Situations where the mother's language and father's language are both spoken in the home and children become simultaneous bilinguals are common. Obviously there is also a large range of English proficiency, depending on when and where English was learnt and how much exposure students have had to the language. Some L2 English students have the proficiency of L1 speakers, while others have very low English proficiency. On average, however, L2 English speakers are less proficient than mother-tongue speakers and the comparisons of item p-values for these two groups of students should reflect genuine difficulties with interpreting the language of the questions.

A think-aloud rerun of the LIN103Y examination was undertaken with four 2006 students and nine 2007 students. Letters were sent to all second-semester 2006 and 2007 students requesting them to participate in the MCQ research after the examination. A few students contacted me to volunteer, and bearing in mind the need for both genders and a range of different ages and home languages to be represented, I phoned several other students until I made up the final 13. The students ranged in age from their early 20s to 84, which is not unusual for a Unisa course. Of these 13 students, eight were women and five were men, which is approximately representative of the gender makeup of the whole LIN103Y student group. For convenience, all 13 students were based in the Johannesburg and Pretoria areas, meaning that rural students with possibly very low English proficiency were underrepresented. All students signed the consent form beforehand (see Appendix C) and were paid for their time if necessary.

Comment [PJ1]: Stats?

In 2006, in the first set of interviews, I decided to interview two L1 and two L2 speakers: NG and DS were L1 English speakers, CL was a Zulu speaker who was taught in English from Grade 8, and DD was a Tswana speaker who was taught in English from Grade 5. Subsequently, however, because the problems experienced by the L2 speakers were of more relevance to the study, I decided to interview mostly L2 speakers in 2007: AJ spoke English as L1 and the other seven students spoke English as one of their additional languages. The home languages of these seven students were as follows: AZ Xhosa, SD Sesotho, ZL Afrikaans, TH Sepedi, ET Tsonga,

NN Zulu and MD Lingala, a language spoken in the Democratic Republic of Congo. One student, YL (L1 Afrikaans), volunteered to participate in a think-aloud rerun of the June 2007 MCQ examination, and although this examination paper was not analysed in this study, her comments will also be referred to. The overall sample of three L1 speakers and 10 L2 speakers was therefore approximately representative of the linguistic makeup of the whole LIN103Y student group.

3.6 The choice of research methods

In the following section I give some consideration to why I have chosen particular methodologies. The discussion will refer again to some of the diverse approaches and methods discussed in Chapter 2 that have been used to research the issue of MCQ difficulty. As shown below, the study will draw on several theoretical and atheoretical traditions. The quantitative measures are discussed further in 3.6.1 below followed by a justification of the qualitative methodology of the study in 3.6.2.

3.6.1 Quantitative research methods

The quantitative measures used in the study are justified and explained further in sections 3.6.1.1 to 3.6.1.4 below, beginning with item analysis of MCQs that flout item-writing guidelines, measures of readability and vocabulary load, and classification in terms of cognitive complexity.

3.6.1.1 Item analysis

The statistical data consisted of Unisa LIN103Y MCQ results for the October 2006 and October 2007 examinations. This data is generated automatically after every MCQ assessment and provided as printouts to lecturers, and contains inter alia two traditional measures used in classical test theory for analysing MCQ items (see also section 2.5):

1. Item **facility** or p-value (percentage of students who answered correctly).
   This is the traditional measure of the difficulty of an MCQ. For example, Q1 2006 was answered correctly by 88% of students and therefore has a facility of 88%, indicating a very easy question. Items with p-values below 50% were considered to be ‘difficult’ and were analysed for linguistic causes of difficulty.

2. Item **discrimination** (item-test point-biserial correlation coefficient)
   This measures the extent to which the MCQ discriminates between better and weaker students on a scale between -1 and 1. Higher discrimination is indicative of better questions. For example, Q1 2006 had a rather low discrimination of 0.256 and didn't separate better from weaker students as almost everyone got this question right.

These two measures were chosen because the Unisa assessment system provides these (and other CTT) statistics for every MCQ assessment and because they are familiar to most educators and researchers in education measurement and continue to be widely used as indicators of item quality (e.g. Sireci, Wiley & Keller 1998).

This study attempts to look beyond simple measures of difficulty and discrimination to the comparative patterns of right answers in groups of L1 and L2 students. I therefore also made use of data from the Unisa assessment system database, SQL, listing each student's home language and actual MCQ choices for every examination question. The latter data was split into two groups on the basis of the home language indicated by the student on the Unisa registration form: mother-tongue English students (those who select English as their mother tongue on the registration form (L1)) and second-language speakers of English (those who select any other language, South African or foreign, as their mother tongue on the registration form (L2)). A third measure was then added to the two mentioned above:

3. The **difficulty differential** is the difference in p-values for L1 and L2 students on a particular item. Calculating the difficulty differential is necessary to answer the second research question: 'What kinds of MCQ items present particular problems for L2 speakers?'. For example, Q1 2006 was answered correctly by 88% of students. 93% of L1 students got this question right and 85% of L2 students got this question right, yielding a difficulty differential of 8%. This score is then compared with the average score difference between L1 and L2 students on the test as a whole: 76% - 62% = 14%. Because 8% is lower than the average difference of 14% between the two groups, this question can be viewed as not

> disadvantaging L2 students. A difficulty differential over 25% (10% higher than the overall average test score difference between L1 and L2 speakers) was taken as an indication of excessive difficulty or unfairness for L2 students and these items were subjected to further analysis to see if there were linguistic issues to which this could be attributed. The range of values, in theory, is anything between -100% and 100%. Items with a negative difficulty differential are easier for the L2 speakers than for the English L1 group, a possible but unlikely scenario. Ideal questions would have no or only a low difficulty differential, i.e. approximately the same as the overall score difference between L1 and L2 students.

In my view, the difficulty differential measure is an important addition to a holistic notion of item quality in a context where L1 English students are in the minority. It is basically a measure of item bias or Differential Item Functioning as explained in Chapter 2, contrasting the performance of English L1 students (the reference group) with L2 students (the target group) to get an indication of which questions were proving particularly difficult for L2 speakers. The performance of subgroups of examinees on an item has a bearing on the fairness of the test as a whole. Standard Unisa statistics do not provide any measure of DIF. However, as mentioned in Chapter 2, Haladyna (1994:162) mentions a legal case in the United States where a difference between the item difficulty for black pupils and white pupils that was greater than the observed test score difference between these two groups was taken as evidence of Differential Item Functioning. Although there are statistically sophisticated methods of calculating DIF, this relatively simple method was selected for the purposes of the study as DIF is only one of the sources of evidence used here to identify difficulties L2 students experience in answering MCQs. A high difficulty differential can provide an important pointer to L2 students' language-related difficulties, which can then be corroborated by student interviews and by further analysis of the items.

This same methodology was used recently by Dempster and Reddy (2007) who found that the average difference in science MCQ scores between learners at African schools (almost all L2

speakers of English) and racially mixed schools (both L1 and L2 speakers) was around 10%. A subset of questions were classified as 'difficult to comprehend' in that there was difference of over 22% between the two groups. These questions were further analysed for passives, logical connectives, qualifiers including adverbs, adjectives, prepositional phrases and embedded clauses, nominalisations, words that have dual meanings and questions phrased in the negative.

Together, the three quantitative measures listed above can be used to provide a statistical picture of item quality (Sireci, Wiley & Keller 1998). Firstly the average values will be calculated for the examinations as a whole, secondly these three measures will be calculated for each item, high difficulty items and high difficulty differential items will be analysed further, and these measures will also be used to validate eight MCQ guidelines. The item quality (average p-value, discrimination and difficulty differential) of questions that disregard the guidelines will be compared with the average values in an attempt to identify whether violation of each of these item-writing guidelines leads to more difficult MCQs for L1 and L2 students.

3.6.1.2 MCQ writing guidelines

We saw in section 2.3.1 that MCQ item design is a well-established field of inquiry in its own right. Haladyna (2004:vii) makes reference to the 'short yet rich history of efforts to improve MC item writing', dating back to the early twentieth century when MCQs were introduced. Though lacking a theoretical basis, there is a wealth of guidelines, working principles and procedures on how best to write effective multiple-choice items (e.g. Haladyna 1994:61-86, Kehoe 1995a, Brown, Bull & Pendlebury 1997, Haladyna, Downing & Rodriguez 2002, Thompson, Johnstone & Thurlow 2002, Stiggins 2005:101-102, Fairbairn & Fox 2009). These include suggestions about content, format, layout, review procedures, and the language that should be used in MCQs.

The eight item-writing guidelines discussed below were selected for empirical analysis because these seemed to me to be the guidelines that related most to the language of MCQs (as opposed to content, format, layout etc.). An indication of how adherence to each guideline will be quantified in the study is given after each guideline below:

(a) Avoid negative words such as *not* or *except* in the stem.

Stems or answer choices containing any negative words such as *no, not* or *false* or negative morphemes such as *un-, dis-* were identified.

(b) State the stem in a question format instead of an incomplete sentence format.

Question stems were full-sentence stems that ended with a question mark. Incomplete statement stems included items like Q58 in (f) below where the answer choices completed the partial sentence in the stem and cloze-type stems with a missing word somewhere in the middle of the stem.

(c) Simplify vocabulary and aim for maximum readability.

The issue of readability and vocabulary load was measured in terms of the density of AWL words (AWL words as a percentage of total words) in each MCQ and the Dale-Chall (1995) readability score as described in section 3.6.1.2 below.

(d) Try and keep options brief.

Brevity was measured in terms of words per question. Questions of over 50 words (50% more than the average of 33) were categorised as long.

(e) Avoid answer choices that are too similar.

The only study I am aware of that has attempted to measure 'answer choices that are too similar' is Rupp, Garcia and Jamieson (2001). What they termed 'lexical overlap' between the correct answer and the distractors was measured by counting how many out of the three distractors had at least one content word in common with the correct answer. In my view, a single overlapping content word is unlikely to confuse students and is insufficient for claiming that answer choices are too similar. Overlap within distractors is also just as likely to be confusing as overlap with the correct answer. In my view the best way to measure answer choices that are too similar is to rate options with at least half of their words in common as 'too similar'. This implies that an item may have two or more than two answers that are 'too similar'. The example below has four very similar answer

choices as [1] and [2] share more than half their words, and [3] and [4] share more than half their words, as do [2] and [4] and [1] and [3].

14. A child who uses the word aeroplane to refer to aeroplanes and helicopters is

[1] overextending the word *aeroplane*
[2] overextending the word *helicopter*
[3] underextending the word *aeroplane*
[4] underextending the word *helicopter*.

A and *an* are counted as the same word, but noun phrases where the only similarity is the article are not counted as similar.

(f) Try and keep options grammatically parallel.

Questions for which all the answer choices were not of the same phrasal type were classifed as being grammatically non-parallel. The issue of grammatically parallel options is interesting and I have not seen any studies on this. I believe this is an issue as it affects whether one can read the item efficiently (as four possible endings to a single sentence or four possible answers to a single question stem) or inefficiently (if one has to keep rereading the stem to recast the sentence). The following example is not grammatically parallel as the first three options are noun phrases while option [4] is a verb phrase, so that the word *are* in the stem now has to be read as an auxiliary verb rather than as the main verb:

58. The girls in the example above are probably

[1] balanced bilinguals (NP)
[2] semilinguals (NP)
[3] monolinguals (NP)
[4] using a divergent strategy to stress the differences between them. (VP)

(g) Avoid All of the above (AOTA)

Items with AOTA as the final answer choice were identified.

(h) Keep None of the above (NOTA) to a minimum.

Items with NOTA as the final answer choice were identified.

Since statistical data can shed only limited light on MCQ difficulty, these item quality measures will be fleshed out with data from other sources including readability analysis as described in section 3.6.1.3 below and student interviews as described in section 3.2.

3.6.1.3 Readability and vocabulary load

One of the most commonly used quantitative measures of the linguistic difficulty of a text is its readability score. As explained in section 2.5.1, extensive research since the 1920s has indicated that, despite its broad and complex nature, the readability of a text can be predicted fairly successfully using simple measures such as word frequency and sentence length. A readability formula enables users to predict the difficulty of a text based on a few short samples. Vocabulary load and syntactic complexity are acknowledged to be the most robust predictors of readability (Homan, Hewitt & Linder 1994:350).

The Dale-Chall readability formula, first published in 1948 and revised in 1995, was selected for use in this study because it is one of the most accurate and reliable of the readability formulas according to Klare (2000:22-23) and DuBay (2004:53) and has been extensively validated since its introduction over 60 years ago. The Dale-Chall Readability Formula is based on only two factors, word familiarity and sentence length, which are a good indication of lexical and syntactic complexity.

Sentence length in the Dale-Chall readability formula is calculated by selecting several 100-word samples and counting the number of complete sentences (ending with a full stop, question mark or exclamation mark) in a 100-word sample. Although the Dale-Chall formula is intended for use on 100-word text samples and MCQs tend to be (sometimes much) less than 50 words long, the formula can be used for items shorter than 100 words simply by counting the number of words and prorating up to 100 words (Chall & Dale 1995:7-8). For example, an item consisting of a single 33-word sentence converts to 3 completed sentences in 100 words. Following the rules of the formula, acronyms and numerals are counted as single words and hyphenated attributive adjectives like *one-word* (in *one-word stage)* are counted as two separate words (Chall & Dale 1995:8).

The issue of word familiarity in the Dale-Chall Formula is calculated by counting the number of words not on the 3000-word list known to 80% of American 4th graders (Chall & Dale 1995:16-29). An equivalent list of words familiar to South African students does not exist. However, apart from a few American English terms (e.g. *billfold, boxcar, burro* etc.), which would not occur in non-American texts anyway, the list is simple and generic enough to be applicable in other English-speaking contexts, particularly if obvious equivalents are replaced as necessary (e.g. *South African* for *American*, *braai* for *barbecue*, *aeroplane* for *airplane*, etc.). Regular morphological variants of the words on the list (those with a maximum of one of the suffixes *'s, -s, -es, -ies, –d, -ed, -ied, -ing, -r, -er, -est, -ier, -iest*) are counted as familiar. Proper names are counted as unfamiliar the first time they occur. Numerals are counted as familiar, while acronyms, fractions, decimals and percentage signs as unfamiliar. For example, Q56 2000 had 1 completed sentence and 9 words not on the familiar list:

56. The **linguistic phenomenon** illustrated above is known as

[1] **borrowing**
[2] **codeswitching**
[3] **positive transfer**
[4] **negative transfer**
[5] a learner **variety**.

The number of complete sentences and unfamiliar sentences in 100 words are then converted to American grade-equivalent reading levels using the tables in the normal way. This yields a Dale-Chall readability score of 16, the highest possible level, for the MCQ above. As pointed out in section 2.5.4, the American grade reading levels are unlikely to reflect South African reading levels, and the grade level score might not be valid when applied to short texts (Allan, McGhee & van Krieken 2005:6). Even if Allan, McGhee and van Krieken (2005:6) are correct, however, I would argue that Dale-Chall readability scores would still provide a reliable way of comparing the readability of MCQ items *with each other*.

A promising alternative readability formula, The Homan-Hewitt Readability Formula (Homan, Hewitt & Linder 1994), has the advantage of being specifically designed for single-sentence test items but was not used in this study, partly because it is very new and has not been extensively

validated. Like the Dale-Chall readability formula, the Homan-Hewitt formula is based on word familiarity for American students (according to Dale & O Rourke's 1981 *The living word vocabulary*) and the readability scores represent American grade levels.

A second problematic aspect of the Homan-Hewitt formula, in my view, is their method of classifying unfamiliar words. HH classifies words as unfamiliar if they differ by more than two letters from the headword listed in *The living word vocabulary*. So, for example, if *follow* was a familiar word, *following* would not be, unless it too was listed. This is highly arbitrary and counter-intuitive in my opinion as there is considerable evidence that we tend to know not just individual words but rather word families and have enough morphological abilities at university-level to make a fair attempt to interpret words on the basis of the known morphemes and regular suffixes (Nagy et al. 1989, Coxhead 2000). Schmitt and Zimmerman (2002:145) comment that 'knowing one member of a word family undoubtedly facilitates receptive mastery of the other members', although L2 students may not be able to *produce* all the appropriate noun, verb, adjective and adverb derivatives of a word. As Anderson and Davison (1988) point out, 'it may seem intuitively obvious that long, rare words are an important cause of text difficulty, but close analysis shows that this intuition is open to serious question, e.g. some rare derived words like *caveman* or *rustproof* are easy to understand because they are derived by regular morphological rules and therefore predictable' (1988:27). The Dale-Chall readability formula (1995) is more sensitive in this respect, as it considers as familiar words all morphological variants of listed words with the following common suffixes *(-'s, -s, -es, -ies, -d, -ed, -ied, -ing, -r, -er, -est, -ier, -iest*) (Chall & Dale 1995:16).

A further criticism of HH is their procedure for counting words per clause (or per sentence). Counting words per sentence is not a simple matter for MCQs, since they often consist of an incomplete sentence stem and four possible options that complete the sentence as in the following example from Homan, Hewitt and Linder (1994:354):

Goods and services provided to people by the government are paid for with ____.

(a) credit
(b) insurance
(c) taxes
(d) interest

Homan, Hewitt and Linder (1994:354) count this as four separate sentences, each 14 words long. Each sentence contains one clause or T-unit (Hunt 1965), so 14 words for each sentence option. Separate HH readability calculations for each sentence option yield 3,7, 3,2, 3,0 and 3,2 which translates to an average HH readability score of 3,3 for the item. I disagree with this view as I contend that the majority of students would read this as one sentence with comma-delineated options rather than returning to reread the stem after reading each option. A more realistic count would therefore see this as a 17-word single sentence: *Goods and services provided to people by the government are paid for with credit, insurance, taxes, interest.* The Dale-Chall readability formula would count MCQs of this type as single sentences, since there is only one full stop.

In light of all these considerations, the Dale-Chall readability formula was deemed the most appropriate for this study. Readability scores for each of the 160 multiple-choice items were calculated using the new Dale-Chall readability formula (1995). These should not be interpreted as reflecting South African grade levels (or even of American grade levels) but do give an indication of the relative readability of the individual items.

A further indicator of readability will be used in the current study in an effort to compare readability measures and improve reliability. Lexical density, the ratio of lexical items to grammatical items, is a useful readability measure for multiple-choice items since it doesn't rely on 100-word samples and has been shown to be a good predictor of readability, particularly for the evaluation of English passages intended for non-native readers of English (see Harrison & Bakker 1998). Lexical density is usually measured in terms of the number of lexical items per clause, but will be tailored in my case to university-level MCQs by calculating AWL density per question, i.e. the number of words on Coxhead's (2000) Academic Word List as a proportion of the total words. The academic words in the 160 MCQs in the study were identified using Haywood's (n.d.) AWL Highlighter website, which automatically identifies words from the

AWL (Coxhead 2000) in any pasted text of your choice and highlights them in bold. The MCQs in this study were analysed at level 10, and therefore included all the academic words on the AWL. For example, Q56 2006 contained one of the highest AWL densities at 0,35 as six of its 17 words appear on the AWL:

56. The linguistic **phenomenon illustrated** above is known as

[1] borrowing
[2] codeswitching
[3] **positive transfer**
[4] **negative transfer**
[5] a learner variety.

3.6.1.4 Cognitive complexity

In order to answer the first research question, it was necessary to classify questions into varying levels of cognitive complexity (the controlled variable) before attempting to investigate the relationship between readability and student performance at each level of cognitive complexity. One of the most well-used measures of the cognitive complexity of an assessment item is Bloom's taxonomy of educational objectives (Bloom 1956). It includes six major categories – knowledge, comprehension, application, analysis, synthesis, and evaluation – and continues to be used as a well-known standard tool for classifying test items.

I believe that the issue of cognitive complexity and linguistic difficulty have not been properly teased apart, as studies tend to focus either on student performance in relation to cognitive complexity only (e.g. Kropp & Stoker 1966 and Fairbrother 1975 cited in Seddon 1978) or to student performance in relation to readability only (e.g. Homan, Hewitt & Linder 1994, Hewitt & Homan 2004, Dempster & Reddy 2007). I have not come across any studies that analyse student performance in relation to both readability and cognitive complexity. It is surely to be expected that readability will not correlate with p-values if questions range from easy recognition to complex analysis, and equally that cognitive levels will not correlate with p-values if readability issues confound results. But it is possible that readability and cognitive complexity interact in complex ways, i.e. that students manage adequately until both are high, for example. In the present study my aim was to provide an exploratory investigation of the effects of Bloom's six-

level hierarchy not just on facility, but also on the discrimination and difficulty differential of questions and also to clarify the relationship between 'readable' and 'less readable' questions and 'difficult' and 'less difficult' questions in terms of the cognitive demands of the question (see section 4.6.4 below).

Each item was evaluated by three independent raters in terms of the cognitive complexity required to answer it (cf Bowman & Peng 1972, Hancock 1994). The raters were given one or both question papers and asked simply to provide a rating from 1-6 for each MCQ by comparing the demands of the question with Bloom's taxonomy of cognitive levels where level 1 required recall, level 2 comprehension, level 3 application, level 4 analysis, level 5 synthesis and level 6 evaluation. This is in line with methodology used by Hancock (1994), Ghorpade and Lackritz (1998), Martinez (1999), Paxton (2000) and Fellenz (2004). In this study the raters of the 2006 paper included myself, an external Independent Examinations Board (IEB) educational consultant, and a Unisa Professor of Education. The same raters were used in 2007 except that a Linguistics colleague was used instead of the Professor of Education as a result of the difficulty experienced by the latter in rating questions outside her discipline. Inter-rater reliability of the Bloom ratings was calculated using Fleiss' free-marginal kappa and ratings on which two of three raters agreed were used to rank questions according to various cognitive levels of difficulty. Item quality measures were then calculated separately for items at each cognitive level in order to establish whether questions at lower levels on the hierarchy were easier than questions at higher cognitive levels on the hierarchy (Bloom 1956:18). My research aimed to test this prediction empirically, as Hancock (1994) has done. Hancock (1994) showed that test items do not necessarily follow a general trend of increasing difficulty as the level of cognitive complexity increases, concluding that cognitive complexity and item difficulty are distinct attributes that are often not even correlated. My prediction is that linguistic difficulty may often be to blame for this lack of correlation, and that there may be interesting interactions between student performance, readability and cognitive complexity.

3.6.2 Qualitative research procedures: Interviews

The most qualitative aspect of the research involved exploring through interviews what the difficulties are in an MCQ examination from the students' viewpoint. Individual interviews that

included a think-aloud protocol probed students' actions, strategies and problems as they completed 80 MCQs. This method has the advantage of giving students a voice in exploring the issue of MCQ difficulty and particularly in answering my second and third research questions: What kinds of MCQ items present particular problems for L2 speakers? and What contribution do linguistic factors make to these difficulties?

Section 3.6.2.1 below gives some background to the advantages, aims and limitations of this methodology with respect to multiple-choice, while section 3.6.2.2 describes the think-aloud procedures used in this study.

3.6.2.1 Think-aloud methodology

Test-taker feedback can include either general examinee feedback about the face validity and fairness of the test or test method (e.g. Nield & Wintre 1986, Roberts 1993, Paxton 2000, Duffield & Spencer 2002, Struyven, Dochy & Janssens 2005), or so-called 'cognitive validation approaches' such as think-aloud protocols which involve more item-specific verbal reports (cf Norris 1990, Farr, Pritchard & Smitten 1990, Cohen 2007). The think-aloud protocol method was selected because I felt that it could provide important and detailed information that needs to be taken into account in a holistic view of MCQ difficulty. Although think-aloud protocols are highly context-specific and may be difficult to generalise from, the perspective that they could offer on the *process* of answering test questions rather than just the product, the test score, was important for my study and had precedents in the literature. Farr, Pritchard and Smitten (1990) were among the first to make use of the think-aloud method in order to identify test-taking strategies and difficulties encountered by 26 College students during MCQ reading comprehension tests. They justified their choice of method as follows:

> Rather than be bound by the use of product information such as test answers to infer process, this study sought to gain more direct data bearing on the test-taking process.
> (Farr, Pritchard & Smitten 1990:213)

Some of the issues that think-aloud protocols can shed light on in the case of multiple-choice assessment include the following: Where do students misunderstand the instructions, the question or the answers? Do problems come from the text or from some other source? If they are

located in the text, do they come from the design, style, organization, coherence or content (DuBay 2004:57)? Do students pick the right answers for the wrong reasons? What reasoning processes do students employ (Connolly & Wantman 1964)?, To what extent do they employ strategies such as narrowing down the possible options by eliminating the implausible, working backwards from response options, or double checking their own answers against response options (Martinez 1999:211)?

Support for the think-aloud interview methodology comes from Haladyna, Downing and Rodriguez (2002) who comment that

> The think-aloud procedure can provide insights into the cognitive processes underlying a student's encounter with a test item, but its limitation is the time it takes to collect data and evaluate the findings.
>
> (Haladyna, Downing and Rodriguez 2002:329)

Messick (1989) recommends that verbal reports of examinees' thinking on MCQ critical thinking test items ('analysis of reasons') are an important source of evidence for validating tests. Like other types of interviews, think-aloud protocols are most valid when the interviewees are representative of the group as a whole, are put at ease upfront and give written permission for their participation. For her part, the interviewer needs to listen carefully, record responses verbatim and keep her reactions to herself as far as possible (cf Farr, Pritchard & Smitten 1990, Ormrod & Leedy 2001:159-160).

Think-aloud protocols can either be 'introspective' (or 'concurrent') – asking subjects to explain their thinking as they go along, or 'retrospective' – asking subjects afterwards to explain why they chose the answers they chose. Norris (1990:42) reports on an empirical study in which he investigated the effect of verbal elicitation on introspective as opposed to retrospective different procedures for eliciting verbal reports, concluded that both of these yielded essentially the same information about the quality of subjects' thinking. Norris believes that introspective think-aloud protocols require less interviewer-subject interaction than retrospective requests for reasons (Norris 1990:53) and would therefore be preferable when more than one interviewer is used. As the only interviewer in this research, I intended to do both, i.e. ask students to think aloud as they

answer 80 questions and, if this is insufficient, also to explain their reasoning after selecting their answer.

While examinees' verbal reports of thinking on test items do provide direct data on the test-taking process, Norris (1990:41) points out that this data is only relevant if it does not alter students' thinking and performance from what it would have been while taking the test under normal silent pen-and-paper examination conditions. His research compared student performance when answering MCQs silently and with verbal reports (including thinking aloud or explaining each answer after making a choice (Norris 1990:46)). He found that verbal reporting, however it is elicited, does not generally affect students' thinking or performance and that this method is therefore a useful and legitimate way to validate an MCQ assessment. Norris (1990:54) suggests that subjects are unlikely to have much to say about recall items, but will have more to say on items requiring deliberative rather than automatic thinking, for example items requiring problem-solving, critical thinking or reading comprehension.

Research has shown that students vary in their ability to identify and explain the problems they experience. DuBay (2004:57) points out that 'In both usability testing and reading protocols, some subjects are more skilled than others in articulating the problems they encounter.' Pretorius (2005:41) explains that weak readers tend to have poor comprehension and recall and find it difficult to pinpoint where exactly they have difficulties understanding a text. Their reading behaviour is inappropriate in relation to the demands of the reading task. Conversely, good readers have well-developed metacognitive abilities in that they are much better able to monitor comprehension while they read and adopt repair strategies when comprehension breaks down:

> [Good readers] can usually pinpoint where in the text they experience difficulties, and they are also good at picking up inconsistencies in the text, recognising poorly written or incoherent texts, and perceiving the author's intentions. When they encounter comprehension problems, they backtrack on the text to reread sections and check their interpretation. They may even reread a section repeatedly until they feel satisfied with their understanding of it. They also tend to have fairly good recall of the main points in a text after reading it.
>
> (Pretorius 2005: 41)

3.6.2.2 Think-aloud procedures used in the study

A think-aloud rerun of the LIN103Y examination was undertaken with four 2006 students and nine 2007 students. I conducted the interviews between November 2006 and December 2007. Individual interviews took between one and three hours, at a venue of the student's choice (their homes, their offices, my office, the university cafeteria, coffee shops etc.). The think-aloud portion of the interview took between 30 minutes and two hours. The interviews were held no more than five weeks after the examination to make sure that students still remembered most of the course content.

I videotaped the first two interviews and then decided to switch to micro-audiotape for the next 11 because the equipment was simpler and less intrusive and because visual aspects of the think-alouds did not add much to the spoken interviews. All the interviews were subsequently transcribed verbatim from the tapes, omitting long digressions of a social nature, of which there were many. The students' utterances were indicated with their initials and my own contributions were indicated as P: in the transcripts. The students' marked-up examination papers with their MCQ answers were used together with the transcripts to determine the number of correct answers or 'think-aloud mark'. The transcripts were also analysed for evidence of the types of difficulties and misunderstandings and their causes in both L1 English and L2 English students. The questionnaire and standardised instructions used in the think-aloud interviews are provided in Appendix D.

In each case an initial structured interview was conducted in order to identify the student's various languages at home and at school, how long he or she had been studying at Unisa and elsewhere. Students were also asked to give their general opinion of MCQ as an assessment method, how difficult they find MCQ assessment and how they found the LIN103Y examination they had just written. Each of these last three questions was answered on a 5-point Likert scale and students were asked to justify their choice by explaining further. Students were also asked whether or not they had had enough time to finish the paper.

The introductory interview was followed by the following scripted explanation by the researcher of the purpose of the study and of what students should do:

> I want you to go through the paper as you would in an exam, talking me through the paper as you complete each question and explaining the thought processes that help you come to a decision about the right answer. For each question please
> - Mark the answer you think is correct
> - Mention any problems you have with the question
> - For each of the case studies, please state which topic/section of the Guide it relates to, e.g. language acquisition or men and women's language

I took notes during the interviews of the gist of their comments in case of equipment failure or unclear portions of the recording. I had to rely entirely on these notes in the case of the interview with DS where the video camera didn't work at all, and to a lesser extent in the CL and TH interviews where ambient noise made the audiotape unclear. I also took notes of my own observations during the think-alouds, for example of students referring back to the texts, rereading questions several times, taking a long time to answer a question, showing obvious signs of tiredness etc.

The interviews frequently became quite animated and informal, with students commenting on the course, the study guide, their backgrounds, giving their own examples of linguistic phenomena under discussion or just chatting about unrelated issues. Since Unisa students seldom have contact with their lecturers, many saw this as a valuable opportunity to discuss the course and ask for clarification. Although I attempted to keep a neutral face or simply nod whether their answers were right or wrong, the students tended to become concerned if I showed little reaction to an answer and probe further as to whether their reasoning was right or wrong and why.

3.7 Conclusion

This chapter has dealt with the choice of research design and the various theoretical perspectives that have influenced the study. It has outlined both the quantitative and qualitative approaches that will be brought together in the investigation of MCQ difficulty for English mother-tongue and for L2 Linguistics students. The quantitative and qualitative findings of the study will be presented in detail in Chapters 4 and 5 respectively.

Chapter 4
Quantitative results

4.1 Introduction

This chapter presents results of the quantitative portion of the investigation into linguistic aspects of MCQ difficulty. Section 4.2 provides a discussion of how the MCQs were classified in terms of their linguistic characteristics, followed in section 4.3 by results relating to the item quality of various types of questions (as measured by difficulty, discrimination and the difficulty differential between L1 and L2 students). Section 4.4 provides data and discussion relating to the extent to which the guideline relating to maximising readability was met and how readability impacted on difficulty, discrimination and the gap between L1 and L2 student scores. Section 4.5 considers the extent to which the empirical data support the other seven item-writing guidelines, namely those relating to incomplete statements, long questions, negative questions, very similar answer choices, grammatically non-parallel answer choices and AOTA and NOTA answer choices. This section also considers the statistical results relating to items that required reference back to the text. Section 4.6 considers the cognitive levels of the questions in terms of Bloom's taxonomy and the extent to which these impacted on item quality statistics for both more readable and less readable questions. Section 4.7 takes a closer look at the linguistic characteristics of the most difficult questions and the questions with the largest score differential between L1 and L2 students, and section 4.8 provides a conclusion to the chapter.

4.2 Item classification

For the purposes of the study the MCQs were classified according to various linguistic criteria relating to their item type, stem type and option type. The initial descriptive classification in this section provides an overview of the structure of the 2006 and 2007 examination papers and enables a comparison to be made between the item quality (difficulty, discrimination and difficulty differential) of various types of items in the remainder of the chapter.

The 80-item examination papers for 2006 and 2007 are provided in Appendices A and B respectively. A total of 18 of the items were identical in 2006 and 2007. There were nine scenario or case study texts in each paper, ranging from a few of lines of data to full-page extracts from academic articles. These were followed in each case by a set of 3 to 11 items relating to the topic raised in the case study. Before each reading passage students were instructed to 'Read the following case study and then answer Questions X to Y. Some of the questions relate directly to the case study while others test your knowledge of the relevant concepts in a more general way'. Haladyna (2004:84) supports the use of this kind of 'context-dependent item set', which he defines as an introductory stimulus followed by 2 to 12 items related to the stimulus, in order to increase the number of application and analysis questions in a multiple-choice assessment.

Table 4.1 below provides a comparative view of the different item types, stem types and option types in each of the two examination papers. The figures represent the number of MCQs of that particular subtype out of a total of 75 items in 2006 and 76 items in 2007 (see section 4.3.1 below for a discussion of discarded items).

Table 4.1 Item types, stem types and option types		**2006**	**2007**
Total		75	76
Item type	3 answer choices	0	2
	4 answer choices	44	47
	5 answer choices	31	27
	Long items (over 50 words)	14	14
	Text comprehension items that require the student to refer back to a passage	27	30
Stem type	Incomplete statement	46	50
	Missing word	1	1
	Question	28	24
	Completed statement	0	1
	Negative stem	15	11

Option types	At least one negative option	13	16
	Double negative (negative stem and one or more negative options)	5	4
	No similar answer choices	39	30
	2 similar answer choices	18	18
	3 similar answer choices	5	12
	4 similar answer choices	11	9
	5 similar answer choices	2	7
	Options not grammatically parallel	7	11
	AOTA	5	5
	NOTA	7	3

As the item-type portion of Table 4.1 indicates, the number of options was not consistent. Each paper included some five-option items (and two three-option items in 2007) although the majority of the questions had four options (59% in 2006 and 62% in 2007). These were not contrasted statistically in this study as this has been done extensively in recent years (e.g. by Andres & del Castillo 1990, Bruno & Dirkzwager 1995, Rogers & Harley 1999). Findings overviewed in Rodriguez (2005) tend to suggest that three options is usually sufficient as the fourth and fifth options tend to be selected by very small percentages of students and thus serve only to lengthen the test without improving discrimination (see earlier discussion in section 1.2.2.1). A total of 57 items (over one-third of the items) were text-based comprehension items that required the student to refer back to a case study text to find information or identify linguistic phenomena.

As far as stem type was concerned, the items were mostly incomplete statements, with one missing-word stem in each paper and one completed statement stem in 2007. The remainder of the items (about one-third of the items) were question stems. A fair number (20% in 2006 and 14% in 2007) were negative stems containing *not* or *false* or negative adjectives and adverbs such as *unconscious* or *unconsciously.*

The items were also classified according to their option types, including items with at least one negative option, and 'double negative' items which had the added complication of a negative stem *and* at least one negative option. Nearly half the 2006 items (48%) and 61% of the 2007 items had two or more similar answer choices. Similarity was defined for the purposes of the study as sharing at least half their words in the case of multiple-word options, and sharing a word stem in the case of single-word options. There were also 20 AOTA and NOTA items and a few items where the options were not all of the same phrase type and that were therefore not grammatically parallel.

4.3 Item quality statistics

A statistical analysis of the examination results yielded a discrimination index for each item (see section 4.3.1 below) and a measure of the facility of each item (see section 4.3.2). According to Knowles and Welch (1992) and Sireci, Wiley and Keller (1998), these two quantitative measures can be used to provide a picture of overall item quality. In order to ascertain the difference in performance between L1 and L2 students on each item, the difficulty differential was calculated by comparing the average test score for the L1 English students with the average test score for the L2 English students (see section 4.3.3 below). The average difference was used as a standard against which the L1/L2 difference on individual test items could be compared (a similar methodology was used to compare L1 and L2 performance in Dempster & Reddy 2007). Together, these three quantitative measures were used to provide a picture of overall item quality for various categories of items for comparative purposes. In most of this chapter the 2006 and 2007 statistics are reported separately in order to provide an indication of whether the data follows similar patterns in both years. In a few cases where the number of items were small, however, the data from the two years was conflated.

4.3.1 Discrimination

The average discrimination index (the item-total point-biserial coefficient) for the 2006 examination questions was 0,343, with a standard deviation of 0,174. The average discrimination for 2007 was very similar at 0,373, with a standard deviation of 0,172. The discrimination for individual questions ranged from -0,331 (Q26 2006) to 0,7 (Q5 2006). As explained in Chapter 2, higher discrimination is better, while negative or near-zero discrimination is indicative of a

poor question that does not effectively separate better from weaker students. Eleven questions had a discrimination below the recommended 0,125 (Kehoe 1995b, Kilpert 2008), with six displaying negative discrimination, but it should be noted that very easy questions cannot by definition score highly on discrimination, as the difference between the highest and lowest scorers cannot be large when almost everyone gets these questions right. This accounts, for example, for the low discrimination (0,010) on Q52 2006, which had a facility of 86%.

All six questions with negative discrimination were discarded prior to the statistical calculations reported in the remainder of the chapter. According to Kehoe (1995b), items with negative discrimination should be discarded because weak students do better on these items than students with higher overall marks in the test. The reliability of student rankings can thus be improved by discarding questions with negative discrimination (Kehoe 1995b). Discarded questions in this category were Q37, Q24 and Q26 (2006) and Q10, Q43 and Q50 (2007), in the latter two cases owing to a wrong answer key having been provided on the memorandum.

Other items which had to be discarded were Q16 (2006), which had no right answer, and Q76 (2006) and Q22 (2007) which were discarded before the final statistics were run due to errors in the questions. This left a total of 75 questions in 2006 and 76 questions in 2007. The average discrimination after discarding the problematic questions listed above was within the recommended range, at 0,372 for 2006 and 0,405 for 2007.

4.3.2 Facility

In 2006, the average percentage obtained for the paper was 67,7%, with a standard deviation of 15,97. The facility of individual items ranged from 33% correct (Q70) to 97% correct (Q12). After discarding the five problematic questions, facility averaged 69,1%. In 2007, the average percentage obtained for the examination as a whole was almost identical to 2006 at 67,8%, with a standard deviation of 20,8. The facility of individual items ranged from 20% correct (Q78) to 97% correct (Q19). After discarding the four problematic questions, facility averaged 70,2%.

Of the 151 remaining questions, only 17 had a facility of below 50%. These ‘difficult’ questions ranged in facility from 20% (Q78 2007) to 49% (Q72 2006 and Q76 2007). It should be pointed

out that there are many possible reasons for low facility, some of which are beyond the scope of the present study. These include questions with no right answer (like the discarded Q16 in 2006) or a wrong answer key provided (like the discarded Q43 and Q50 2007). The students may also have genuine misunderstandings about the study material or the material could have been poorly explained in the Guide. It is also possible that students simply didn't study certain (especially the last) sections of the study material. These issues are a fact of life in most MCQ examinations, but are beyond the scope of the present study, which will focus on linguistic reasons for question difficulty that are inherent in the wording of the questions themselves. This issue will be taken up in the remainder of the chapter after a discussion in section 4.3.3 below of how L2 students fared in comparison to L1 students.

4.3.3 L1 – L2 Difficulty differential

On average, the L1 speakers scored 78% in both years and the L2 speakers 63% in 2006 and just under 64% in 2007. The average score difference between L1 and L2 students (what I term the 'difficulty differential') was therefore 15% and 13,8% respectively. The difficulty differential is obviously due to a number of factors, including but not limited to the poorly-resourced and dysfunctional South African government school system for black children prior to 1994 and a systemic and ongoing problem with teaching reading at schools (see discussion in section 1.2.1). However, in the case of individual MCQs where the difficulty differential is considerably higher than average, it is almost certainly due to problems in in interpreting the language of the test question.

It was decided to take a difficulty differential of 25% or more as the cutoff point at which questions would be investigated further, i.e. where the difference between L1 and L2 students was at least 10% higher than the average difference between these two groups. Dempster and Reddy (2007) used a similar methodology in their investigation of items with a score differential of 22% or more between L1 and L2 South African students (the average score difference was 10%). In 2006 18,6% of my test items had a difficulty differential of 25% or more, in the worst case up to 54%, thereby strongly favouring L1 speakers over L2 speakers. In 2007 14,5% of the items had a difficulty differential of 25% or more. Both tests can be said to be unfair in some respects to L2 speakers, with 2007 slightly less problematic than 2006. The items with a high

difficulty differential are listed in increasing order in Table 4.2 below. These represent questions that were particularly hard for L2 students:

Table 4.2 Questions with a difficulty differential of 25% and over

2006			2007		
Question	**Difficulty differential**	**Facility**	**Question**	**Difficulty differential**	**Facility**
23	26%	54%	62	25%	58%
31	26%	76%	32	25%	68%
71	27%	51%	49	27%	70%
69	29%	48%	73	27%	60%
7	29%	48%	77	28%	66%
70	30%	33%	75	30%	50%
80	30%	51%	56	31%	70%
35	31%	46%	47	33%	56%
10	32%	65%	16	34%	62%
38	33%	60%	17	36%	50%
78	36%	67%	76	38%	49%
61	37%	57%			
5	47%	42%			
53	54%	49%			

The preponderance of high-difficulty differential questions very late in the question papers (2006 Q69, Q70, Q71, Q76, Q78, Q80 and 2007 Q73, Q75, Q76, Q77) suggests that L2 students may have been struggling to complete the paper and therefore guessing the last few questions, with L1 students doing much better on these questions. Another possibility is that many L2 students didn't make it to the end of the Guide in their examination preparation and therefore did considerably worse than L1 students in the final sections of the work, or simply were less interested in the later study units on gendered language and sign language (see discussion in section 5.7.5).

As can be seen in the rather low facility figures in Table 4.2 above, the questions with a high difficulty differential tend to be hard not only for L2 students but for all students. The correlation coefficient of the facility and difficulty differential data sets provides an indication of the relationship between the two properties. The downward trend of the scatterplot in Figure 4.1 illustrates the negative correlation coefficient of -56,1% between facility and difficulty differential in 2006. Figure 4.2 illustrates a negative correlation of -25% between facility and difficulty differential statistics in 2007. In both cases there was a trend for the more difficult questions to display a greater difficulty differential, i.e. a greater score difference between L1 and L2 speakers than the easier questions. L2 speakers therefore struggled more than L1 speakers as questions became more difficult.

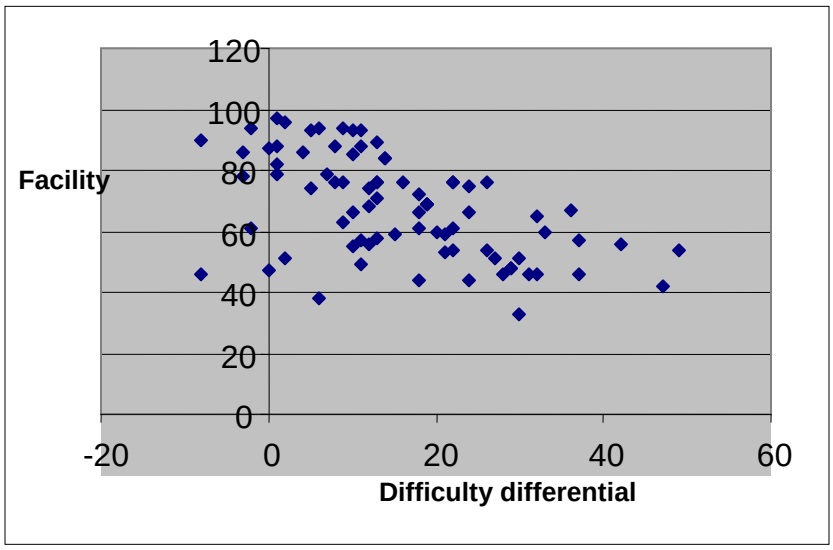

Figure 4.1 Scatterplot of facility versus difficulty differential 2006

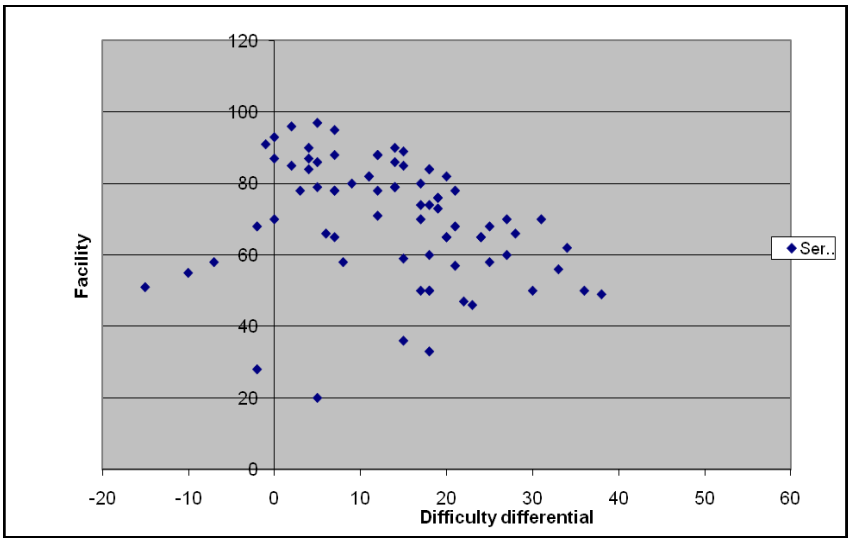

Figure 4.2 Scatterplot of facility versus difficulty differential 2007

The outliers to the left of the vertical line indicate the 11 questions with negative difficulty differential, in other words where L2 students did *better* than L1 students. The extent of the negative difficulty differential ranged from -1% (Q30 2007) to a substantial -15% (Q44 2007). Although the questions with negative difficulty differential in 2006 (Q9, 34, 52, 73 and 74) were all easy questions, with facility above 74%, this wasn't the case in 2007, where the facility of the questions with negative difficulty differential (Q14, 30, 31, 44, 51 and 61) ranged from 28% (Q14) to 91% (Q30).

It is not easy to identify patterns here as to why these 11 questions may have been easier for L2 than L1 students. However an attempt to analyse possible reasons is important in the context of my study in that it may provide insight into ways to make questions fairer for L2 students. One possibility is that the African language data in four of the questions (Q52 2006, Q30 2007, Q31 2007 and Q61 2007) helped L2 speakers despite the fact that the translations provided should have made these questions clear to all students. For example:

52. The Xhosa terms *isosara* 'saucer' and *irum* 'room' are examples of

*[1] borrowing
[2] codeswitching
[3] interference
[4] convergence.

Four of the items (Q14 2007, Q51 2007 and Q9 2006) were negative items asking 'Which of the following statements are **false**?' or containing negative answer choices like Q44 2007 (negatives highlighted in bold) :

44. Genie's case shows that

[1] a child's first language is acquired easily after puberty
*[2] it is **im**possible to achieve proficiency in a first language after puberty
[3] a second language **cannot** be successfully learned before puberty
[4] a second language **cannot** be successfully learned after puberty
[5] None of the above.

This item had the highest negative difficulty differential, with 40% of L1 students answering correctly and 55% of L2 students answering correctly. The issue here may have been that testwise students elected not to choose [2] because of the overgeneralisation implicit in the word 'impossible', preferring the incorrect NOTA option instead (see further discussion in section 5.5 and 5.7.3).

Despite the unnecessary and undesirable complexity associated with double negatives (cf Cassels & Johnstone 1980, Johnstone 1983, Tamir 1993), the double negative Q9 2006 (see below), which had a negative stem and a negative answer choice [4], was answered correctly by 84% of L1 speakers and 92% of L2 speakers. The double negative appears therefore to have been more confusing for L1 speakers than for L2 speakers in this particular case.

9. Which of the following statements is **false**?

[1] Language acquisition follows similar developmental stages in all children irrespective of the language they acquire.
[2] All children acquire the language to which they have been exposed.
[3] A child has the potential to acquire any language.
[4] Exposure to language is not necessary for language acquisition.

Q73 2006 (see below) was a complex 'Type K' MCQ, with combinations of choices offered. The very low discrimination (0,071) also indicates that this question is problematic. This format is long, unnecessarily complex and is not recommended by Haladyna (2004:80):

73. Which of the following is an articulatory feature of sign language signs?

[1] facial expression
[2] palm orientation
[3] location of the sign relative to the body
[4] [2] and [3] are correct
*[5] [1], [2] and [3] are correct

It was hoped that further indications of the nature of the difficulty in these and other questions would be obtained from the think-aloud interviews (see section 5.7).

4.3.4 Summary

The three measures used in this study to provide an indication of item quality were discussed in the section above, namely discrimination, facility and difficulty differential. These measures were very similar in 2006 and 2007, with facility averaging 69,1% and 70,2%, discrimination averaging 0,372 and 0,405 and the L1 – L2 difficulty differential averaging 15% and 13,8% respectively. Consideration was also given to the questions with negative difficulty differential, where L2 students outperformed L1 students. The average item quality measures above will be used for comparative purposes in the discussion of particular subgroups of items presented in the rest of the chapter.

4.4 Readability

Item-writing guidelines recommend that test-setters should keep items as readable as possible, to meet the needs of the weakest readers in the group. In the following section, I focus on the readability measures of sentence length, unfamiliar words, density of academic words and Dale-Chall readability score in relation to item quality statistics. Readability scores for each of the 80 multiple-choice items were calculated using the new Dale-Chall readability formula (Chall & Dale 1995), based on the number of words per sentence and the number of unfamiliar words per sentence. These values were converted to 100-word equivalents as explained in section 3.6.1.3 and then to American grade-equivalent reading levels using the Dale-Chall tables in the normal way.

4.4.1 Sentence length

The average sentence length was identical in both examination papers at 32,8 words, ranging from 13 to 90 words in 2006 and 10 to 94 words per item in 2007. MCQs are by their nature often long sentences as they frequently offer four or five possible endings for an incomplete stem. Items that were 50% longer than average (50 words or more) were classified as long items. The 14 long items in each examination paper are investigated further in section 4.5.2 below.

4.4.2 Unfamiliar words

In each item there was an average of nine unfamiliar words (words not on the 3000-word list) (Dale & O'Rourke 1981, Chall & Dale 1995), ranging from one to 25 in 2006 and one to 21 in 2007. In 2006 facility (question difficulty) did not correlate with the number of unfamiliar words on the Dale-Chall list (correlation 2,8%). In 2007 question facility had a better correlation with the number of unfamiliar words on the Dale-Chall list at -20,7%. This indicates, not surprisingly, that questions with more unfamiliar words tended to be more difficult, which is why the proportion of unfamiliar words is used as an important indicator of readability in most readability formulas including the Dale-Chall formula discussed in 4.4.4 below. The mixed results for 2006 and 2007 mean, however, that no clear pattern was discernible as regards the relation between readability and facility using this measure.

4.4.3 Academic words

Since the multiple-choice assessment was at first-year university level, the density of academic words per item using Coxhead's Academic Word List (Coxhead 2000) was used here as an additional way of comparing the lexical difficulty of individual items.

AWL words are 'unfamiliar' by definition as they exclude the 3000 most common word families in general use (West 1953, Coxhead 2000) and are common in university-level texts but certainly not in the fourth-grade texts which form the basis of West's (1953) list of familiar words. A high number of academic words in a question will therefore lead to a high number of unfamiliar words and a high readability score. Most discipline-specific terminology would not appear on either the AWL (Coxhead 2000) or the list of familiar words used by Chall and Dale (1995), but should be fully explained in the study material and therefore 'familiar'. However, there is no

doubt that the discipline-specific words are difficult for students to learn and internalise, although mastery of the vocabulary is an important and unavoidable part of the knowledge required in any content subject.

The Academic Word List words (Coxhead 2000) in each multiple-choice item were identified using the AWL Highlighter website (Haywood n.d.). The AWL Highlighter automatically identifies words from the AWL in any pasted text of your choice and highlights them in bold. The results of this procedure can be seen in the bold words in the examination papers in Appendix A and B. For example, Q28 in the 2006 examination contained the highest number of academic words (13 tokens and 10 types, given the repetitions of *period*):

28. The critical age **hypothesis** states that

[1] children are better second-language learners than **adults**
[2] the younger a person is when he or she starts to learn L2, the more likely it is that he or she will **attain** near-native competence.
[3] there is a **specific period** for first language **acquisition,** after this **period** it is difficult and maybe even impossible to **acquire** language
[4] in order to become bilingual, children need special **instruction** during a **specific ‘**critical' **period**
[5] there is a **period** between birth and puberty when a second language is usually **acquired.**

There were an average of 2,7 AWL words per question in 2006 and 2,1 AWL words per question in 2007, ranging from zero to a maximum of 13 academic words in a single question (Q28 2006 above). The AWL density per question (AWL word tokens as a percentage of total words) ranged from 0 to 0,43 (Q24 2007), averaging 0,08 in 2006 and 0,07 in 2007. Question difficulty correlated very weakly with the density of AWL words at -1,5% in 2006 and -6,5% in 2007, indicating a very slight trend in both years for the questions with more academic words to be more difficult (lower facility). This mirrors the slight trend reported in 4.4.2 above for the questions with more unfamiliar words to be more difficult, although this applied only to the 2007 data.

Items with double the average AWL density (0,15) (where at least 15% of the words were AWL words) were presumed to be 'very difficult' in terms of vocabulary load and were investigated further, as were an intermediate level of 'difficult' items where at least 12% of the words were AWL words. The item quality measures for these 'high AWL density' items are given in Table 4.3 below.

Table 4.3 Item quality measures for high AWL density items

	2006			**2007**		
Item type	**AWL density over 0,12**	**AWL density over 0,15**	**Overall**	**AWL density over 0,12**	**AWL density over 0,15**	**Overall**
Total number	18	12	75	17	8	76
Average facility	66,1%	64,1%	69,1%	67,8%	68,5%	70,2%
Average discrimination	0,405	0,391	0,372	0,440	0,501	0,405
Average difficulty differential	15,3%	16,1%	15,0%	15,1%	18,5%	13,8%

Table 4.3 indicates that in both years items with high AWL density (above 0,12) were more difficult than average, more discriminating than average and had a larger difficulty differential than average. Items with double the average AWL density (above 0,15) had an even higher average difficulty differential. This finding implies that a heavy load of academic vocabulary can disadvantage L2 students and increase the L1 – L2 score gap.

AWL density is therefore a possible measure that can be used prior to a test to get a quick indication of the questions that could usefully be simplified to reduce the gap between L1 and L2 scores. Both AWL words and number of words are easy to count using the *AWLhighlighter* website (Haywood n.d.) and Microsoft Word respectively, and if a benchmark maximum level is set (say 0,15 or twice the average AWL density value), only a small number of questions would have to be reviewed. Even more practical would be to look at all questions with more than six AWL words (since the average question length is 33 words, questions with six or more AWL

words are fairly likely to have an AWL density of 0,15 or more). Substituting some of the AWL words in just these questions with more familiar synonyms or paraphrases would probably help lessen the gap between L1 and L2 scores. This is an area that could be addressed by future experimental research.

4.4.4 Readability scores

As can be seen in Table 4.4 below, readability scores for individual items ranged from Reading Level 2 (a standardised reading level appropriate for American second-graders) to Reading Level 16 (a standardised reading level appropriate for American postgraduate students), with 54 items (36%) at the highest reading level. The examinations could therefore be said to be difficult in terms of the reading proficiency required. This large number of items at Reading Level 16 is partly due to the long sentences that are a feature of MCQs and partly due to the high number of unfamiliar words. It should be borne in mind that technical linguistic terminology, although classified as unfamiliar according to the Dale-Chall definition, should be familiar to students by the end of a semester of linguistics and is also an intrinsic part of the content that needs to be tested.

As a result of the methodology of counting the number of sentences in questions of less than 100 words (10 - 94 words per item) and converting these to the number of sentences that would occur in 100-word samples, it can be argued that there was a systematic inflation of the final readability score. This is because very long sentences would probably be followed by some shorter-than-average sentences in genuine 100-word samples (cf Allan, McGhee & van Krieken 2005). The reading levels should for this reason not be taken at face-value (e.g. 13-15 = a standardised reading level appropriate for American undergraduate students) but should be used primarily to compare the readability of individual items with each other.

In 2006, the Dale-Chall readability score of the items correlated only very slightly with facility at -8,6%, showing a slight tendency for higher readability scores to be associated with lower facility (more difficult questions). In 2007 there was a positive correlation coefficient of 12,4%, suggesting that questions with higher readability scores (*less* readable questions) were easier. These conflicting results suggest that the effect of readability is subordinate to other aspects of a

question that are more important to its overall facility, for example the cognitive complexity of the question (cf also Hewitt & Homan 2004, Dempster & Reddy 2007).

An attempt was made to shed more light on the relationship between readability and facility by calculating item quality statistics separately for groups of questions with similar readability scores. Table 4.4 below groups questions into one of five readability levels: below 8, 9-10, 11-12, 13-15 and 16. The item quality statistics for items at various grouped readability levels are indicated below.

Table 4.4 Readability level of items

	2006				**2007**			
Dale-Chall reading level	**No. items**	**Ave. facility**	**Ave. discrim-ination**	**Ave. difficulty differenti al**	**No. items**	**Ave. facility**	**Ave. discrim -ination**	**Ave. difficulty differenti al**
2	1	75,3%	0,345	11,7%	1	68,5%	0,455	13,5%
3	1				0			
4	2				2			
5-6	4				6			
7-8	7				6			
9-10	14	63,4%	0,325	12%	11	68,5%	0,438	13,3%
11-12	9	67%	0,394	20,1%	16	64,9%	0,373	12,6%
13-15	11	66,1%	0,403	18,5%	6	77,5%	0,392	10,7%
16	26	70,5%	0,393	15,2%	28	73,1%	0,397	15,5%

Item quality statistics for 2006 show that the most readable questions (Reading level below 8) were the easiest at an average facility of 75,3%. They were also lower in discrimination (at 0,345) than the less readable questions (11-12 and above). The 2007 statistics are surprisingly different, with the most readable questions (below 8) proving to be *more* difficult at 68,5% than the least readable questions (level 13-15 77,5% and level 16 73,1%). The questions pitched

below readability level 8 also surprisingly had higher discrimination (0,455) than the least readable questions (0,397). This indicates again that the relationship between readability and question facility is not a direct one, as it is confounded by the cognitive level at which questions are pitched, from simple recall to more complex tasks such as extrapolating, identifying reasons for assertions, or applying abstract concepts to unseen real-life case studies.

While there were no clear patterns as regards difficulty and discrimination for more readable and less readable questions, the difficulty differential showed a clearer pattern. The more readable items (below level 8) showed a lower average difficulty differential (11,7% in 2006 and 13,5% in 2007) than the level 16 items which averaged a difficulty differential of 15,2% and 15,5% respectively. The consistent picture evident here is that L2 students do worse than L1 students by a bigger margin when readability scores are high (i.e. when questions are less readable). This supports the guideline to make items as readable as possible.

It should also be pointed out that the AWL density measure described above was more useful than Dale-Chall readability scores in its ability to reflect question difficulty and difficulty differential. Calculating just the percentage of AWL words in an MCQ was a better indicator of its difficulty than the more complex calculation associated with Dale-Chall readability scores, which involved counting words, counting the number of sentences, counting the number of unfamiliar words not on a given list, converting to a 100-word equivalent and reading the resulting scores off a table to obtain a Reading Level score. Items with high (above 0,12) or very high (above 0,15) AWL density were more difficult, more discriminating and had a larger difficulty differential than average, while the same could not be said of the high readability level items. High Dale-Chall reading levels were reflected only in a higher difficulty differential between L1 and L2 students, but not in better discrimination or a lower percentage of right answers. The AWL density measure therefore has some potential as a pre-test check on MCQ readability and fairness.

MCQ guidelines enjoin writers to make their questions as readable as possible, and there is no doubt that this is an important consideration when setting MCQs. The findings above suggest that a heavy load of academic vocabulary, unfamiliar words and very long sentences can

disadvantage L2 students and increase the L1 - L2 score gap. While none of these linguistic features is entirely avoidable in university-level MCQs, fairness will be compromised when these go beyond a certain level.

4.5 Item quality statistics relating to MCQ guidelines

While section 4.4 considered the effect of readability on question difficulty, the following section investigates various other language guidelines relating to MCQs by comparing the item quality measures of facility, discrimination and difficulty differential of items that adhere to the guideline and items that disregard the guideline. Seven other guidelines under investigation are listed below and taken up in sections 4.5.1 to 4.5.7 below:

(a) State the stem in a question format instead of a completion format
(b) Keep options brief
(c) Word the stem and options positively; avoid negative phrasing
(d) Avoid answer choices that are too similar
(e) Avoid All of the above (AOTA)
(f) Avoid None of the above (NOTA)
(g) Try to keep options grammatically parallel

Although it is not a language guideline as such, Haladyna (2004) also advocates the use of context-dependent text-comprehension items, and the statistics for these items are discussed in section 4.5.8 below.

4.5.1 Questions versus incomplete statement stems

MCQ guidelines recommend that questions are preferable to incomplete statements in MCQ stems. Both types of stem were used in the examination papers under investigation, with the majority being incomplete statements (see Table 4.1 above). Item quality measures for items with question stems and incomplete statement stems are given in Table 4.5 below. For the purposes of the statistics, missing word stems were counted as incomplete statements. One 2007 item (Q76) was excluded as its stem was a completed statement.

Table 4.5 Item quality measures for question stems and incomplete statements

	2006			2007		
Stem format	**Question stems**	**Incomplete statement stems**	**Overall**	**Question stems**	**Incomplete statement stems**	**Overall**
Total number	28	47	75	24	51	76
Average facility	69,1%	69,1%	69,1%	66,3%	72,4%	70,2%
Average discrimination	0,357	0,381	0,372	0,405	0,391	0,405
Average difficulty differential	14,8%	15,1%	15,0%	13,8%	13,5%	13,8%

As the above figures illustrate, the combined item quality measures of question stems versus incomplete statement stems differed only marginally. In 2006 the question stems and incomplete statement stems had exactly the same facility at 69,1% and question stems were 0,024 less discriminating. The difficulty differential was almost identical for question stems and incomplete statement stems at 14,8% and 15,1% respectively. In 2007, question stems were 6,1% more difficult than incomplete statement stems on average, while discrimination differed by only 0,014. The difficulty differential was almost identical for question stems and incomplete statement stems at 13,8% and 13,5% respectively. These figures indicate that neither first nor second language students appear to find incomplete statements more challenging than question stems. Discrimination and difficulty differential measures for the two stem types are almost identical, although question stems were harder than incomplete statement stems in 2007. The guideline to avoid incomplete statements in favour of question stems is therefore not supported by the data. This accords with the findings in Sireci, Wiley and Keller (1998).

While the statistics indicated little difference in the item quality of question stems and incomplete statement stems, this does not detract from Haladyna's (2004:108) point that placing the main idea in the stem rather than in the answer choices has the advantage that the test taker knows what is being asked in the item after reading only the stem. This presumably enables the

student to identify the relevant topic faster and access the relevant portion of long-term memory before reading the options.

4.5.2 Long items

MCQ guidelines suggest keeping each item as brief as possible to minimise reading time (Haladyna 2004:106-108). This advice applies both to the stem and the options. The guideline is contravened by questions with irrelevant information ('window dressing') in the stem that distracts students from the problem itself. These types of stems are more informative and longer than required and thereby disregard Grice's maxims of quantity, manner and relevance. The guideline recommending brevity is also contravened by items where material is repeated at the beginning of each answer choice instead of being moved to the stem. For example in Q57 2006 the superfluous repetition of *The linguistic phenomenon illustrated above* resulted in a 67-word item instead of a 52-word item with the repeated portion shifted to the stem:

57. Which of the following statements is **false**?

[1] The linguistic phenomenon illustrated above involves the use of two or more languages in the same conversation.
[2] The linguistic phenomenon illustrated above is a way for speakers to express their common identity as bilinguals.
*[3] The linguistic phenomenon illustrated above is a sign of laziness and linguistic decay.
[4] The linguistic phenomenon illustrated above is often linked to a change in topic.

As shown in 4.4.1 above, the items averaged 32,8 words. Items of 50 words or more were classified as long items, as these had 50% more words than average. The item quality measures for the 28 long items are given in Table 4.6 below:

Table 4.6 Item quality measures for long items (50 words or more)

	2006		2007	
Item type	**Long**	**Overall**	**Long**	**Overall**
Total number	14	75	14	76
Average facility	68,6%	69,1%	61,8%	70,2%
Average discrimination	0,427	0,372	0,461	0,405
Average difficulty differential	17,7%	15,0%	17,3%	13,8%
Average unfamiliar words	14	9	14	9

As one might expect, longer items were slightly more difficult than average at 68,6% in 2006 (0,5% below the average facility) and 61,8% in 2007 (8,4% below average). They were more discriminating than average and had above average difficulty differential (2,7% above average in 2006 and 3,5% above average in 2007). The high difficulty differential suggests that long items are causing more problems for second language students than first language students. One possible explanation for the high difficulty differential is that long items have more unfamiliar words than average (14 as opposed to 9), which adds to their comprehension difficulties. It is also possible that reading to the end of a long item while trying simultaneously to remember the stem overloads the limited capacities of working memory. In terms of Perfetti's Verbal Efficiency Theory of reading (1988), the various cognitive processes that are required to read and understand a text share a limited resource pool. Within this theory, overall comprehension is impeded when more attention is required to decode a question and retain it in working memory (Perfetti 1988, Pretorius 2000). Since the amount of reading practice and the degree of automatisation of processes affects the extent to which a process requires attention, L1 readers might be at an advantage over L2 readers with respect to reading speed and attention demands, resulting in better comprehension in non-optimal circumstances such as this. There is therefore a case to be made for limiting items to under 50 words. Because many MCQs consist of single sentences, reducing the length of items will also reduce sentence length and improve readability.

4.5.3 Negative items

Negatives such as the *not* in Q61 2006 below are said to be best avoided as they make items more difficult to comprehend and are easy to miss (Cassels & Johnstone 1980, Johnstone 1983, Harasym et al. 1992, Tamir 1993, Haladyna 2004:111).

61. Which of the following does **not** generally contribute to language shift?

*[1] the structure of the language
[2] attitudes towards the language
[3] increased contact between languages
[4] persecution of speakers of the language
[5] migration of speakers to urban areas.

Roberts (1993) mentions that trick questions often contain a confusing use of the negative that actually turns out to be crucial. An example of this is Q11 2006, where a tiny but critical negative morpheme (*un-*) was easy to overlook in options [1] and [2]:

11. Which of the following is a typical characteristic of caretaker speech?

[1] flat, unchanging intonation
[2] short, ungrammatical sentences
*[3] frequent repetition
[4] long sentences

It is standard practice to highlight negative words such as *not, never* and *false* typographically in item stems (Haladayna 2004:117). This was done by using bold type for the words **not** and **false** in item stems in both examination papers, but not for any other negatives in the stem or the answer choices. (Note that these words are not bold in Appendix A and B as only AWL words are in bold, but they were in bold in the original examination papers.)

In 2006 there were 23 negative items in total, of which the item quality measures are listed in Table 4.7 below. These included 15 questions with negative stems, namely seven stems with *not* (Q5, Q33, Q40, Q61, Q66, Q68, Q77), five with *false* (Q9, Q49, Q57, Q72, Q78), and one (Q44) asking students to identify *disadvantages*. The negative adjectives and adverbs *unconscious* and

informal were included in the stem of Q6 and *unacceptable, unspoken* and *unintentionally* in Q31. These were counted as negative items, although it is possible that negatives are not a homogeneous group in terms of the way they function in MCQs. There were nine items with one negative answer choice, namely Q5[3] *do not*, Q9[4] *not*, Q26[2] *not*, Q27[4] *never*, Q28[3] *impossible*, Q41[4] *unstable*, Q48[2] *other than*, Q78[3] *misunderstandings*, 80[3] *reject the compliment by denying that it is true.* Three items had two negative answer choices, namely Q11[1] *unchanging* and Q11[2] *ungrammatical*, Q45[2] *not* and Q45[4] *unrelated*, 62[1] *disability* and Q62[2] *imperfectly*. Note that Q5, Q9 (see below) and Q26 and Q78 were double negatives, with negative stems and at least one negative answer choice (negatives are highlighted in bold):

9. Which of the following statements is **false**?

[1] Language acquisition follows similar developmental stages in all children irrespective of the language they acquire.
[2] All children acquire the language to which they have been exposed.
[3] A child has the potential to acquire any language.
[4] Exposure to language is **not** necessary for language acquisition.

In 2007 11 of the questions had negative stems, namely three stems with *not* (Q67, Q72, Q73) and eight with *false* (Q5, Q12, Q14, Q21, Q33, Q47, Q51, Q59). There were also ten items with one negative answer choice, namely Q5[3] *unconsciously*, Q12[3] *unconsciously*, Q21[2] *are not*, Q27[2] *is not*, Q42[1] *do not*, Q46[5] *no*, Q58[2] *other than*, Q72[1] *fail to*, Q76[4] *no* and Q79[2] *no.* There were 6 items with two or more negative answer choices, namely Q11[2] *a negative attitude* and Q11[4] *unwilling*, Q34[2] *little or no* and Q34[4] *cannot*, Q44[2] *impossible*, Q44[3] *cannot*, Q44[4] *cannot*, Q48[2] *cannot*, Q48[3] *no*, Q48[4] *no*, Q56[2] *are not*, Q56[4] *unrelated*, Q68[1] *disability*, Q68[2] *imperfectly.* Q5, Q12, Q21 and Q72 had the double complexity of negative stems and at least one negative answer choice. There were therefore 23 negative items in total, of which the item quality measures are listed in Table 4.7 below:

Table 4.7 Item quality measures for negative items

	2006		2007	
Item type	**Negative item**	**Overall**	**Negative items**	**Overall**
Total number	23	75	23	76
Ave. facility (%)	66,1	69,1	65,8	70,2
Ave. discrimination	0,407	0,372	0,413	0,405
Ave difficulty differential (%)	18,3	15	14,7	13,8

Table 4.7 indicates a consistent picture where negative questions are both more difficult and more discriminating than average. This suggests that negative questions can be effective questions in terms of sorting better from weaker students. The average difficulty differential of the negative questions is also considerably higher than the average difficulty differential of the papers as a whole, indicating that negative items are more difficult for L2 speakers than L1 speakers. The implication here is that not only are negative questions generally harder than positive ones, as the literature suggests (e.g. Cassels & Johnstone 1980, Johnstone 1983, Harasym et al. 1992, Tamir 1993, Haladyna 2004:111), but the difference is greater for L2 than for L1 speakers of English. The small advantage of negative questions (higher than average discrimination) is therefore offset by the disadvantages (higher difficulty differential and the possibility of students missing the negative words). The guideline to avoid negatives is therefore supported by the data.

In order to test my intuition that negative MCQs are not a homogeneous group and that some types of negatives are more difficult than others, statistics were calculated separately for the items with negative stems as opposed to negative options as well as for ‘double negative’ items. Results are presented in Table 4.8 below:

Table 4.8 Item quality measures for negative (N) stems, negative options and double negatives

	2006					2007				
Item type	**Any N**	**N stem**	**N option**	**Double N**	**Over-all**	**Any N**	**N stem**	**N option**	**Double N**	**Over-all**
Total number	23	15	13	5	75	23	11	16	4	76
Ave. facility (%)	66,1	62,4	68,4	61,0	69,1	65,8	56,3	69,9	56,0	70,2
Ave. discrimination	0,407	0,429	0,392	0,433	0,372	0,413	0,462	0,387	0,443	0,405
Ave difficulty differential (%)	18,3	20,2	16,6	19,4	15	14,7	17,1	14,1	18,5	13,8

The facility statistics indicate that items with negative stems are more difficult than items with negative answer choices. Items with double negatives are the hardest of all. This accords with Cassels and Johnstone's (1980) findings that double negatives result in poor student performance. Johnstone (1983:115 cited in Tamir 1993:311) suggests that double negatives may take up to four times the working memory of positive equivalents. In terms of discrimination, double negatives and negative stems were more discriminating than negative answer choices. Difficulty differential statistics show double negatives and negative stems have a larger differential between L1 and L2 students than negative answer choices.

The three item quality statistics for items with negative options differ only very slightly from the overall average values, so these questions do not appear to pose particular problems. It is only items with negative stems and double negatives (items with a negative stem and at least one negative answer choice) that differ considerably from the average values, being harder, more discriminating and displaying a higher difficulty differential. These results may be interpreted as supporting the guideline to avoid negatives, where negatives is understood to refer specifically to negative stems and double negatives but not to negative answer choices.

4.5.4 Similar answer choices

Roberts (1993) suggests that very similar answer choices may be unnecessarily confusing to students and require closer reading to differentiate the options from one another. In this study, answer choices that shared at least half their words (such as answer choices [1] and [5] in Q7 2007 below) were classified as similar. The singular and plural of the same word were counted as the same word, as were *a* and *an.* Single-word answers that shared a morphological stem were also counted as being very similar as these are often confused by students. For example, *dialect* and *sociolect* were counted as very similar answers because of the shared stem *-lect* but *borrowing* and *codeswitching* were not as the shared *–ing* is an affix. Discipline-specific terminology (including linguistics terminology) tends to make heavy use of this kind of shared-stem-different-affix contrast to make important semantic distinctions, resulting in pairs of words like *convergence* and *divergence*, *monolingual* and *bilingual,* and groups of similar words like *dialect, idiolect, ethnolect* and *sociolect.*

The number of very similar answer choices per item ranged from 0 to 5. Items with 2+2 similar answer choices were counted as having four similar answer choices. Items with 2+3 similar answer choices like Q7 2007 below were counted as having five similar answer choices:

7. In Mr Dlamini's case, learning German involves

[1] *first language acquisition*
[2] second language learning
[3] spontaneous language learning
*[4] foreign language learning
[5] *third language acquisition.*

Less than half the items (45% in 2006 and 39% in 2007) had no similar answer choices. A total of 41 items in 2006 (55%) and 46 items in 2007 (61%) had at least two similar items and a fifth of the items (21%) had three, four or five similar options (see Table 4.1 above). Table 4.9 below compares the item quality statistics for items with no similar options, two similar options and more than two (three, four or five) similar options.

Table 4.9 Item quality measures for 0 similar, 2 similar and 3 or more similar answer choices

	2006				2007			
Option type	**0 similar**	**2 similar**	**3 or more similar**	**Overall**	**0 similar**	**2 similar**	**3 or more similar**	**Overall**
Total number	34	19	22	75	30	17	29	76
Average facility	69%	69,6%	68,8%	69,1%	68,9%	66,1%	73,9%	70,2%
Average discrimination	0,363	0,380	0,379	0.372	0,382	0,456	0,386	0,405
Average difficulty differential	15,4%	15,6%	13,7%	15,0%	13,1%	18,9%	11,5%	13,8%

In 2006 facility was almost identical for no similar, two similar, and three or more similar answer choices. Discrimination was also very similar for these groups. The difficulty differential was in fact lowest for items with three or more similar answers at 13,7%, suggesting that L2 students are not performing worse than average when three or more of the answer choices are similar. 2007 showed a similar pattern, with items with three or more similar answer choices displaying the highest average facility, the lowest discrimination and the lowest difficulty differential. Items were therefore not harder when there were three or more similar answer choices. Rather surprisingly, the statistics in both 2006 and 2007 suggest that questions with two similar answer choices are harder for L2 speakers than questions with three or more similar answer choices (15,6% difficulty differential as opposed to 13,7% in 2006 and a substantial 18,9% difficulty differential as opposed to 11,5% in 2007). It therefore seems to be L2 speakers that are particularly at a disadvantage when two of the answer choices are similar. This could be due to misreading small distinctions, difficulty distinguishing or remembering similar forms and meanings or accidentally picking the wrong one of two similar answer choices when transferring answers to the mark reading sheet. Three or more similar answer choices did not affect the item quality negatively and do not need to be avoided, but it may be better for the sake of the L2 speakers to avoid two similar answer choices. This surprising finding may be linked to the

observation in the think-aloud protocol (see section 5.7.2) that L2 speakers had particular difficulties recalling and distinguishing paired technical terms like *immersion* and *submersion*, forgetting on average 9,5 words each as opposed to an average of 6,7 words for each L1 student. This finding could be followed up, for example to see if lexically similar and syntactically similar answer choices both follow the same pattern, with two similar answer choices causing higher average difficulty differential than multiple similar answer choices.

4.5.5 AOTA

According to suggested MCQ guidelines, AOTA should be avoided because AOTA is almost always correct when it is offered as a choice. As Haladyna (1994:78, 2004:117) points out, the use of this option may help test-wise test takers and reward students for partial information such as knowing that two of the three options offered in addition to AOTA are correct. The item quality measures for the ten AOTA questions are given in Table 4.10 below:

Table 4.10 Item quality measures for AOTA items

	2006		**2007**	
Item type	**AOTA**	**Overall**	**AOTA**	**Overall**
Total number	5	75	5	76
Average facility	80,6%	69,1%	67,4%	70,2%
Average discrimination	0,318	0,372	0,451	0,405
Average difficulty differential	13,6%	15,0%	18,6%	13,8%

These results are quite different for 2006 and 2007, with 2006 students finding the AOTA easy and not especially discriminating. Since the difficulty differential is below the average of 15,7%, these items do not disadvantage L2 speakers. In 2007, AOTA questions were harder than average at 67,4%, and displayed good discrimination and above average difficulty differential at 18,6%. These apparently contradictory figures can be explained by separating AOTA items where AOTA is the key from AOTA items where AOTA functions as a distractor. Like negative MCQs, it appears that AOTA items are not homogeneous in terms of item quality and patterns of

student performance. In 2006, AOTA was indeed the correct option in four of the five items (Q31, Q63, Q67, Q79) and functioned as a distractor in Q68. In 2007, AOTA was the correct option in two of the five items (Q15, Q69) and the wrong answer in the case of Q13, Q21 and Q73.

Insight into the way that AOTA functions when used as a distractor can be obtained by looking at the patterns of student responses. In the two examples below, the figures alongside each answer choice indicate first the percentage of students choosing that particular option, and second the average overall score of students choosing that option:

13. Kim's accent will signal that she is an L2 speaker of Kiswahili. The term accent refers to distinctive

0,9	41%	[1]	idiomatic phrases
0	0%	[2]	vocabulary items
1,8	40%	[3]	grammatical differences
72,7	58%	*[4]	pronunciation
24,8	45%	[5]	All of the above.

21. Which of the following statements is **false**?

46,2	61%	*[1]	L2 proficiency is directly related to the number of years of L2 study.
34,2	49%	[2]	Learners may benefit from a silent period where they listen to L2 but are not required to speak it.
1,8	42%	[3]	L2 teachers can help students by using gestures, pictures and contextual clues to clarify aspects of the language.
17	49%	[4]	All of the above.

In both these items a fairly high number of students (24,8% and 17% respectively) are choosing the incorrect AOTA answer choice. However these students are averaging below 50% for the test as a whole. This suggests that some weaker students are picking AOTA every time they see it. These items also illustrate the point that MCQs tend to have at most three functioning answer choices where a functioning answer choice is defined as one which is selected by at least 5% of students (see Haladyna 2004, Rodriguez 2005).

Item quality statistics for AOTA as key and AOTA as distractor (combined for 2006 and 2007) are compared in Table 4.11 below:

Table 4.11 Item quality measures for AOTA as key and AOTA as distractor

Item type	**AOTA as key**	**AOTA as distractor**	**Overall 2006**	**Overall 2007**
Total number	6	4	75	76
Average facility	83,6%	59,5%	69,1%	70,2%
Ave. discrimination	0,306	0,501	0,372	0,405
Average difficulty differential	11,8%	22,5%	15,0%	13,8%

Although the numbers of AOTA items are small, these figures still show clearly that AOTA items where AOTA is the right answer are easy at 83,6% facility, have below-average but still acceptable levels of discrimination and do not disadvantage L2 speakers. This seems to support Haladyna's contention that AOTA items reward test-wise test-takers who choose AOTA whenever it occurs and also rewards those with partial knowledge (1994:78, 2004:117). In contrast, AOTA items where AOTA functions as a distractor are more difficult than average at 59,5% facility, more discriminating than average at 0,501 and with above-average difficulty differential at 22,5%, are particularly difficult for L2 speakers. However the high discrimination suggests that they may be useful in distinguishing knee-jerk AOTA selectors from thinking students. In light of this data, my own view is that AOTA questions are fine in moderation. Test setters need to accept that they are easy questions if AOTA is the key and ensure that AOTA is not always the right answer.

4.5.6 NOTA

There is some debate about the value of NOTA as an answer choice in MCQs (see Knowles & Welch 1992, Dochy et al. 2001), but according to some guidelines, NOTA should be kept to a minimum because it is rarely a real option (Haladyna 1994:8, Haladyna 2004:116-117). Testwise students often therefore simply ignore NOTA as an option. Table 4.12 below gives the relevant item quality measures for the ten NOTA items:

Table 4.12 Item quality measures for NOTA items

	2006		2007	
Item type	**NOTA**	**Overall**	**NOTA**	**Overall**
Total number	7	75	3	76
Average facility	66,3%	69,1%	64,7%	70,2%
Ave. discrimination	0,387	0,372	0,220	0,405
Average difficulty differential	17,1%	15,0%	-4,3%	13,8%

NOTA questions are more difficult than average, but the discrimination and difficulty differential statistics are mixed, showing above average discrimination and difficulty differential in 2006 and below average discrimination and difficulty differential in 2007. In 2007 two of the three items had negative difficulty differential, indicating that L2 students did better than L1 students on those questions. Again it was hoped that a separation of NOTA as key from NOTA as distractor might yield interesting results which have not been highlighted much in the literature: NOTA functioned as a distractor in Q36, Q39, Q43 and Q48 2006 and Q44 and Q58 2007. NOTA was the correct option in Q22, Q23, Q38 2006 and Q31 2007 (see below):

31. Which of the following is the most appropriate description of Zulu?

0	0	[1]	a dead language
6,9	39%	[2]	a dying language
37,7	55%	[3]	a national language
54,7	56%	*[4]	None of the above

This example of NOTA as key shows that though more than half of the students (averaging 56%) selected NOTA as the right answer, a fair number (37,7%) of fairly good students (averaging 55%) selected option [3], possibly because they ignored NOTA as a genuine possibility here. Table 4.13 below provides the relevant data for items with NOTA as key versus NOTA as distractor for 2006 and 2007 combined:

Table 4.13 Item quality measures for NOTA as key and NOTA as distractor

Item type	NOTA as key	NOTA as distractor	Overall 2006	Overall 2007
Total number	4	6	75	76
Average facility	51,7%	75,2%	69,1%	70,2%
Average discrimination	0,341	0,334	0,372	0,405
Average difficulty differential	13,7%	8,6%	15,0%	13,8%

The figures here suggest that items with NOTA as key are very difficult items (facility 51,7%), with average discrimination and slightly below average difficulty differential. Items with NOTA as distractor are much easier than average at 75,2%, have below average discrimination at 0.334 and below average difficulty differential at 8,6%. In both cases the figures suggest that these are psychometrically acceptable items that do not discriminate against L2 speakers and that there is no statistical reason to avoid NOTA items.

The figures do suggest however that NOTA is being ignored as an option by some (but not the best) students, making NOTA as key items tricky items and NOTA as distractor fairly easy items with one-option less than face value. I suppose the test-setter's moral dilemma here is whether it is acceptable to have NOTA as key, knowing full well that many students are going to ignore it in their deliberations, or whether this is in fact a good way to separate thinking candidates from knee-jerk-reaction candidates.

4.5.7 Grammatically non-parallel options

Haladyna (2004:116) recommends that items should be homogeneous in grammatical structure to avoid accidentally giving away the right answer, usually the one that stands out as different from the others. I was also interested in finding out whether grammatically non-parallel items were more difficult for students or displayed a greater difficulty differential than average. It is at least possible that these items require students to read the stem several times in order to parse them correctly and that they may place a heavier burden on working memory than questions

where options are all phrases of the same type. An example of an item with grammatically non-parallel options is Q58 2006 which had noun phrases for all options except option [4], which was a verb phrase:

58. The girls in the example above are probably

*[1] balanced bilinguals
[2] semilinguals
[3] monolinguals
[4] using a divergent strategy to stress the differences between them.

There were 16 items of this type (Q28, Q58, Q71, Q74 and Q80 in 2006 and Q14, Q15, Q18, Q42, Q48, Q51, Q56, Q71, Q72, Q73 and Q80 in 2007). Item quality statistics are given in Table 4.14 below:

Table 4.14 Item quality measures for grammatically non-parallel options

	2006		**2007**	
Item type	**Non-parallel**	**Overall**	**Non-parallel**	**Overall**
Total number	5	75	11	76
Average facility	63%	69,1%	68,4%	70,2%
Average discrimination	0,393	0,372	0,398	0,405
Average difficulty differential	15,4%	15,0%	11,5%	13,8%

These figures indicate that grammatically non-parallel options are more difficult than average (6,1% more difficult than average in 2006 and 1,8% more difficult than average in 2007). Discrimination figures are approximately the same as the average values and the difficulty differential is also approximately average (0,4% above average in 2006 and 2,3% below average in 2007). Statistically, these items therefore do not seem to pose particular problems for students. Although the statistics do not indicate any cause for avoiding items with grammatically non-

parallel options, it is possible that students are having to reread the stem several times to recast the sentence and that these might be time-consuming items and hence more difficult. It was hoped that the think-aloud interviews might shed additional light on this issue.

4.5.8 Context-dependent text-comprehension items

As seen in Table 4.1, 57 items were comprehension items that required the student to refer to a passage (RTP) to find information or identify linguistic phenomena being described. Context-dependent questions tend to have the advantage of being higher-order application or analysis questions as they present novel scenarios for students to consider. The item quality statistics for these items are provided in Table 4.15 below:

Table 4.15 Item quality measures for context-dependent text-comprehension (RTP) items

	2006		2007	
Item type	**RTP**	**Overall**	**RTP**	**Overall**
Total number	27	75	30	76
Average facility	62,9%	69,1%	72,5%	70,2%
Average discrimination	0,392	0,372	0,363	0,405
Average difficulty differential	16,0%	15,0%	13,1%	13,8%

In 2006 the text-comprehension items were 6,2% more difficult than average at 62,9%, slightly more discriminating than average at 0,392 and had 1% higher than average difficulty differential at 16%. This accords to some extent with the findings in Case, Swanson and Becker (1996) where scenario-based items were found to be slightly more difficult but not more discriminating than average. However, this pattern was not repeated in 2007, where the text-comprehension items were 2,3% easier than average at 72,5%, slightly less discriminating than average and had below average difficulty differential. These conflicting results basically indicate no clear pattern as regards facility and discrimination of comprehension-type questions, but do show that this format does not discriminate unduly against L2 speakers. The average difficulty differential between L1 and L2 speakers differed by less than a percent in text-comprehension questions as

opposed to the average difficulty differential for the test as a whole. Overall, therefore, the context-dependent questions proved to be psychometrically useful questions that did not disadvantage L2 speakers.

4.5.9 Summary

By way of summary of sections 4.4 and 4.5, Table 4.16 below compares the effect of violating each of the abovementioned guidelines on the average facility and difficulty differential in an attempt to ascertain which of these have the greatest effect on question difficulty for all students or specifically for L2 speakers:

Table 4.16 Item quality measures for item-writing guidelines

Item type	**2006**		**2007**	
	Average DD (%)	**Average facility (%)**	**Average DD (%)**	**Average facility (%)**
All items	15,0	69,1	13,8	70,2
Readability level 16	15,2	70,5	15,5	73,1
AWL density over 0,12	15,3	66,1	15,1	67,8
AWL density over 0,15	16,1	64,1	18,5	68,5
Incomplete statement stem	15,1	69,1	13,5	72,4
Long item of 50 words or more	17,7	68,6	17,3	61,8
Negative stem	20,2	62,4	17,1	56,3
Negative option(s)	16,6	68,4	14,1	69,9
Double negative	19,4	61,0	18,5	56,0
2 similar answer choices	15,6	69,6	18,9	66,1
3, 4, or 5 similar answer choices	13,7	68,8	11,5	73,9
AOTA as key	11,8	83,6	Combined with 2006	
AOTA as distractor	22,5	59,5	Combined with 2006	

NOTA as key	13,7	51,7	Combined with 2006	
NOTA as distractor	8,6	75,2	Combined with 2006	
Context-dependent items requiring reference to passage	16,0	62,9	13,1	72,5
Grammatically non-parallel options	15,4	63,0	11,5	68,4

This table can be summarised in the form of a figure categorising item types into one of the four following categories: fair and not difficult, unfair and not difficult, difficult but fair, difficult and unfair. 'Difficult' in this case implies below average facility for all students, and 'fair' implies average or below average difficulty differential between L1 and L2 students. The overall averages for 'All items' in the first line of the table are therefore used as the benchmark for comparative purposes. Item types for which there were mixed results in 2006 and 2007, namely two similar answer choices and context-dependent items are not included in the figure:

	NOT DIFFICULT	**DIFFICULT**
FAIR	Incomplete statement stems Negative answer choices 3, 4, or 5 similar answer choices AOTA as key NOTA as distractor Items at Level 16 readability	Grammatically non-parallel answer choices
UNFAIR		High AWL density Long questions Double negatives Negative stems AOTA as distractor NOTA as key

Figure 4.3 Categories of MCQ items based on facility and difficulty differential statistics

In the ‘fair and not difficult’ category were incomplete statement stems, negatives in the answer choices, 3, 4, or 5 similar answer choices, AOTA as key, NOTA as distractor and items at the highest readability level, level 16. These were not more difficult than average, nor did they appear to discriminate unduly against L2 speakers by displaying a difficulty differential more than 1,7% above average. It would appear therefore that none of these types of items need to be avoided in the interests of fairness.

Items with grammatically non-parallel answer choices were ‘difficult but fair’ as their difficulty differential was approximately average. No action is required for these items although Haladyna (2004) prefers answer choices to be grammatically parallel so that none of the answer choices stands out.

Items with high AWL density, long items, double negatives, negative stems, items with AOTA as distractor and NOTA as key were ‘difficult and unfair’. All of these were more difficult than average and had above average difficulty differential. These items are therefore particularly difficult for L2 speakers, widening the gap between L2 and L1 performance on the test. (The low facility and high difficulty differential associated with items with two similar answer choices in 2007 was not replicated in the 2006 statistics and is therefore probably not a major problem for L2 speakers, although more research would help clarify this issue).

In terms of the magnitude of the difficulty, the most difficult items appeared to be NOTA as key, probably because students do not expect NOTA to be the key and ignore it in their deliberations. Items with AOTA as distractor were almost as difficult for the converse reason, namely that many students expected AOTA to be the key and picked it without paying much attention to the other answer choices. Also very difficult were double negatives and items with negative stems. The worst offenders in terms of widening the difficulty differential between L1 and L2 students were AOTA as distractor, double negatives, negative stems, long items and items with an AWL density above 0,15. Since AOTA as distractor is a structural rather than a linguistic cause of difficulty, I believe there is a case for retaining a small number of mixed AOTA items in a test (as argued in section 4.5.5 above). However, the above findings constitute strong reasons for avoiding double negatives and negative stems, keeping items brief (below 50 words) and

keeping academic words to a minimum, for example by revising the wording of questions containing more than six academic words as discussed in section 4.4.3 above.

While linguistic recommendations such as these are one side of the story, questions are also obviously easier or harder in terms of the cognitive demands they make on students. Section 4.6 below therefore reports on the findings as regards the cognitive complexity of the MCQs and the effect of cognitive complexity on item quality statistics.

4.6 Cognitive measures of predicted difficulty

Few if any efforts have been made in the literature to disentangle the interactional effects of two aspects of MCQ difficulty, namely readability and cognitive complexity. Studies have tended to focus on only one or other of these aspects in relation to question facility. For example, readability and facility were explored in Homan, Hewitt and Linder (1994), Hewitt and Homan (2004) and in Dempster and Reddy's (2007) investigation of item readability and primary school science achievement, while Seddon (1978) provides an overview of attempts to relate question facility to Bloom's cognitive levels. My attempt here was to provide an exploratory investigation of the effects of Bloom's six-level hierarchy not just on facility, but also on the discrimination and difficulty differential of questions (see section 4.6.3 below) and also to clarify the relationship between 'readable' and 'less readable' questions and 'difficult' and 'less difficult' questions in terms of the cognitive demands of the question (see section 4.6.4 below).

In order to achieve these aims, each item was evaluated by three independent raters in terms of the cognitive complexity required to answer it (cf Bowman & Peng 1972, Hancock 1994). The raters were given one or both question papers and asked simply to provide a rating from 1-6 for each MCQ by comparing the demands of the question with Bloom's taxonomy of cognitive levels where level 1 required recall, level 2 comprehension, level 3 application, level 4 analysis, level 5 synthesis and level 6 evaluation. The raters were myself, an external Independent Examinations Board (IEB) educational consultant, a Unisa Professor of Education for the 2006 paper, and a Linguistics colleague for the 2007 paper. All the raters were therefore experienced university teachers, had been trained as assessors and had at least a Master's degree.

4.6.1 Inter-rater reliability

Exact agreement between all three raters as to the Bloom category of an item was obtained on only 14 items (19%) in 2006 and 13 items in 2007 (17%). Inter-rater reliability of the Bloom ratings was calculated using Fleiss' free-marginal kappa, a statistical measure for assessing the reliability of agreement between a fixed number of raters tasked with assigning categorical ratings to a number of items (Fleiss 1971). The measure calculates the degree of agreement in classification over that which would be expected by chance, and is expressed as a number between 0 and 1. Fleiss' free-marginal kappa was calculated for the Bloom ratings of the three raters using an online kappa calculator (Randolph 2008), yielding a result of 0,20 for 2006 and 0,19 for 2007. There are no generally accepted guidelines for interpreting the value of kappa, as kappa will be higher when there are fewer categories, but based on their personal experience interpreting kappa values, Landis and Koch (1977) suggest that a kappa value between 0 and 0,2 indicates 'slight agreement', tending in this case towards the 'fair agreement' category (0,21 – 0,4).

This result was somewhat surprising in view of the claims in the literature that the Bloom levels of MCQs and other types of questions can be assessed with high inter-rater reliability by experienced teachers (Hampton 1993, Hancock 1994, Fellenz 2004). For example, Seddon (1978) reports on studies undertaken in the 1960s with 85% agreement between 3 raters.

While the levels of agreement between all three raters in my study were low (at 17% and 19%), agreement between two of the three raters was obtained on an additional 44 items in 2006 and 48 items in 2007, giving an acceptable 79% agreement overall. This level of agreement approaches that recorded by Hancock (1994), who observed that he and a colleague familiar with his course could achieve around 90% inter-rater agreement. Bloom (1956:51) explains that 'it is considered essential to know, or at least make some assumptions about, the learning situations which have preceded the test' and also for raters to attempt to solve the questions themselves. Lecturers on a particular course are therefore best positioned to rate the Bloom level of their questions. One of the expert raters in my study who was not familiar with the course made the following comments about the difficulty of the task:

> I found it extremely difficult to allocate the taxonomic levels as I am not familiar with the content. I have done it a great deal with the Life Sciences because I know what the students have to do to answer the question, but it was difficult for me as it was completely out of my field. You will probably get more reliable feedback from someone who is familiar with the Linguistics curriculum.
>
> (Personal communication 2010)

A more homogeneous group of raters, familiar with both Bloom's taxonomy and the course in question, would probably have improved the inter-rater reliability in this study. Other possible reasons for disagreement between raters are suggested by Fellenz (2004), who observed that it is harder to achieve consensus between raters at higher Bloom levels as higher-order questions require lower-level skills as well (Fellenz 2004:709) and a level 4 question is therefore also a level 1, 2, and 3 question. As Bloom (1956) acknowledges, it is also possible that questions could be answered in different ways by different students, e.g. by recalling an example in the study guide (level 1) or by applying one's knowledge to come to the same conclusion (level 3).

The items about which all three raters disagreed (17 in 2006 and 15 in 2007) were excluded from the analysis that follows. The analysis therefore relates to the 119 (79%) items about which there was agreement on the Bloom level between at least two of the three raters.

4.6.2 Distribution of the various Bloom levels

All the items in the study were classified into the first four of Bloom's levels by all raters, supporting the contention that multiple-choice tends to test the first four cognitive levels - knowledge, comprehension, application and analysis (Martinez 1999, Govender 2004, Stiggins 2005). There is disagreement in the literature about whether synthesis and evaluation can be tested using multiple-choice, with the dominant view being that MCQs can test all levels with the exception of synthesis (see for example Orey n.d.). While some experts (e.g. Govender 2004) believe that synthesis and evaluation can be tested to a limited extent using MCQs, it is difficult to imagine how synthesis could be tested using multiple-choice since it requires novel production of some type. Likewise, evaluation (in the sense in which Bloom used the term) would require unprompted critique and justification of one's views and would be difficult if not impossible to assess with MCQs. In my opinion, the limited synthesis and evaluation that MCQs are reputed to be able to test is in fact analysis or application. This view is supported by the fact that none of

the literature in which questions have been rated reports on MCQs at Bloom levels 5 or 6 and that none of the raters in this study assigned levels 5 and 6 to any item.

The distribution of items on which there was agreement between two or all three raters was as follows:

Bloom level 1 - recall

56 questions (47%) required recollection of a term, concept, definition, fact or generalisation.

Bloom level 2 - comprehension

35 questions (29%) required understanding or use of a term, concept, fact or generalisation

Bloom level 3 - application

19 questions (16%) involved applying one's knowledge to a new example using recalled criteria

Bloom level 4 - analysis

9 questions (8%) involved analysing elements or relationships between elements including logical deduction and identifying advantages or disadvantages.

This distribution reflects the expectation (put forward, for example by Martinez (1993) and Paxton (2000)), that MCQs tend to test predominantly lower-order (level 1 and 2) knowledge and skills. Three-quarters of the questions in my study fell into the two lowest categories.

4.6.3 Bloom levels and item quality statistics

Statistics on average facility, discrimination and difficulty differential for the four cognitive levels are given in Table 4.17 below.

Table 4.17 Item quality measures for Bloom's levels (1 knowledge, 2 comprehension, 3 application and 4 analysis)

	2006 and 2007			
Bloom level	**1**	**2**	**3**	**4**
Total number on which category agreement was reached	55	35	19	9
Average facility	71,9%	72,2%	66,0%	60,1%
Average discrimination	0,384	0,394	0,366	0,437
Average difficulty differential	14,1%	15,2%	14,5%	20,1%

What this table indicates is that overall, question difficulty at Bloom levels 1 or 2 is very similar at around 72%, but increases considerably to 66% for application questions and even more to 60% for analysis questions. Discrimination of questions at Bloom levels 1-3 is fairly similar but increases at level 4 to 0,437, while the difficulty differential increases from around 14 or 15% at levels 1-3 to around 20% at level 4. In sum then, levels 1 and 2 appear fairly similar as regards item quality and the demands made on L1 and L2 students, but there appears to be a drop in facility for all students at level 3, and again at level 4, as well as a big increase in discrimination and difficulty differential at level 4. Analysis questions in this study therefore proved much more difficult for L2 students than for L1 students and increased the gap between these two groups.

A similar finding with respect to question facility and discrimination of MCQs at various Bloom levels was observed by Benvenuti (2010), who showed that comprehension, application and analysis MCQs were not only more difficult but also more discriminating than recall questions. Considerable research efforts in the 1960s (see Seddon 1978 for a review) were invested in attempts to adduce statistical evidence to validate the hierarchical nature of Bloom's taxonomy, with limited success. Seddon (1978) claims that there is no direct relationship between facility and question complexity, that average facilities cannot be used to validate Bloom's taxonomy, and that correlation coefficients and other statistical procedures are required. However, the evidence presented in Table 4.17 above suggests that knowledge and comprehension questions are easier than application questions which in turn are easier than analysis questions. While I agree with Seddon's view that there is no direct relationship between facility and the cognitive

demands of an MCQ, as other aspects of difficulty such as poor readability or ambiguity may contribute to low facility, the data presented above do provide support for Bloom's taxonomy by showing that Bloom levels relate not only to facility but also to discrimination and to the score gap between L1 and L2 speakers. The ordering of categories 1-4 is as predicted by Bloom's taxonomy and therefore provides empirical evidence in support of the model.

The following section attempts to map the combined effects of Bloom levels and question readability on student performance. Some of the statistics relating to AWL density and Dale-Chall readability are therefore reconsidered here for questions at Bloom levels 1-4.

4.6.4 Combined effects of Bloom levels and readability

Since the AWL density of items averaged 0,075 (see Table 4.3 above), items with a density of 0,12 and above were classified as having high AWL density. Low AWL density items for the purposes of the following analysis had a density of 0,11 and below, with most items falling into this category. Results are given in Table 4.18 below:

Table 4.18 Combined effects of AWL density and Bloom's cognitive level (1 knowledge, 2 comprehension, 3 application and 4 analysis)

	Low AWL density (0,11 and below)				**High AWL density (0,12 and above)**			
Bloom level	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Number of items	42	27	12	9	12	8	7	0
Average facility	71,9%	74,3%	65,2%	60,1%	71,9%	65,4%	67,6%	
Average discrimination	0,379	0,375	0,327	0,437	0,403	0,456	0,432	
Ave difficulty differential	13,7%	15,5%	13,5%	20,1%	15,7%	14%	16,1%	

This table provides evidence that both Bloom level and the density of academic words play a role in making questions more difficult and more discriminating. The hardest questions, with an average facility of less than 68%, are analysis and application questions, as predicted by Bloom's model. Comprehension questions proved much harder when coupled with high AWL density, with an average facility of 65,4% in comparison to the low AWL density comprehension questions at 74,3%. However, the facility of recall questions was unaffected by AWL density.

The most discriminating questions (over 0,400) are those at Bloom level 4 and at all the Bloom levels where there is simultaneously a high density of academic words. This provides further evidence in support of Pretorius' (2000) and Cooper's (1995) findings that students' vocabulary levels are an important indicator of academic success. Those high discrimination questions that best separate weaker from stronger students in Linguistics are questions which contain a large proportion of academic words. Conversely, those students who can understand and correctly answer questions with a high density of academic words are the students who are doing better in the test as a whole.

The highest difficulty differential (20,1%) is at Bloom level 4 and to a lesser extent at Bloom level 3, where there is simultaneously a high AWL density (16,1%). (There were no questions with high AWL density at Bloom level 4.) Comparing questions at the same Bloom level was also interesting, with recall questions proving equally difficult, but more discriminating and 2% more difficult for L2 students when the AWL density was high. Comprehension questions proved almost 9% more difficult on average when coupled with a high density of academic words, and also more discriminating. However the difficulty differential was slightly lower (14% as opposed to 15,5%). Application questions were 2,4% easier when combined with a high AWL density but this surprising finding could be due to the small sample size of only 7 questions in this category. Again, application questions with a high AWL density were more discriminating and had a 2,6% higher difficulty differential than application questions with fewer academic words.

These results suggest that both cognitive levels and vocabulary levels play a part in making questions more difficult, and that when both of these demand more of readers, the number of

right answers will decrease. Questions with dense academic vocabulary *and* high cognitive demand are placing a particular burden on L2 as opposed to L1 speakers, increasing the gap in student scores between these two groups. In order to triangulate this finding, the question of the interaction between readability and cognitive demand was approached in a slightly different way by investigating the combined effects of Dale-Chall readability score and Bloom level, where the readability score incorporates both vocabulary level and average sentence length. Table 4.19 below reflects the findings.

Table 4.19 Combined effects of Dale-Chall readability level and Bloom's cognitive levels 1-4 (1 knowledge, 2 comprehension, 3 application and 4 analysis)

	Low readability score (up to 11-12)				**High readability score (13-16)**			
Bloom level	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
Number of items	29	23	6	6	23	12	13	3
Average facility	67,1	73,6	52,8	62,6	77,5	69,6	72,2	55,0
Average discrimination	0,373	0,396	0,341	0,417	0,389	0,390	0,378	0,479
Ave difficulty differential	14,6%	15,0%	7,7%	14,3%	13,5%	15,4%	17,6%	31,7%

The results here are rather mixed, but the difficulty differential in the last row is interesting, indicating approximately average or below average difficulty differential for all the more readable questions. For the less readable questions, the difficulty differential escalated rapidly from a below average 13,5% at Bloom level 1, to above average 15,4% at level 2, 17,6% at level 3 and a huge 31,7% at level 4. Again this points to the double burden of cognitive complexity and low readability combining to heavily disadvantage L2 speakers.

Because the samples above were small, with just three to six items in some categories, the statistics for Bloom levels 1-2 and 3-4 were conflated as shown in Table 4.20 below:

Table 4.20 Combined effects of Dale-Chall readability level and Bloom's cognitive levels 1-2 (knowledge and comprehension) and 3-4 (application and analysis)

	Low readability score (up to 11-12)		**High readability score (13-16)**	
Bloom level	**1-2**	**3-4**	**1-2**	**3-4**
Number of items	52	12	35	16
Average facility	70.0%	57,8%	74,8%	68,9%
Average discrimination	0,384	0,379	0,389	0,397
Ave difficulty differential	14,8%	11%	14,1%	20,3%

This table shows a similar pattern of results to Table 4.19, suggesting that a low readability score coupled with a low Bloom level leads to 'easy and fair' questions, while a low readability score with a high Bloom level leads to 'difficult but fair' questions. The cognitive level of the question is thus the most critical factor affecting question facility. Low readability questions at Bloom levels 3-4 were not harder for L2 students than for L1 students, displaying a below average difficulty differential of 11%. More readable questions (up to level 12) were therefore fair to all students.

The results also suggest that a high readability score has little effect on lower order questions. These remain 'easy and fair' with an average facility of 74,8% and approximately average difficulty differential (14,1%). However, the final column indicates that a high readability score coupled with the cognitive demands of a Bloom level 3-4 question results in reduced facility of 68,9% on average (more wrong answers from all students) and a very high difficulty differential of 20,3%. These are therefore both 'difficult and unfair' in that making higher demands on both students' cognitive problem-solving abilities and their comprehension abilities (with respect to the sentence length and vocabulary level of items) appears to be negatively affecting the chances of a correct answer for L2 students even more than L1 students. The findings above are summarised graphically in Figure 4.4 below:

	READABLE (up to Dale-Chall level 12)	**LESS READABLE (Dale Chall levels 13-16)**
BLOOM LEVELS 1-2	Easy and fair	Easy and fair
BLOOM LEVELS 3-4	Difficult but fair	Difficult and unfair

Figure 4.4 Interactional effects of readability and cognitive complexity on question facility and difficulty differential

The general conclusions to be drawn from the findings reported in this section are that the cognitive level of an MCQ question is not always easy to pinpoint, and that MCQs, while they are certainly able to test application and analysis, in this case tested predominantly recall and comprehension. The Bloom levels were shown to have an effect not only on facility, but also on discrimination and the score gap between L1 and L2 speakers, with analysis questions proving much more difficult for L2 students than for L1 students. The cognitive level of the question was shown to be more important than readability in affecting question facility. However, an interaction effect was evident whereby a high readability score had little effect on lower order questions, but a noticeable effect on higher-order questions, decreasing the number of right answers and increasing the gap between L1 and L2 students. This ties in with Kemper's (1988:47) finding that readability affects text comprehension only when texts are already difficult to understand because they require high levels of inferencing. In the interests of fairness, attention therefore needs to be paid to ensuring that higher-order questions are readable by the weaker readers in the group, while for lower-order questions, student results are less sensitive to difficult wording and long sentences.

Section 4.7 below offers a reconsideration of the most difficult MCQs in the light of the quantitative measures and guideline contraventions associated with each item as well as the Bloom level of the items.

4.7 Linguistic characteristics of the most difficult questions

This section revisits the low facility and high difficulty differential items discussed in sections 4.3.2 and 4.3.3 above and attempts to identify linguistic reasons that contribute to their difficulty. Low facility (the most difficult) questions are discussed in section 4.7.1 and high difficulty differential items (those with the biggest score gap between L1 and L2 students) in section 4.7.2 below.

4.7.1 Low facility questions

The most difficult questions in 2006 (those below 50% facility) ranged from 49% (Q72) to 33% (Q70). In 2007, questions below 50% ranged from 49% (Q76) to 20% (Q78). Only four of these 17 most difficult items were analysis and application questions (Bloom levels 3-4). However, all of them ignored one or more of the item-writing guidelines (incomplete statements were not counted as guideline contraventions since the item analyses for incomplete statement stems and question stems were shown to be very similar). Six of the items violated two of the MCQ guidelines, three violated three guidelines and Q5 2006 (see below) violated four guidelines, containing 55 words, a double negative (*cannot* in the stem and *not* in option [3], a reading level of 16 and similar answer choices [2] and [3]:

5. The data above indicates that children cannot accurately reproduce sentences that are above their current level of language development. This kind of evidence suggests that

   [1] children are born with innate knowledge of the vocabulary of their language
   [2] children learn language by imitation
   *[3] children do not learn language simply by imitation
   [4] children imitate adults' language errors.

Among the 17 most difficult items there were six long items (50 words or more) and eight negative items, of which four were double negatives and three were negative stems. Five items had readability levels above Grade 13 and four had an AWL density of 0,12 or more, which is also an indication of low readability. Nine items had two or more similar answer choices. Eight of the most difficult items, including Q5 above, were context-dependent questions that required students to refer back to the passage before answering. Although this is not a violation of an

item-writing guideline, paging back or rereading sections of the text before answering the item does add another layer of complication.

The characteristics listed above suggest that a piling up of factors that make comprehension more difficult (reference back to a text, similar answer choices to disentangle, negatives, high readability scores, a high density of academic words and general wordiness) leads to very difficult questions that less than half the class (in some cases less than a third of the class) will be able to answer. Multiple guideline violations should therefore be avoided when setting MCQs.

4.7.2 High difficulty differential questions

The 25 questions with the highest difficulty differential (25% and over) ranged from 26% (Q23) to 54% (Q53) in 2006, and from 25% (Q62) to 38% (Q76) in 2007. Although it is difficult to identify clear patterns, 22 of the items contravened one and usually more than one of the item-writing guidelines (excluding the guideline relating to incomplete statements which was not counted). Of the 25 questions, six contravene two guidelines simultaneously, six contravene three guidelines and two contravene four guidelines. These multiple violations make it difficult to pinpoint the exact nature of the problem, but do combine to cause particular difficulty for second language speakers. For example, the highest difficulty differential was in Q53 2006 (see below), which was answered correctly by 87% of L1 students and only 38% of L2 students. This question had 50 words, a Dale-Chall reading level of 16 and a Bloom level of 4 (analysis). This is therefore one of those questions alluded to in section 4.6.4 above, where a high cognitive level and a high readability score combine to make the question overwhelming for L2 students.

53. One advantage of choosing Xhosa as South Africa's only official language would be that

    [1] Xhosa is an international language
    [2] it could lead to tension between ethnic groups
    [3] it would benefit mother-tongue Nguni speakers at the expense of other language groups
    *[4] it would be cheaper than an 11-language policy.

Of the 25 questions with high difficulty differential, 11 had a readability score of 13-15 or higher and therefore a large number of unfamiliar words and/or long sentences, 11 contained at least

two very similar answer choices and nine contained negatives. Six items were 50 words long or longer, six had an AWL density over 0,12 and four contained answer choices which were not grammatically parallel. It is also noteworthy that of the two AOTA and two NOTA questions with high difficulty differential, three are of the ‘unexpected’ type (one with AOTA as distractor and two with NOTA as the key). Again these findings suggest that multiple violations of MCQ guidelines should be avoided to prevent L2 students from experiencing unnecessary comprehension difficulties.

4.8 Conclusion

The present study set out to identify which kinds of MCQs were ‘difficult’, the contribution that linguistic factors made to these difficulties and to compare the performance of L1 and L2 speakers on two 80-item MCQ examinations. Facility, discrimination and difficulty differential data were mapped against the linguistic characteristics as well as the readability and cognitive levels of various subgroups of items.

There was a trend for the more difficult questions to display a greater score difference between L1 and L2 speakers than the easier questions. The average difficulty differential was 14,4%, but approximately a sixth of the test items had a difficulty differential of 25% or more and thus favoured L1 English speakers quite noticeably over L2 speakers. One of the aims of the study was to investigate which linguistic characteristics of questions were associated with an increased difficulty differential and therefore unfair to L2 students.

The first guideline that was investigated was the guideline regarding maximising the readability of questions. While question difficulty did not correlate significantly with either the readability, the number of unfamiliar words or the number or density of AWL words, items with high AWL density (above 0,12 and particularly above 0,15) were shown to be more difficult, more discriminating and with a larger difficulty differential than average. AWL density is therefore a possible measure that can be used prior to a test to identify questions that could usefully be simplified to reduce the gap between L1 and L2 scores. There was also some indication that more readable questions (Dale-Chall levels 7-8 and below) lessened the gap between the scores

of L1 and L2 students. Aiming for maximum readability is therefore beneficial to all but particularly to L2 students.

In addition to readability, seven other guidelines relating to the language of MCQs were tested empirically by comparing the three item quality measures for items that disregarded these guidelines with the average values. Item-writing guidelines that were confirmed by the data include avoiding items over 50 words long and avoiding negative stems. Long items had a high difficulty differential, suggesting that long items are causing more problems for L2 students than L1 students, probably by overloading working memory. The guideline on avoiding negative questions was also supported (although these do have the advantage of high discrimination). Negative questions were harder than average and had a larger gap between L1 and L2 scores. On closer investigation it appeared that negative stems and double negatives were more problematic than negative answer choices and should be eliminated as far as possible during item-writing and subsequent editorial review of questions. Adherence to these two guidelines would therefore be particularly beneficial for L2 students.

Item-writing guidelines that were *not* confirmed by the data included avoiding incomplete statements in favour of question stems, avoiding similar answer choices, keeping answer choices grammatically parallel. These types of items, together with context-dependent items that required reference back to a text, were not more difficult than average, nor did they appear to discriminate unduly against L2 speakers. These types of items therefore do not need to be avoided in the interests of fairness.

In terms of the magnitude of the difficulty caused by contraventions of the various item-writing guidelines, the most difficult items were those with NOTA as key or AOTA as distractor, followed by double negatives and items with negative stems. The worst offenders in terms of widening the difficulty differential between L1 and L2 students were items with AOTA as distractor, double negatives, negative stems and more than 50 words. The statistics showed very clearly that AOTA and NOTA items function differently depending on whether AOTA/NOTA is the key or a distractor. Since AOTA as distractor and NOTA as key are structural rather than linguistic causes of difficulty, and are associated with high discrimination, I believe there is a

case for retaining a small number of mixed AOTA and NOTA items in a test. However, the above findings constitute strong reasons for avoiding double negatives and negative stems and keeping items brief (below 50 words).

All the items in the study were classified into the first four of Bloom's levels by all four raters, with most items at level 1 and progressively smaller proportions of items at levels 2, 3 and 4. The 79% of items on which there was agreement between two of the three raters on each examination were analysed further. Questions at Bloom levels 1 and 2 were fairly similar as regards item quality and the demands made on L1 and L2 students, but there was a drop in facility for all students at level 3, and again at level 4, as well as a big increase in discrimination and difficulty differential at level 4. Analysis questions are therefore proving much more difficult for L2 students than for L1 students and increasing the gap between these two groups.

Data that reflected the combined effects of readability and Bloom level showed that questions with dense academic vocabulary *and* high cognitive demand are placing a particular burden on L2 as opposed to L1 speakers, increasing the gap in student scores between these two groups. Questions with a large proportion of academic words were best able to separate weaker from stronger students in the discipline, reinforcing the importance of high levels of vocabulary as an indicator of academic success. A high Dale-Chall readability score was shown to have little effect on lower order questions but a high readability score coupled with the cognitive demands of a Bloom level 3-4 question resulted in more wrong answers from all students and a very high difficulty differential of 20,3%. Making higher demands on both students' cognitive problem-solving abilities and their comprehension abilities therefore appeared to negatively affect the chances of a correct answer for L2 students even more than L1 students.

A reconsideration of the 17 most difficult questions showed that all of these violated one or more of the item-writing guidelines (excluding stem vs incomplete statement which did not appear to have any effect). The data suggests that a piling up of factors that make comprehension more difficult (reference back to a text, similar answer choices to disentangle, negatives, high readability scores and general wordiness) leads to very difficult questions that less than half the

class will be able to answer. Multiple guideline violations are therefore to be avoided in setting MCQs.

Although it was difficult to identify clear patterns, most of the 25 items with high difficulty differential (25% and over) also contravened one or more of the item-writing guidelines. Suggestions for reducing the difficulty differential include keeping items brief (less than 50 words long) and adhering to the item-writing guidelines regarding readability and negative stems. In particular, results suggest that multiple violations of MCQ guidelines should be avoided to prevent L2 students from experiencing unnecessary comprehension difficulties.

Chapter 5 will report on the findings of the think-aloud interviews with 13 individual students and will attempt to focus in more detail on the reasons for question difficulty and wrong answers and on the role that language plays in this regard. This will offer a complementary perspective to the findings presented in this chapter.

Chapter 5
Qualitative results

5.1 Introduction

This chapter reports the results of the qualitative interviews with 13 students who had just completed the LIN103Y course. Obviously 13 students is a small sample, and the results reported below should be viewed only as case studies. However, the opportunity for students to reflect on MCQ tests in their own words offers an important additional perspective on both the validity of the test and on the types of problems that students experience. Cohen (2007) and Fairbairn and Fox (2009:17) stress the growing awareness of the importance of test-taker feedback for test validation and for a more nuanced understanding of how tests are functioning than statistical analysis can provide.

Section 5.2 below provides a brief description of the method adopted and provides some discussion of the validity of this method and the compilation of the sample. Profiles of the 13 students and their approach to answering MCQs are provided in section 5.3. The results of the Likert-scale questions on student perceptions of multiple-choice assessment are then reported in section 5.4. Section 5.5 describes the students' MCQ-answering strategies, and section 5.6 compares reported difficulties and observed difficulties in the think-aloud interviews. An attempt to classify and clarify the nature of the difficulties experienced by students is made in section 5.7 to 5.9, with section 5.7 focusing on difficulties related to readability, section 5.8 on difficulties related to other guideline violations and section 5.9 on other types of difficulties. Section 5.10 provides a conclusion.

5.2 Method

As described in section 3.6.2.1, test-taker feedback can include either general examinee feedback about the face validity and fairness of the test or test method (e.g. Nield & Wintre 1986, Roberts 1993, Paxton 2000, Duffield & Spencer 2002, Struyven, Dochy &

Janssens 2005), or verbal reports on specific test items (cf Norris 1990, Farr, Pritchard & Smitten 1990, Cohen 2007). Both of these types of feedback were elicited in the qualitative interviews and are reported on in sections 5.4 and 5.5-5.9 respectively. Support for the think-aloud interview methodology as a way of probing student's reasoning and of validating tests comes from Haladyna, Downing and Rodriguez (2002:329), Messick (1989) and Norris (1990:41) among others. Some of the issues that think-aloud interviews can shed light on in the case of multiple-choice assessment include students' answering strategy and difficulties with the instructions, design, style, organisation, coherence and content (Martinez 1999:211, DuBay 2004:57).

As described in section 3.6.2.2, the interviews followed a set format, namely 10-15 minutes' elicitation of information on the students' linguistic and educational background, followed by oral completion and discussion of three Likert-scale questions – What is your general opinion of MCQ as an assessment method? How difficult do you find MCQ as an assessment method? and What was your opinion about this particular question paper that you wrote a few weeks ago? This was followed by a scripted explanation by the researcher of the purpose of the study and of what the students were expected to do (see 3.6.2.2).

A consent form was then signed and a think-aloud protocol of the examination was undertaken using a copy of the examination paper they had written 2-5 weeks before. I took notes during the interviews of their comments and of my own observations regarding long pauses, reference back to text passages, rereading of questions and other non-verbal cues. The think-aloud portion of the interview took between 30 minutes and two hours for each student.

5.2.1 Compilation of the sample

In terms of the compilation of the sample of students interviewed, three of the 13 were L1 English speakers and the others spoke English as a second language. This is approximately representative of the LIN103Y student body as a whole, which self-reported 33% English home language in 2006 and 26% in 2007. Five of the students were

male and eight were female, which also mirrors the predominantly female student body in first-year Linguistics at Unisa. All students in the think-aloud interviews in the current study experienced the same test-taking problem, namely 2-5 weeks of elapsed time between the actual examination they had studied for and the think-aloud rerun of the examination. It should also be noted that the students in the study were also probably more interested in the course than average as they all voluntarily accepted my open offer to participate in the research.

The students' official results for the LIN103Y course had not been finalised at the time of the interviews, but as it happened, 12 of the 13 students passed the examination (and the course as a whole), with average to good marks ranging from 60% to 92%. One student failed the MCQ examination with a mark of 48% (see Table 5.1 below.) Although the students did well, they can be considered a fairly representative sample in that their average mark was 73% while the overall average mark of the entire LIN103Y class was 68%. The benefit, however, of this slightly above average sample of students is that good students tend to be more metacognitively aware and able to identify and articulate their concerns more easily than weaker students can (Pretorius 2005:41, Norris 1990:211). Students' awareness of their own comprehension difficulties was revealed, for example by the fact 16 of the 25 questions with the highest difficulty differential (over 25%) were identified by the students in the think-aloud protocols as being 'tricky'. The students' realisation and detailed explanation of their own answering strategies and points of confusion was a feature of most of the interviews, with only the two weakest students in terms of MCQ examination scores offering little in the way of verbal explanations of their reasoning. These two students often offered an answer choice without a reason, and struggled to explain their reasoning process when prompted.

5.3 Student profiles

Brief descriptions are provided below of each of the students' backgrounds and their MCQ-answering strategies. The 3 mother-tongue English interviewees are discussed in 5.3.1 and the 10 L2 students in section 5.3.2 below. The score achieved for the think-aloud assessment is given in bold underneath each student profile. Questions that each

student identified as ‘difficult’ in some way are also listed and will be explored in more detail in sections 5.7-5.9 below.

5.3.1 English L1 students interviewed

NG (2006) was a mother-tongue English speaker who also speaks Italian at home. He was studying a BA part-time and already had a degree in film studies. His overall initial strategy in answering MCQs was to read all texts straight through as he came across them and underline important information. He then read the questions in order and referred back to the text if necessary. He generally went through each option, crossed out the implausible ones, marked possible answers with an arrow, then went back to evaluate the possible ones, i.e. using successive processes of elimination. He always writes the correct answer in the margin of his question paper and later double-checks these against the completed mark reading sheet. He can be considered to be a ‘test-wise’ student and had given a lot of thought to multiple-choice answering strategy, e.g. *‘You need to be careful with negative questions, but the highlighted **not** is fair’.*

87% - Reportedly difficult questions (2006) Q4 Q5 Q9 Q17 Q18 Q26 Q28 Q35 Q44 Q68 Q72

DS (2006) was an 84-year-old mother-tongue English speaker who had been studying towards a BA for the last 10 years. He used to speak some Afrikaans for work purposes but rarely does so any more now that he is retired. He tries to select Unisa courses that are assessed using MCQs because his memory of terms and facts is no longer good and he finds written examinations significantly more difficult than MCQs. His overall MCQ strategy was to read the text, then read each question. He tends to read all the options quickly, then rereads the stem before deciding. He makes no marks on the question paper. He skipped the ones he didn’t know and came back to these later. He checks the time after each ten questions and calculates beforehand how much time he has per ten questions. He referred back to the passage more than strictly necessary. The video camera didn’t work in this interview and so my notes are the primary data source here. He verbalised very little during the think-aloud protocol, and so his rationale was often

unclear to me as the researcher. Even when prompted to explain his answer choices, his answers tended to be along the lines of *'Because it's what I can remember'*.

61% - Reportedly difficult questions (2006) Q4 Q29 Q41 Q53

AJ (2007) was a mother-tongue South African Indian English speaker who attended English-medium schools. She understood Gujerati and Afrikaans and was studying a BA Communication Science full time. She underlined key words on the paper, then answered the questions quickly and briefly, crossing out the inappropriate answer choices systematically, e.g. the true ones in a false question or vice versa. She sometimes stopped reading when she found what she thought was the right answer. There was a very apparent time difference between the recall questions which she answered in 10-20 seconds and the application questions. She said she always checks her answers after completing the paper.

76% - Reportedly difficult questions (2007) TEXT E Murchison Q25 Q28

5.3.2 L2 students interviewed

CL (2006) was a mother-tongue Zulu speaker who was taught in English from Grade 8. She works in banking administration and also understands Sesotho and Afrikaans. She was studying a BA after having completed a Diploma in Language Practice. CL's overall initial MCQ strategy, according to her own report, is to read the entire question paper for approximately half an hour before beginning to answer. She then reads each passage once or twice before attempting the questions. She reads each question and looks back at the passage to search for the correct answer if necessary. The interview was taped in unfavourable conditions with a radio blaring so the researcher's notes were used as the main data source. CL had an additional test-taking difficulty in the form of severe toothache, which impaired her concentration during the think-aloud interview, especially towards the end. She took a long time to answer negative questions.

71% - Reportedly difficult questions (2006) Q11 Q26 Q29 Q39 Q55 Q61 Q66 Q69 Q73

DD (2006) was a mother-tongue Tswana speaker, who was taught in English from Grade 5. She also spoke Sesotho, Zulu and understands Afrikaans. She worked as a Personal Assistant at a parastatal and was in her first year of a BA Creative Writing, studying part-time. Her overall MCQ strategy was to read the text, reread it if necessary, then read each question once or twice before answering. She referred back to the reading passages where necessary. She said she normally struggles to finish MCQs in the allotted time, although this was not the case for the test in the current study.

61% - Reportedly difficult questions (2006) Q10 Q15 TEXT C Q16 Q17 Q18 Q21 Q22 Q26 Q28 Q32 Q33 Q45 Q50 Q51 Q55 Q59 Q60 Q62 Q68

YP (Semester 1 2007) was a mother-tongue Afrikaans speaker, whose father was English and who spoke English to them at home. She was therefore a balanced bilingual. She went to an Afrikaans school for the first 11 years, switching to an English-medium school for her matric year. Her overall MCQ strategy is to read the instructions first, then mark the end of each item-set with a horizontal line before reading the text and related questions. She believed that *'narrowing your questions down into sections … helps a lot. Otherwise your mind's gonna start jumping to the other questions and you should be focused on what you're doing.'* She tended to follow the lines of the reading passage with a pen to aid concentration. She generally read the options twice *'because sometimes they are very similar'*. She circled the numbers of the questions she was not sure about and went back to these when she reached the end of each section. Other than these circles and circles for correct options, she made no marks on the paper. She found paging back to refer to the passage on a previous passage irritating and reported finding it difficult to locate the question again afterwards.

91% - Reportedly difficult questions (S1 2007) Q3 Q6 Q10 Q17 Q18 Q21 Q24 Q27 Q31 Q32 Q41 Q50 Q57 Q60 Q74 Q76

ZL (2007) was a mother-tongue Afrikaans speaker who learned English at age 3 and was a balanced bilingual with L1 English accent and fluency. She attended a dual-medium school with both English and Afrikaans as languages of instruction. She was studying a BA Languages and Literature and already had a BA (Music). She worked as an editor for

the Department of Education, checking and translating examination papers. When answering multiple-choice questions she reads the stem quickly and then goes systematically through each option, marking each one T or F as she goes. She was a test-wise student who marked the paper a lot, circling important terms in the case studies or stems and usually writing the number of her chosen answer in the lefthand margin. She was particularly wary of negative stems, circling them every time and reminding herself to *'concentrate on "false"'*. She answered very quickly (and the two-hour exam reportedly took her 20 minutes), commenting that '*I never leave any out and I never go back cos else I get confused*'.

93% - Reportedly difficult questions (2007) Q2 Q12 Q16 Q22 Q25 Q29 Q41 Q46

ET (2007) was a Tsonga mother-tongue speaker who attended a school where Tsonga was the medium of instruction, with a switch to English in Grade 5. He was highly multilingual, being fluent in Sesotho, Xhosa, Zulu, Shona, Sepedi, Tswana, Afrikaans and German (he studied in Germany for a year). He worked for the government on a government publication. He was noticeably tired at the interview after a long day. He made few markings on the paper, except circling the chosen answer and occasionally underlining an important word in the stem. He spent a lot of time justifying his choice of answer in detail. He clearly struggled when questions were longer.

71% - Reportedly difficult questions (2007) Q5 Q17 Q25 Q26 Q46 Q47 Q55 Q56 Q62 Q68 Q75 Q76

AZ (2007) was studying a BA and was a Xhosa speaker, also fluent in Zulu, English, Afrikaans, Sesotho and Siswati. Xhosa was the language of instruction till Grade 7, after which he attended an English-medium high school. He worked for Unisa as a data capturer. The interview took place over two days due to his work commitments. Because he worked part-time as a peer tutor in the Unisa Reading and Writing Centre he had thought consciously about MCQ-answering strategy and could be considered a test-wise student: *'I always think of the book when I approach the MCQs because they are tricky. You'll find maybe this one can be this one, maybe you eliminate this one and this one and get to two, then you think about the book, what is in the study guide? Then I choose the*

one that I think I've seen in the study guide.' He underlined or circled keywords and negative words like *false, impossible* etc. and crossed out the numbers of the options that he had eliminated. If he knew the answer he didn't read all the options: *'Questions like these you don't have to confuse yourself with the options. You know what they are asking. Go straight to it.'*

83% - Reportedly difficult questions (2007) Q9 Q14 Q18 Q23 Q25 Q28 Q31 Q34 Q46 Q72 Q73

MD (2007) was a BA (Court interpreting) student whose L1 was Lingala (the national language of the Democratic Republic of the Congo). He was also fluent in French, Otetela (a dialect of Lingala), English and Malagasy. French was the language of instruction at his school in the DRC, but he was taught in English after immigrating to South Africa in Grade 8. He had a French accent when speaking English. He read very slowly and thoughtfully, following with a pen and underlining or circling important phrases as well as those he didn't understand. He misread several words, including *primarily* as *primary*, *colloquial* as *colonial*, *production* as *pronunciation*, *intelligible* as *intangible*, etc. He worked very systematically through each of the options in each question, considering all of them before making his choice. But he marked each one with little lines or dots instead of T or F as he went along, so he got confused sometimes about which ones were true.

66% - Reportedly difficult questions (2007) Q2 Q10 Q13 Q17 Q33 Q35 Q44 Q51 Q56 Q64 Q69 Q70 Q71 Q72 Q73 Q74 Q77 Q78

NN (2007) was a mother-tongue Zulu speaker with a South African English accent. She learnt through the medium of English from Grade 4. She was a government employee studying a BA (Communication Science) part-time. She underlined a few key terms but made few marks on the question paper, except sometimes a tick on the right answer. She read options quickly and silently or in a quick mumble then answered quite quickly. She seldom revisited her initial answer.

67% - Reportedly difficult questions (2007) (not including forgotten term) Q5 Q9 Q31 Q73 Q75

TH (2007) was a Venda speaker who worked for a media liaison and PR company. She was highly multilingual, speaking 9 of South Africa's 11 official languages. She was an experienced student, having already completed a BCom (Hons) degree before starting her BA. She often went directly to the answer without reading through all the options. She made almost no markings on the paper except a little line near the selected option, circling one or two key phrases and making one or two notes. However, for the 'Which of the following is false?' questions she tended to be very systematic, marking each option with a tick or cross as she went through it. She ticked each section of the text as she completed it. The interview took place after work and she was noticeably tired towards the end. The tape was of very bad quality due to background noise and a humming airconditioner.

86% - Reportedly difficult questions (2007) TEXT A Susan Q5 Q6 Q22 Q23 Q35

SD (2007) was a Sesotho L1 speaker who worked for a cell phone company. She spoke six South African languages and studied through the medium of English with Sesotho as a school subject. She really disliked multiple-choice, finding she had to cram for the exam and often feeling pressured for time. She made no markings on the paper, not even to indicate the answer she had chosen. She was easily distracted, very slow to decide and offered little explanation of her rationale. She tended to relate the issues to people and situations she knows and wanted to chat at length about her linguistic observations more than focus on the MCQs.

50% - Reportedly difficult questions (2007) Q5 Q6 Q7 Q21 Q27 Q33 Q76

While all 13 students read the text passages and answered the questions in the order in which they appeared, the students differed with regard to the time they took to complete the paper and the degree to which they marked up the paper while answering. All of them passed the 'examination' with at least 40 out of 80 questions correct, but all found some of the questions difficult. The assessment records of the 13 students are discussed in section 5.3.3 below, focusing on the degree of alignment between the scores obtained by

each student for the genuine MCQ examination and for the think-aloud protocol of the same MCQs.

5.3.3 Assessment records

The students' assessment records for LIN103Y and for the think-aloud interview are given in Table 5.1 below. The problematic questions identified in Chapter 4 were discarded from the calculations:

Table 5.1 Assessment records and think-aloud results for students interviewed

Student interviewed	Written assignment mark	MCQ exam mark	Think-aloud exam mark
NG (L1 English)	85%	81%	87%
DS (L1 English)	75%	60%	61%
CL (L1 Zulu)	0% (not submitted)	62%	71%
DD (L1 Tswana)	90%	62%	61%
YP (L1 Afrikaans)	80%	86%	91%
ET (L1 Tsonga)	73%	67%	71%
AZ (L1 Xhosa)	60%	86%	83%
AJ (L1 English)	67%	80%	76%
NN (L1 Zulu)	67%	76%	67%
TH (L1 Venda)	0% (not submitted)	80%	86%
ZL (L1 Afrikaans)	87%	92%	93%
MD (L1 Lingala)	67%	70%	66%
SD (L1 Sesotho)	0% (not submitted)	48%	50%

The last two columns indicate striking similarity between the results of the actual MCQ exam and the think-aloud rerun. The difference between the two attempts was 9% or less for all the students and 6% or less for 11 of the 13. This observation supports the validity of the think-aloud method for shedding light on students' cognitive processes. The

method in this case yielded very similar test results to a genuine test-taking situation. This supports the findings of Norris (1990) described in 3.6.2.1 above.

The results of the Likert-scale opinion questions on multiple-choice assessment are reported in section 5.4 immediately below, followed in sections 5.5 to 5.8 by the results of the think-aloud interviews.

5.4 Student opinions of MCQ assessment

A 5-point Likert scale was used to gauge students' opinions on (a) how much they liked MCQ as an assessment method (b) the difficulty of MCQ assessment in general and (c) the difficulty of the LIN103Y exam they had just written. Students were also asked in each case why they gave the answer they gave. (The protocol is provided in Appendix D.) Results are reported below:

On a 5-point Likert scale, the students' general opinion of MCQ as an assessment method was mostly positive, ranging from Neutral to Really Like. Only one student, SD, really disliked the MCQ format. This was the same student who also failed the examination.

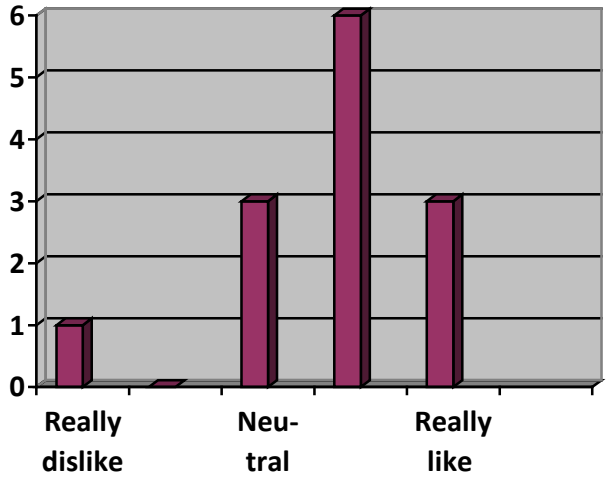

Figure 5.1 What is your general opinion of MCQ as an assessment method?

Students' opinions of MCQ assessment varied from difficult to very easy. However, most students said that in general, MCQ was easy or neutral, i.e. neither difficult nor easy as an assessment method:

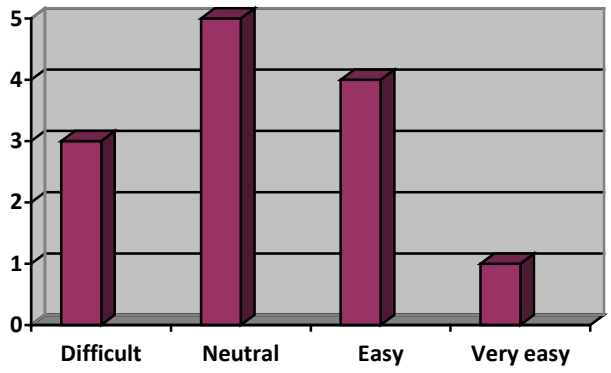

Figure 5.2 How difficult do you find MCQ as an assessment method?

All students said they had managed to finish the LIN103Y examination in the time allowed (80 questions in 2 hours), some of them in less than an hour and only one with no time to spare. The students rated the particular question paper they had written as easy (4 students) or neutral (neither difficult nor easy) (9 students):

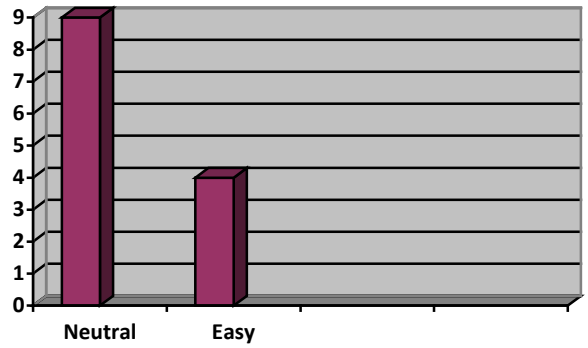

Figure 5.3 What was your opinion about the difficulty of this particular LIN103Y MCQ exam?

The above results reflect a generally positive attitude towards MCQ assessment and its fairness. The students find MCQ questions challenging and a good test of how well they know the material:

> DS '*It is reasonable. I can cope with it. It gives me pleasure to fathom out the right answer.'*
>
> ET *'You have to think logically and apply knowledge … It's, you know, the questions are such that they test a person's understanding of the subject and expose you if you haven't studied sufficiently.'*
>
> MD *'It depends how you study.'*
>
> NN '*The questions are structured so that you really need to know your work.'*
>
> TH '*You can't guess. You have to know*'.

Students commented on the fact that MCQs jog your memory by giving you options and requiring recognition rather than recall of terms and facts:

> DS *'I think it's an excellent test. It helps with recall, I have a problem with this. MCQs help me remember the words and are stimulating. It's an effort and a challenge to choose the right answer.'*
>
> ET **'***The thing is it gives an option to remember what one has forgotten. A sense of balancing and common sense is required.'*

Some students reported finding MCQ tests less demanding than written tests:

> ZL '*You feel that you're cheating cos the answers are right there. You don't have to actually remember*.'
>
> MD *'For me, you get all the answers there, unlike in written questions where you have to think.'*
>
> DD *'It doesn't have a lot of work. It is less effort*.'

Struyven, Dochy and Janssens' (2005) investigation of students' perceptions about assessment in higher education showed a similar finding, namely that students prefer the

multiple-choice format to essay-type assessment, associating it with lower perceived difficulty, lower anxiety and higher expectations of success (2005:336). These higher expectations of success are often justified. For example, two South African studies showed that pass rates were higher for both L1 English and L2 English students on MCQ than essay examinations, indicating that students, on average, are advantaged by the multiple-choice format (Curtis, de Villiers & Polonsky 1989, Miller, Bradbury & Wessels 1997).

Despite their generally positive perception of MCQ assessment, some of the students in the present study reflected on the degree of precision required by MCQs and acknowledged that MCQs can be tricky at times:

> CL '*They are short questions but yet you must be really specific and exact.'*
>
> DD '*It is actually tricky more than difficult.'*
>
> ET '*I got confused at some points. Logical guesswork was needed. It gave an opportunity to engage thoughts.'*
>
> SD '*The answers are so similar. It's so tricky to pick the precise answer. I take longer to look at it and think clearly.***'**

This is in line with research such as Nield and Wintre (1986), Roberts (1993) and Paxton (2000), which suggests that students perceive true-false and multiple-choice test formats as confusing and tricky compared to other types of test questions. Paxton (2000:114) suggests that some of her South African students reported finding MCQ tests disempowering and would prefer to express their understanding in their own words. Five of the students in the present study expressed similar sentiments, but in less negative terms:

> NG '*…I mean there're certain elements that just come out in essay writing that don't come out in multiple choice and I think with multiple choice, it's very much either right or wrong whereas with essays you can tell, I mean there's more that you can tell about how the person has learnt the work.'*

> AZ *'…it doesn't give us an opportunity to express our understanding of the module.'*
>
> CL *'If it is not MCQ you have more freedom. MCQ answers are either right or wrong.'*
>
> TH *'I'd prefer writing. You can throw in stuff. But essay marking is more subjective, which can be a problem. You have to study harder for MCQs.'*
>
> SD '*It restricts me. I want to get an opportunity to give my insight on the subject. For MCQ I have to cram. I take longer to finish the exam.'*

Although I didn't specifically ask students whether they preferred MCQ or written assessment, the above comments suggest that opinions seemed to be split fairly evenly, with five preferring written exams. Only one student really disliked MCQs and, at the other extreme, one student has such problems with memory that he can't pass written tests and deliberately chooses his courses to ensure that there is MCQ assessment in the exam. Students therefore seem very aware of the double-sided nature of MCQ assessment, of its pros and cons. None of the students underestimated the difficulty of multiple-choice assessment and multiple-choice questions and all the students I interviewed studied for the examination and gave every question due consideration.

5.5 MCQ-answering strategies

All 13 students in the study read or at least refreshed their memories by skimming the passages before attempting the questions (unlike the students in Farr, Pritchard and Smitten (1990), some of whom read the MCQs before the passages). Some of the students underlined key words in the passage as they read and some made notes as they read. On analysing transcripts of similar think-aloud interviews to the ones in the present study, Farr, Pritchard and Smitten (1990) devised a detailed typology of what actually happens when a student answers a multiple-choice examination. Some of the question-answering strategies they observed and which were also observed in all 13 interviews in the present study included the following:

- rereading questions to clarify, e.g. NG Q5 '*OK so this is a statement. When I get a statement I reread them just to make sure I understand*.'
- looking back in passages to search for specific words or phrases (Reference back to the passage is identified as RTP in my interview transcriptions.), e.g. YP Q23 'RTP. *OK I'm going back because I can't remember what they do*.' In a few cases students referred back more than necessary or failed to refer back (possibly because they remembered the relevant information from the passage).
- self-adjustment strategies such as adjusting one's reading rate or temporarily skipping a question, e.g. YP Q10 '*But ja, if I don't know it I circle it, come back to it. But I don't go through the whole paper though. I just go to where I'm supposed to* [the end of the questions on that particular passage], *so I basically finish off the section before I carry on.'*
- thoughtful consideration of various answer choices, e.g. DS '*It was just a matter of considering the most correct or best answer.'*

Martinez (1999) claims that 'test-wise' students – those with more test experience and conscious strategies for answering MCQs – are more capable with regards to 'thoughtful consideration of answer choices', for example by capitalising on information embedded in the response options, narrowing down the possible options by eliminating the implausible, and double checking their answers (Martinez 1999:211, Durham 2007:3). Several of the students interviewed in the study displayed this ability to strategise consciously about what their experience of MCQs had taught them and about what the lecturer might have intended.

In my opinion, Farr, Pritchard and Smitten's (1990) strategy of 'thoughtful consideration of various answer choices' is in fact a cover term that can be further specified to include several substrategies I observed in the think-aloud interviews. It should be noted that some of these are useful strategies that would be approved of by Durham (2007) and by the various internet sites that offer MCQ tips, while others are not to be recommended when answering MCQs:

(a) Using a process of logical elimination to get down to a subset of options, e.g.

> NG Q2 *'I mean in this case, where I can't remember, the ones that I definitely know aren't correct, I'd cross out.'*
>
> DS Q2 *'Options [1] and [2] are too young.'*
>
> DD Q6 *'It's between [2] and [3]'*

(b) Guessing between the remaining possible answers, e.g.

> NG Q76 '*Here I'd go eeny meeny between [1] and [2]'*.
>
> DD Q30 *'Between [3] and [4], because they ARE bilinguals. I'll say subtractive.'*

Although there was a fair amount of guessing taking place, it tended to be educated guessing between the two most plausible answers, often using own experience or background knowledge. I saw no evidence of mindless (but test-wise) strategies like picking the longest option, picking [3] when in doubt, etc.

(c) Selecting the option that has a term, morpheme or synonym that matches the stem, e.g.

> DD Q46 *'I was guessing, and then I said it because of like I see this 'society' here and 'classes' so I just saw that it's social, so that's why I chose 'sociolect'. But this is how we guess it.'*
>
> P: *No, but it's a good strategy. You're using the other words to try and give you clues. It's right. It's good.*
>
> NN Q16 *'I didn't remember what diglossia means. But maybe the fact that … they made mention of the word 'regional' so this would basically be about territorial* [multilingualism].'

A less sensible example of this strategy was observed in the students' answers to Q37 2006, which had a facility of 46%. Of the four 2006 students interviewed,

three got Q37 wrong because they identified the key word in the question as 'L2' ('a second or additional language'), then referred back to the passage for the actual word L2, rather than for other phrases with the same meaning. This supports Dempster's (2006:329) observation that '(a)necdotal evidence from learners and university students indicates that matching terms in the question and in the answers is a widespread practice when students do not understand the question, or do not know the answer.'

(d) Using general knowledge or personal experience to answer a question, e.g.

> CL Q80 *'I mean the cultures I know it's* [i.e. the best answer to a compliment is] *just a simple thank you.'*
>
> CL Q77 *'We* [women] *normally use long sentences and men would just be very brief in what they say.'*
>
> YP Q13 *'It's pretty much common sense. Cos I firmly believe that even though you can learn a language, you are not going to be proficient if you don't speak it. Ask me, I study Latin.'*
>
> MD Q34 *'I would take [1] cos if you are physically abused there are chances that you could be wild in a sense that you would practise back what is done to you.'*

This was actually a very common strategy and shows that students are actively thinking up analogous examples and situations as they answer MCQs, rather than just passively recalling information. All the students wanted to discuss examples of language use in their own communities during the interviews. Of course the strategy of giving one's own opinion based on one's own experience doesn't always result in a right answer, but it is a useful way to make an educated guess that will improve one's chances of a right answer.

(e) Not choosing an answer because that answer had been selected previously, e.g.

> DD Q51 *'This is the third time* [that the question has been about dialects]. *And I know that it's dialect but I'm thinking like no no no, you can't have*

dialect all the time, so you start answering wrong, but you know exactly the right answer.'

NG Q25 *'I'm hesitant to say integrative because that was the answer in the last section* [Q17] *so I'm thinking 'Would they ask it twice?'* (laughs)

(f) Not choosing an answer because it contains an absolute statement or an overgeneralisation (see also section 5.7.3 below) , e.g.

NG Q9 *'Now this is where, ja, this question could be a little bit confusing. I think this "all"* [circles *all* in option 2, 'All children acquire the language to which they have been exposed'] *would lead me to say that that's incorrect because it's so broad.'*

NG Q5 *'I look at the options first, and in this case [4], without the word* ***can****, it comes across, at this stage of my thinking, as an absolute statement, which is why I would be cautious of it.'*

MD Q44 *'It's not impossible. Impossible. The word "impossible" is not a word you use, maybe "difficult" but it said "impossible", that means it can never happen, which is not true.'*

(g) Not choosing an answer because of unfamiliarity with its meaning (see also section 5.7.2 below), e.g.

NG Q52 *'Cos I can't remember what codeswitching is, so even though I know I could be wrong, I'd go with what I remember.'*

NG Q18 '*Also with this one what I think, why this one was difficult is I've never heard of a "one-person-one-language strategy" so I'm not too sure whether this new terminology I should just sort of try figure out what it means, you know, by looking at the words and seeing if it makes sense referring back to the text, or whether I should just ignore it.'*

(h) Choosing AOTA when it occurs, e.g.

NG Q79 *'From experience, when I see All of the above, a lot of times it is the right answer.'*

(i) Selecting the first right answer and then not reading the rest of the question, e.g.

> DD Q55 '*This one. [1]. As a lazy student I didn't even go through it* [the rest of the question]. *I just saw 'the husband's name' and I chose it.'*
>
> TH Q34 '*1, 2, that's it, I can't go further. [2]'*

This strategy can lead to wrong answers when AOTA is actually the right answer, as happened with DS Q31, Q63, Q67 and Q79, TH Q21 and many other examples.

(j) Not considering NOTA as a serious option.

One of the most difficult questions in the present study, 2006 Q22 (facility 38%), had NOTA as the *right* answer. All four of the 2006 students interviewed got it wrong, due to guessing or making assumptions on the basis of common sense rather than evidence in the text. Part of the difficulty of this question in my view is also the fact that the students failed to consider NOTA as a real option. Whether or not NOTA question types should be avoided is debatable, but if NOTA is used, it should sometimes be the key.

This section has described some of the MCQ-answering strategies that students utilised during the think-aloud interviews. Since the research questions in the present study are focused on identifying difficulties in MCQs, this was one of the main aims of the think-aloud interviews, and is taken up in more detail in section 5.6.

5.6 Observed versus reported difficulties in the MCQs

The interviews yielded three ways of identifying difficult questions, namely (a) students themselves reporting having difficulty with a question, (b) observed difficulties such as misreading, paging back and forth, a very long answering time, etc. and (c) a wrong answer coupled with an explanation. The difficult questions identified by these three methods obviously overlap, as any or all of these indicators may have occurred on the same question. But between these three indicators it should be possible to pick up most of

the difficulties that students encountered, and more importantly, to get some insight into the nature of these difficulties.

Questions that students found difficult were typically identified by way of a comment that the question was difficult/hard/confusing/tricky/unfair/nasty etc. or by questioning the question or the lecturer's intention. For example,

> DS Q53 '*1 – it's not international. 2 isn't an advantage. [4]. I don't like that question.*'
>
> DS Q41 '*Are most of them bilingual? No they aren't, there are 135 different languages. It is an unfair question*.'
>
> SD Q7 '*This is the tricky part. It's very tricky. I think it's probably because you take it as, you were right in saying, a question you answer with um your perspective in mind, and* [this might differ from] *what the lecturer had in mind when she or he was asking this question*.'

Each student identified between 4 and 20 of the 80 questions as being tricky or difficult in some way, with the good students tending to identify more questions as tricky than the weaker students. This is in line with findings that good students tend to be good readers who are very aware of when their reading comprehension is less than optimal or when inferences are required (Pretorius 2000a).

The difficulties that Farr, Pritchard and Smitten's (1990) students reported experiencing while answering MCQs to test their reading comprehension were lack of interest in the passages, lack of familiarity with the passage content, too many details in the passage, inferential questions that were too difficult, ambiguous or irrelevant questions, unfamiliar vocabulary, test anxiety and difficult situational context. As we shall see, the students in the present study reported all of these difficulties with the exception of test anxiety, irrelevant questions, and lack of familiarity with the topics of the passages. As regards the last two points, the questions in the LIN103Y exam related either to the reading passages or back to the issues in the study guide, and although most of the passages were unseen,

they dealt with topics covered in the course. Text anxiety was not an issue here as the examination had already been written and their fates were already decided:

> *P: 'I'm not trying to stress you out here.'*
> *SD 'It's fine, this is much easier than the exam because in the exam you're so pressured to pass.'*

In many cases, questions manifested their difficulty indirectly rather than in direct comments from students. These observed difficulties tended to take the form of ejaculations like *sjoe* or *eish* or *shit* or heavy sighing. There was also evidence of students misunderstanding, misreading or misinterpreting questions (see 5.7.1 below). Most commonly, however, difficulties manifested themselves in long pauses and long deliberation (see discussion in 5.7.4 below).

There were multiple causes of difficult/tricky questions according to the students interviewed in this study. Reported and observed difficulties relating to readability will be discussed in section 5.7 below, followed by a discussion of reported and observed difficulties relating to the violation of other MCQ guidelines in section 5.8 and other miscellaneous difficulties in section 5.9.

5.7 Difficulties relating to readability

Readability difficulties noted in the think-aloud interviews included instances where students misread the question, difficulties understanding unfamiliar words and issues relating to lack of clarity, slow reading times and lack of interest. These issues are taken up in sections 5.7.1 to 5.7.5 below.

5.7.1 Misread questions

A few examples of misreading were observed, mainly by the L2 students, but these seldom affected students' understanding of the question or the logic of the answering process. For example, TH misread *psycholinguists* as *psychologists* on two occasions and misread NOTA as AOTA but subsequently corrected her error. NN misread *transitional* as *transactional*. MD was the weakest reader and mistook several words for other words

during his reading. In addition to several misread words which he subsequently self-corrected, he misread *primarily* as *primary*, *colloquial* as *colonial*, *production* as *pronunciation, intelligible* as *intangible*, *severe* as *several*, *introvert* as *introvect, wh-questions* as *who-questions*.

On several occasions the students experienced momentary confusion that was subsequently clarified by rereading the question, sometimes several times. An example of a homonym causing momentary confusion was ET Q60 '*The Xhosa used? Oh the Xhosa language'*, where 'Xhosa' was initially taken to mean the Xhosa people.

One interesting example of a misreading was option [2] in Q11, where '*short ungrammatical sentences'* was viewed as a typical characteristic of caretaker speech by both NG and CL, even though short grammatical sentences is in fact the appropriate characteristic. This seems to be a case where a tiny but critical negative morpheme (*un*-) is overlooked. As Roberts (1993) describes, trickiness can result from 'using a trivial word that may be easily overlooked or a confusing use of the negative that actually turns out to be crucial'. I would suggest that it is not only the short, 'missable' size of the negative morpheme that is problematic but also the position of this small but critical morpheme in the sentence. Here the correct adjective '*short'* is foregrounded and the incorrect adjective '*ungrammatical'* is hidden in the middle of the answer choice where it is easier to miss.

There were three other occasions where incorrect information in the middle of sentences was overlooked by some students. Q5 [1] '*This kind of evidence suggests that children are born with innate knowledge of the vocabulary of their language'* was viewed as true by both DS and DD. In fact, the evidence presented suggests that children are born with innate knowledge of universal aspects of language structure, but not of vocabulary. In my opinion, the trickiness in Q5 (42% facility) lies in option [1] where the first part of the statement 'Children are born with innate knowledge…' is true, and the false part of the statement '… of the vocabulary' is backgrounded somewhat by its position in the middle of the sentence. CL took Q28 [5] 'there is a period between birth and puberty when a

second language is usually acquired' to be true. In fact, this is true of a first language but not of a second language. Again the correct portion of the statement is foregrounded and the incorrect element ('a second language') appears in the middle of the sentence. Finally, DD appeared to ignore the small but critical word 'L1' in the middle of the stem of Q1, answering which language had the most speakers, as opposed to the most L1 speakers.

There was also one case where 'windowdressing' caused a wrong answer. In Q13 the less relevant beginning portion of a long stem distracted ET from what was actually being asked in the final sentence of the stem. The context in the first sentence of the stem put him on the wrong track, causing him to focus on differences between L1 and L2 speakers, not just on accent:

13. Kim's accent will signal that she is an L2 speaker of Kiswahili. The term **accent** refers to distinctive

    [1] idiomatic phrases
    [2] vocabulary items
    [3] grammatical differences
    [4] pronunciation
    [5] All of the above.

The best remedy for this kind of error is close reading and sustained focus on the part of the student. In the same way as students are advised to read *all* the options, students should be advised to read right to the end of every option before making their choice. However, test compilers can assist by avoiding windowdressing, by keeping stems as short and as focused as possible (see Haladyna 2004:108-110) and by avoiding negatives in the middle of sentences.

5.7.2 Unfamiliar words

The unfamiliar words students commented on included linguistic terms they had forgotten, terms they had never seen before, instances where the term used in the question differed from the term used in the study guide, and general academic words they didn't know. Examples of these various lexical difficulties are given below:

- Meaning of term(s) forgotten

It was clear from the think-aloud interviews that by far the majority of wrong answers were because the students had forgotten the definitions of technical linguistic terms, i.e. terms that had been introduced and defined in the study guide and subsequently forgotten by the students or confused with other technical terms:

> DD Q22 '*Substractive?* (sic) *If ever I do remember my terms, that would be easy for me (laughs). Because I don't remember my terms, I don't know.*'
>
> DD Q59 *'Dead language. Why do you have dead language and dying language? Dying language …. Is it a dying language or a dead language? I will say dead language. They have very limited competence in Yaaku.'*
>
> DD Q68 '*I wish I knew what does "lexicon" mean. So I'm gonna do guesswork because the other thing if ever you just don't know one word, you wouldn't be probably able to answer the rest of the question, so I don't know.*'
>
> CL: '*Some of the terms I've forgotten … I can't remember this, immersion programme and submersion programme.*'
>
> ET Q55 '*These are tricky terms, underextension, overextension. They need revision.*'
>
> TH Q6 '*And I had a problem with, I can't really differentiate instrumental and integrative'.*

Linguistic expressions whose meaning was not recalled during the think-aloud exam (students either acknowledged that they had forgotten the word, or assigned a clearly incorrect definition) were as follows:

> **Expressions forgotten by the 3 L1 English students**
>
> *instrumental (x2), integrative (x3), one-person-one-language strategy (x2), diglossia (x2), immersion (x2), submersion (x2), codeswitching, dual language programme (x2), ethnolect, overextension, underextension egocentric speech, semispeaker*

Expressions forgotten by the 10 L2 students

submersion (x7), immersion (x6), instrumental (x5), integrative (x5), diglossia (x5), semispeaker (x4), rememberer (x3), telegraphic stage (x3), near-native (x3), territorial monolingualism (x3), territorial multilingualism (x3), hlonipha (x3), overextension (x3), underextension, dual-language programme (x2), transitional programme (x2), function words (x2), mutually intelligible (x2), holophrase (x2), codeswitching (x2), natural sign language (x2), cued speech (x2), dialect (x2), idiolect (x2), sociolect, ethnolect, early bilingual, semilingual, sociolinguistics, one-person-one-language strategy, language shift, partial language shift, gradual death, pidgin, language acquisition, language learning, additive bilingual, subtractive bilingual, critical period hypothesis, euphemism, lexicon, national language, wild child, manual sign code, finger spelling, wh-questions

While it is to be expected that students would forget many of the terms in the 2-5 weeks intervening between the examination and the interview, the very similar think-aloud and examination results suggests that forgotten definitions was also the major reason for wrong answers in the examination itself. The issue of whether the study material provided clear definitions and sufficient examples of each concept is not addressed here, but it must be acknowledged that there were a large number of technical terms introduced in the study guide. L2 students tended to forget more words than L1 students at an average of 9,5 words each as opposed to an average of 6,7 words for each L1 student.

What is striking about the lists above is that many of the same terms were forgotten by several students. These highly forgettable terms tended to be pairs of terms which were similar in form but contrasted in meaning, such as *instrumental* and *integrative* (alternative motivations for learning a second language), or clusters of contrasting terms such as *submersion, immersion, dual-language programme* and *transitional programme,* which refer to alternative language-medium policies in schools. Contrasting word pairs (or triplets) are a feature of technical vocabulary, and differentiating between very similar terms is a necessary part of academic proficiency. However the large number of forgotten word pairs suggests that extra attention needs to be given to these word pairs in

assignments and in the study material to alert students to the formal differences (e.g. in the morphology) and to enable them to identify and memorise the meaning contrasts.

- Term never seen before:

Expressions in this category included terms that students had simply not noticed while working through the study material, and terms like *calque*, *loan translation* and *linguacide*, which were, rather unfairly, used in the examination but not in the study guide.

> NG Q18 '*Also with this one what I think, why this one was difficult is I've never heard of a one-person-one-language strategy so I'm not too sure whether this new terminology I should just sort of try figure out what it means, you know, by looking at the words and seeing if it makes sense referring back to the text, or whether I should just ignore it.'*
>
> YP Q50 '*You actually used this* [word 'calque'] *twice and I have no idea what it means. So I was like OK well I've never see that word in the Study Guide so obviously it can't mean anything! (laughs) And I didn't have a dictionary with me.'*

- Term used in question and study guide not exactly the same:

In one or two cases students picked up on a slight difference between the term used in the study guide and the examination.

> ZL Q41 '"*Critical age", now in the guide it says "critical period"* [hypothesis]'.

- General academic word not understood

In a very few cases, the L2 speakers did not understand some of the general academic terms, namely *working class*, *idiosyncratic, idiomatic, mimic* (x2) and *location*. It did not appear, however, that poor academic vocabulary per se was a major problem for any of the students in the think-aloud interviews, possibly because all of these students were average to good students. Simpler words would have none the less helped to prevent unnecessary confusion in these questions.

5.7.3 Lack of clarity

Students identified several items as being unclear or confusing due to real or perceived ambiguities in the items, overgeneralisations or hedges in the answer choices, or complex sentence structure. Examples of these various difficulties are discussed below:

- Ambiguity

Students commented that some of the questions were ambiguous, often as a result of unclear deictics such as *this, above*, etc.

> NG Q5 *'What does 'This kind of evidence' refer to?* [Text A or the statement in Q5]?
>
> DD Q21 notes the ambiguity of the term *family* (nuclear or extended), asking *'Of Jane's family, is that what you meant?'*
>
> AZ Q31 *'Ah, this one is vague, man.'*
>
> AJ Q28 *'OK this one I'm not too sure about. What do they mean by the younger members? … I'm not too sure about the younger members. Who are they referring to?'*
>
> AZ Q23 '*Now this one was a trick. …'Discussed in the case study'. In the case study, Penny, you discussed two different types of pupils! Which ones are you referring to?'*
>
> TH Q22 '*The problem that I had is we still have Zulus who are still in Murchison and we have other students who are urban, so I didn't know if you are talking about the urban ones or the Zulu ones who are still* [rural]'.

Both Grice's maxims and accepted item-writing guidelines stress the importance of unambiguous wording and clear references. This is particularly important in the case of MCQ examinations where clarification is unavailable to students. Ambiguous questions can result in examinee anxiety, and often in two right answers and/or very low item discrimination meaning that questions subsequently need to be discarded.

- Overgeneralisations

Another type of difficulty that students recognised concerned questions or answer choices containing overgeneralisations or absolute statements. Since there are always exceptions

to overgeneralisations, this kind of question leads to doubt in the minds of students, and sometimes to two right answers:

> NG Q72 '*I'm wary of absolute statements like "the most beneficial"'.*
>
> NG Q9 '*Now this is where, ja, this question could be a little bit confusing. I think this* ***all*** [circles it, in option [2] 'All children acquire the language to which they have been exposed'] *would lead me to say that that's incorrect because it's so broad. … One of the confusing things that I do find about multiple-choice is that I'm not always sure if the examiner is sort of like trying to catch me out, by dropping in these type of words, and if that's the process that I must take, or if it's a simple process of looking, because I mean you can't tell and especially at Unisa, because it's correspondence, I mean if it's at a normal university you can sort of tell what type of lecturer you have, you know what I'm saying.'*

Grice's maxims enjoin us not to say things for which we lack adequate evidence. In the context of multiple-choice, this often means avoiding opinions and overgeneralisations as well as specific determiners like *'always', never'* and '*completely'*, which are so extreme that they are seldom correct answers (Haladyna 1994:79). Overgeneralisations may confuse students and they also have the disadvantage that they are eliminated as options too easily by test-wise students.

- Hedges

Another category of confusing questions was questions containing hedges like *mostly, mainly* and *primarily,* which refer to rather vague or subjective proportions that caused uncertainty in the minds of students:

> Q4 (2006) In the data above, the children are producing
>
> [1] *mainly* three-word sentences
> [2] incomprehensible sentences
> [3] egocentric speech
> [4] short sentences containing *mostly* content words
> [5] short sentences containing *mostly* function words.
>
> NG Q4 *'I mean I think these type of words* [like '*mostly'* and '*mainly'*], *I think this is where things can get confusing.'*

Q9 (2007) Kim's reasons for learning Kiswahili are *primarily*

[1] instrumental reasons
[2] integrative reasons
[3] interference reasons
[4] internal reasons.

AZ Q9 *'Questions like that are subjective …'*

- Complex phrases

Phrases that students considered difficult to understand are discussed individually below:

Q26 (2006) 'L2 proficiency is directly related to the number of years of L2 study'

DD Q26 '*Somebody who uses English as a second language, they would not be able to get this.'*

Q76 (S1 2007) [3] 'stylistic variation'

YP *'I have no idea what that sentence* [option 3] *means, what stylistic variation means. Oh! The style can vary.*

Q73 (2006) [3] 'location of the sign relative to the body'

CL *'I'll choose [5], however, [3], 'location of the sign relative to the body', I'm sort of like not getting that.'*

All three of these examples contain nominalisations (*proficiency, study, variation, location*), which are abstract and notoriously difficult to interpret (Flesch 1962). Q26 also has a complex noun phrase structure '*the number of years of L2 study'*. It later became clear in the third example, Q73, that the academic use of the familiar words *location* (which usually means 'township') and *relative* (which usually means 'family member') were part of the problem. Homonyms (like *relative* and *location*) are problematic for L2 speakers as the most familiar meaning is the one that tends to be retrieved (as shown, for example, by research such as Schmitt (1988) on difficulties experienced by Spanish-English bilinguals on the US Scholastic Assessment Tests).

Q56 below has the double difficulty of a passive (which is more difficult to interpret than the active form) and no agent mentioned, which also makes it somewhat ambiguous as to who considers Zulu and Xhosa separate languages:

> Q56 (2007) 'Zulu and Xhosa are considered separate languages because'
> ET '*It's a tricky one cos one needs to understand "are considered separate"* [in the stem].'

In Q73 below, the word *lexicon* was not clearly understood, but the abstract, somewhat metaphorical concept of a lexicon shrinking also seemed difficult to conceptualise:

> Q73 (2007) [1] 'Its lexicon shrinks'
> AZ '*Another difficult one. Is not. ..* [skips it. Then comes back after Q74] *Educated guess. OK Let's make a guess now. [1] I don't know how a lexicon shrinks though. How can they shrink?'*

While only a few MCQs were reported as having complex syntax, these examples suggest that nominalisations, passives, homonyms, complex noun phrases and abstractions presented the greatest barriers to comprehension.

5.7.4 Slow answering times

Chall and Dale (1995:80) explain that readable material is material that students can understand, read at an optimal speed, and find interesting. The issue of reading speed and answering time was not addressed directly in this study, but some observations based on the think-aloud interviews can nevertheless be made here. I observed very striking time differences between the recall questions (usually under 30 seconds) and the higher-level and negative questions, as well as those with similar answer choices, many of which took several minutes to answer. Sometimes during a very long pause I would probe with a question, e.g. '*Is this a difficult one? Why?*'

The considerable variation in the time taken to answer questions is illustrated, for example by AJ, who answered the last half of the paper (40 questions) in just 16 minutes,

excluding time taken to read the passages. This is an average time per question of 24 seconds, but the average hides the dramatic difference between the quickest question at 5 seconds and the slowest question at 61 seconds. Of the nine questions that required relatively long deliberation for her (over 30 seconds), seven had very similar answer choices that required close differentiation.

Students also took a noticeably long time over high-level questions such as Q53 2006 (same as Q62 2007):

> 53. One advantage of choosing Xhosa as South Africa's only official language would be that
>
> [1] Xhosa is an international language
> [2] it could lead to tension between ethnic groups
> [3] it would benefit mother-tongue Nguni speakers at the expense of other language groups
> [4] it would be cheaper than an 11-language policy.

This question requires students to evaluate the truth of each of the statements and then, in a subsequent process, decide which one of the true statements is an advantage. This dual processing makes it both more difficult and more time-consuming. AJ took 36 seconds over this one, which was well above her average answering time of 24 seconds. Other students were much slower, e.g. DD Q53 '*Oh, the advantage. It would be cheaper*. (long pause) *Choosing Xhosa would lead to tension*.'

CL, MD and SD were very slow to answer the 'Which of the following is false?' questions, e.g. MD Q33 took 3 minutes 22 seconds, Q50 took 55 seconds, while Q51 took 1 minute 48 seconds. In the latter case he quickly and correctly identified the true statements, leaving only [3] as a possible answer, but still took 1:48 seconds to arrive at a final answer. The rereading and slow answering times that tended to accompany negative items, very similar answer choices and cognitively demanding questions also pointed to the difficulty of these items.

Although some questions required relatively long deliberation, this in itself is not an indication of unfairness. I think there is some scope for further investigation of relative answering times of low-level and high level MCQs to ensure that overall test times are fair, but the often-cited guideline of one question per minute (Haladyna 1994:27) seems reasonable for L1 students where there are no long passages to read. In my opinion a minute-and-a-half per question (as in this examination) would allow L1 and L2 students time to read text passages and would allow for longer deliberation of questions which require reference back to the passage, untangling of negative questions or similar answer choices or simply a longer reading time due to the length of the question.

5.7.5 Interest

It is well-known that readability, interest and recall are intimately related. The interviews reaffirmed the importance of this link, with students reporting having difficulty answering questions on sections of the work that they found uninteresting and therefore hadn't read very thoroughly:

> NG '*OK I remember having trouble with this* [section on instrumental and integrative motivation for learning a second language] *because I didn't study it that well.*'
>
> AJ Q75 '*OK I didn't learn this deaf children section so well. It just didn't click'.*

Some of the students in the think-aloud interviews admitted that they prepared less diligently (and presumably guessed more!) on sections that interested them less. Several did not study the last section of the study guide – on sign language – due to lack of interest in the topic (Q69-Q80 in the 2006 exam and Q75-80 in the 2007 exam). Students also commented that interest plays a part in motivating them to learn (or skim read or skip) certain sections of the work. For example:

> AJ Q20 on hlonipha *'I really enjoyed this cos it was so exciting to learn. I couldn't believe that people do that. Ja I enjoyed that. I really learnt something.'*
>
> AJ Q55 '*I liked that study unit* [on overextension] *(laughs). Cos it's so easy.*'

> SD Q20*'(laughs) I loved this, number 20. The avoidance is hlonipha. This is very oppressive (laughs) and sexist too. And these are the type of things that are slowly drifting away.'*

This section has addressed students' concerns and difficulties related to the readability of questions. Difficulties related to the violation of the other seven MCQ guidelines discussed in Chapter 4 are taken up in section 5.8 below.

5.8 Difficulties related to other MCQ-guideline violations

While there were no particular difficulties reported or observed for MCQs with incomplete statement stems, NOTA or answer choices that were not grammatically parallel, violations of the other guidelines reported on in Chapter 4 did cause difficulties for students. Long items are discussed in section 5.8.1 below, followed by negative stems, similar answer choices and AOTA items in sections 5.8.2 to 5.8.4.

5.8.1 Long items

One student commented on the difficulty of long MCQs and long answer choices, claiming to lose concentration before coming up with a final answer:

> YP Q21 *'OK the one that's false. I think I actually went back to this question because the answers were so long I was losing concentration. So I carried on and then I came back.'*

> YP Q32 *'This one I also went back to. That first sentence was really long. Which is not a bad thing. Cos I mean it helps, one must concentrate eventually on your work.'*
> P: *It's quite interesting that you've always been mentioning the long questions as being particularly tricky. I mean you've done it about three times.*
> YP: *Ja. I lose my concentration half way through the sentence. But if they were all short sentences I probably would have finished in half an hour (laugh)….(long pause for rereading) I'll come back because if I spend too much time on one question I lose what I'm supposed to be doing.'*

The 2007 students also complained about the 337-word (Murchison) case study (Text E) which was over a page long and had 16 associated questions. This was felt to be tiring and unnecessarily confusing:

> AZ *'The tricky one though was the one about KZN* [Text E Qs 18-33 on Murchison]. *This was tricky. Oh I'm lazy to read this. It had many more questions than the other case studies.*
>
> AJ *'I hated this question* [Text E Qs 18-33 on Murchison] *in the exam. It got confusing after a while.'*

Recommendations in this regard would include keeping passages to under, say, 250 words and keeping context-dependent sets to 12 items so that they fit on a single page or double-page spread (cf Haladyna 2004:125).

5.8.2 Negative stems

Comments from both L1 and L2 students indicated that negative stems such as 'Which of the following is false?' are considered tricky and are known to lead to accidental errors:

> NG Q9 *'This I always do underline these* ['false' in the stem] *cos in the past I've made the mistake of looking for the true ones.'*
>
> NG Q28 *'I think that what can happen if you are not concentrating is that you can just look for which statements are correct and forget to refer back to the* [stem].
>
> NN Q5 '*OK* (long pause) *Oh which one is false? Oh you see, that's another tricky thing sometimes.'*

In quite a few cases, the students started off by circling the word 'false', then forgot they were looking for the false ones, then subsequently remembered, e.g.

> ET Q51 *'I think she's using holophrases, 4, No, no, no, sorry this is supposed to be false. [1] cos she's not speaking to herself'.*
>
> AZ Q50 'RTP *[4]. No It's false. I didn't read the questions. Not true, it's false. [5], she does not use commands. [4] is false.*
> P: *Ja, that mistake happens easily I think in an exam situation*.'

These 'near-misses' are a further indication that negatives function as traps for the unwary. In several cases, there were actual misses, e.g. DS probably got three negative questions wrong due to reading too quickly and missing the negative elements. His answers suggest that in Q44 he misread '*disadvantage'* as '*advantage'*, that in Q57 he

misread '*false'* as '*true'* and that in Q66 he may have missed the '*not'*. AJ and AZ did the same on one occasion each. This study showed that negatives clearly caused difficulties and wrong answers and should be avoided if possible. This finding is in line with claims that negative questions lead to worse performance by students (Cassels & Johnstone 1980, Haladyna, Downing & Rodriguez 2002, Haladyna 2004:111).

5.8.3 Similar answer choices

L2 students commented that it was often difficult to decide between very similar answer choices. Because questions with similar answer choices needed to be logically processed and compared, they required close reading, often rereading and longer deliberation:

> CL Q55 *'OK I know it's either 3 or 4. No I'm just mixing up this syllables and same letters (laughs). I'm going to choose [3].'*
>
> DD Q59 '*Why do you have "dead language" and "dying language"* [as options]*? Dying language …. Is it a dying language or a dead language? I will say dead language.*'
>
> DD Q60 *'Is it rememberers or forgetters? I'm not sure which one is it. I will go with .. . Remembers, forgetters. Forgetters.'*
>
> ET Q76 '*Ja, this one was a tricky one. One needed to understand it clearly*.' ET identified many of the questions with similar answer choices as being tricky, e.g. Q46, Q47, Q76.

5.8.4 AOTA

Test-wise students know they need to read all the options of an MCQ before making a final decision, e.g. YP Q22: '*OK, with this one when I read the third one I was already "Oh, that's that" but I always continue reading the rest, just to make sure that I didn't miss something*.' However, I saw evidence on several occasions of students identifying the first correct answer and failing to read the remainder of the question, e.g. NG Q74 didn't read the rest after ticking [1]. CL and MD Q15 each selected answers before reading as far as option [4], AOTA. DS did this four times, missing that AOTA was the right answer in Q31, Q63, Q67 and Q79. This is an argument for avoiding AOTA, in that some students will read option [1], identify it as correct and fail to read the other options,

all of which are also correct. This is particularly likely in an exam situation when students are under time pressure. A second argument for avoiding AOTA questions is that they risk testing test-wiseness rather than knowledge (Haladyna 2004:114). NG commented: '*From experience, when I see AOTA, a lot of times it is the right answer.'*

5.9 Other issues contributing to difficulty of questions

While most of the difficulties mentioned in sections 5.7 and 5.8 above related to the language and wording of questions, and are therefore of most interest in the context of this study, students also experienced difficulties related to layout and content of questions. The only layout issue raised referred to questions requiring reference to a passage on a previous page, while content issues included statements that they had not encountered before in the study guide and questions for which they felt there was insufficient information provided or where there seemed to be more than one right answer. Questions where the expected answer was not there, questions based on hypothetical situations and questions about which students had opinions or general knowledge also proved problematic:

- Questions requiring a page turn and a page back:

Questions where students are required to page back to a passage on a previous page are distracting and can result in students losing their place and their train of thought:

> YP Q6 '*OK I have to say what made this a little bit difficult was you had to turn the page.'*
> P: *'Mmm, not good layout.'*
> YP: *'…and you had to keep going back.'*
> P: *'It's irritating, ja.'*

- Statements that were not in the study guide

The issue here seems to be that students are not sure whether to ignore these unfamiliar statements or to evaluate them on the basis of common sense (as NG commented previously).

AZ Q72 *'Jo, jo, jo. [1]- I don't remember seeing this in the book but it looks like the most appropriate answer.'*

AZ Q21 *'It's the first time I've seen it, in this question paper. It's not in the book.*
P: *'OK. If it's not in the book it's wrong? (laughs)*
AZ *'Yes (laughs) They cannot give us alternatives that you've just made up.'*

- Insufficient information provided

In some cases students felt that too little information was provided, with the result that they had to make assumptions they felt unsure about:

NG *'2, I would never put an answer like that correctly because the evidence doesn't say anything about the speakers' attitude to English, I mean I'm assuming that they have a positive attitude because they use it but I don't want to be too assumptious so I'd be wary of it.'* (note the word 'assumptious' – a once-off coining that may imply that you feel suspicious about your own assumptions!)

ZL Q46 *'Either 2 or 4. It doesn't say anything about pronunciation, so I guess it's [4].'*

- More than one seemingly right answer

More than one apparently right answer was a common problem encountered by students, and meant they had to weigh up the relative merits of these options or resort to guessing between the possible right answers.

AZ Q25 *'Eish. Number 5 and 3. This is another high-level MCQ. … Both of them can be true.'*

- The expected answer is not one of the options

Where students predict an answer before consulting the answer choices, the absence of the expected answer as one of the answer choices leads to confusion and anxiety:

DD Q15 '*I thought this was going to be an easy one. The present, here and now. The thing is when I saw 'here and now', I thought I will see 'present' somewhere here* [in the options]'.

DS Q69 '*Why don't they have lipreading as an option? They practise lipreading, they don't use sign language. OK [2.]'*

- Questions about hypothetical situations

  ET Q62 *[4] That's a tricky one also* (laughs)… *Xhosa as the only language* [i.e. the only official language of South Africa] *is unheard of, but it tests one's understanding.'*

- Questions about which students have their own experience

Questions about which students have some prior experience are difficult because students are unsure of whether they should bring their own knowledge into play or rely purely on the given information:

> AZ Q14 *'You don't have to be concerned about your perspective or your view. You just have to concentrate on this* (points to case study)*.'*
>
> SD Q7 '*OK this is where I start entering into the lecturer's mind. If the lecturer says "Mr Dlamini is a businessman with Zulu as L1 and English as L2. He attended school in Soweto." This is the tricky part. It's very tricky. I think it's probably because you take it as, you were right in saying a question you answer with um your perspective in mind, and what the lecturer had in mind when she or he was asking this question. Because in my mind, um Mr Dlamini attended school in Soweto, he can't just speak Zulu and English. Mr Dlamini is bound to speak a whole lot of other languages including Shangaan and Venda.'*

In sum, the non-linguistic difficulties that students experienced related to layout and content. The six content-related difficulties described here are part of the nature of multiple-choice questions in my view, but MCQ test designers and editors need to ensure that there is one and only one right answer and that students have been given sufficient information to answer the questions. While page-turns are difficult to avoid when there are more than seven questions per passage, they can be minimised by starting each item-set (passage and questions) on a new, preferably left-hand page as Haladyna (2004:125) suggests. From the students' side, these difficulties can be overcome by adequate preparation and 'thoughtful consideration of answer choices', including eliminating implausible options and guessing between the remaining options where necessary.

5.10 Conclusion

This chapter has focused on one-on-one think-aloud interviews in which 13 L1 and L2 students each answered 80 MCQs and explained their reasoning processes. The method yielded very similar test results to a genuine test-taking situation, and shed some light on the strategies adopted and the difficulties encountered while answering MCQs. Despite their mostly correct answers (ranging from 40 to 72 out of 80) and their generally positive perception of MCQ assessment, all the students in the present study acknowledged that MCQs can be tricky at times. The most important findings are reviewed briefly below.

Most of the students interviewed in the study proved to be very aware of their own answering strategies and points of confusion and several displayed an ability to strategise consciously about what their experience of MCQs had taught them. All the students read the text passages first and then attempted the questions in the order in which they appeared. The strategies adopted when answering questions ranged from using logical elimination to get down to a subset of options, guessing between the remaining possible answers, selecting the option that contained a term that matched the question stem in some way, selecting AOTA whenever it occurred and using general knowledge or personal experience to answer a question. There was also evidence that students were ignoring NOTA as an answer choice and were avoiding answer choices that had been selected as the answer to previous questions, that contained overgeneralisations, or whose meanings were unclear or unfamiliar.

While I did not systematically measure the time taken by students to answer each question, there were striking time differences between the recall questions (usually under 30 seconds) and the higher-level and long questions which sometimes took several minutes to answer. An average of a minute-and-a-half per question was found to be sufficient time for both L1 and L2 students in the study to read the text passages, answer the questions and deliberate over the more time-consuming questions.

The interviews yielded three ways of identifying difficult questions, namely reported difficulties or overt complaints from students about questions, observed difficulties such as a very long answering time and, thirdly a wrong answer coupled with an explanation. It was evident that there were multiple causes of these difficulties.

Difficulties relating to readability included instances where students misread the question or overlooked words or information in the middle of sentences. Some questions were unclear as a result of ambiguities, overgeneralisations, hedges, nominalisations, passives, homonyms, complex noun phrases or abstractions, which can be minimised by careful and considerate item-writing and subsequent editorial review. Some questions proved difficult owing to a lack of interest on the part of the student in certain sections of the course material. However, by far the majority of wrong answers were a result of unfamiliar vocabulary such as a forgotten technical term or pairs of similar terms rather than misunderstanding or misreading the question.

In addition to readability difficulties, students also experienced difficulties when MCQ guidelines were not adhered to. These included losing focus while answering long items, experiencing confusion and frustration while attempting to disentangle similar answer choices, missing a backgrounded but critical word or negative morpheme in the middle of a sentence, selecting the first right answer and failing to read on to the end of the question, or identifying a true option in a question phrased in the negative mode.

The only layout difficulty related to questions that required reference to a passage on a previous page. Content-related difficulties included statements that students had not encountered before in the study guide and questions for which they felt there was insufficient information provided or where there seemed to be more than one right answer. Questions where the expected answer was not there, questions based on hypothetical situations and questions about which students had opinions or general knowledge also proved problematic. The six content-related difficulties described here are part of the nature of MCQs, in my view, and can be addressed only by adequate preparation by students and by ‘thoughtful consideration of answer choices’, including

eliminating implausible options and guessing between the remaining options where necessary.

While many difficulties experienced by the three English-speaking and 10 L2 students were the same, e.g. problems with layout, content, missing important words in the middle of a question or forgetting technical terms, there were some areas in which the L2 students appeared to struggle more than the L1 students. Only L2 students mistook words for other words during decoding or failed entirely to understand an academic word or a phrase in a question. The L2 students also tended to forget more terms and definitions than the L1 students and reported more difficulties disentangling similar answer choices.

Study Unit 6 takes up these findings from the qualitative portion of the research and revisits them in combination with the quantitative findings in Chapter 4 in order to reflect further on the nature of difficulty in MCQs and to draw final conclusions.

Chapter 6
Conclusions and recommendations

6.1 Introduction

The purpose of this chapter is to review both the aims and the findings of the present study, focusing on an attempt to integrate and triangulate the findings of the quantitative study and the think-aloud protocols in order to draw further conclusions about the difficulties that students face when answering MCQs. Section 6.2 revisits the context and aims of the study while section 6.3 highlights the methodological, descriptive-analytical, theoretical and applied contributions of this research. This is followed in Section 6.4 by a discussion of the limitations of the study and suggestions regarding further avenues for research. By means of a review of the answers that the study provided to the three research questions, Section 6.5 summarises the quantitative and qualitative findings and attempts to triangulate these to offer a deeper understanding of MCQ difficulty. Section 6.6 offers a summary of the applied findings and practical recommendations for MCQ assessment design and comes to final conclusions as to whether or not the eight MCQ guidelines were supported by the findings, and section 6.7 provides a conclusion to the thesis.

6.2 Revisiting the context and aims of the study

Within the current context of heightened attention to accountability and equity in education, this research was prompted by concern over how the language of MCQ tests might be negatively influencing the test performance of the large number of L2 students in my own first-year Linguistics course and of how this risk might be mitigated. The particular focus of this study was therefore an attempt to find out what kinds of MCQs were proving difficult for Linguistics students at Unisa, focusing on how the language of test questions impacted on fairness, readability, difficulty and student performance for L2 English students in comparison to L1 students. The way in which particular assessment procedures impact on second language speakers of English is underexplored in South Africa (Paxton 2000:114) and the current study was intended to add some concrete findings and recommendations to the debate. The focus was on identifying linguistic aspects of MCQs that were causing disproportionate difficulties for L2

as opposed to L1 students and to find ways to minimise the gap in assessment performance between these two groups. Arising out of an applied problem, the study falls within the growing area of study known as educational linguistics, a special focus area within applied linguistics that addresses a range of themes that touch on both language and education (Hult 2008:13, see also Spolsky 1978, Halliday 1990, Hornberger 2001).

The aims of the study can be classified as theoretical-methodological, descriptive-analytic and applied (see section 1.3), and contributions were made at all of these levels (see 6.3 below), but the three research questions below focused predominantly on the descriptive-analytical aim, namely to shed light on the way that particular kinds of MCQs impact on L2 students:

1. Which kinds of multiple-choice questions (MCQs) are 'difficult'?
2. What kinds of MCQ items present particular problems for L2 speakers?
3. What contribution do linguistic factors make to these difficulties?

These three questions will be answered as far as possible in section 6.5 below, after a discussion of the contributions and limitations of the study in 6.3 and 6.4 respectively.

6.3 Contributions of the study

At a theoretical-methodological level the study contributed to the debate on MCQ validity and fairness for a linguistically diverse student group by drawing together a variety of methodologies from classical test theory, readability research and think-aloud interviews to investigate the notion of MCQ difficulty in a multifaceted way. Several researchers (e.g. Norris 1990, Haladyna 1994, Sireci, Wiley & Keller 1998, Paxton 2000, Rupp, Garcia & Jamieson 2001) have pointed out the need for multiple perspectives on the difficulty and validity of MCQ items – with triangulation of statistical and interview data enabling more sensitive interpretation than is typically encountered in the literature. The present study attempts this methodology, complementing think-aloud interviews with quantitative MCQ data. Although no new methodologies were used, the particular combination of item analysis measures, readability measures and think-aloud interviews has not been used elsewhere. Think-aloud protocols of an MCQ examination were shown to yield very similar test results to a genuine test-taking situation,

lending validity to this methodology as a way of investigating the difficulties encountered by students with MCQs. Coxhead's (2000) Academic Word List was also used in a new way, as a measure of the lexical density of an MCQ. While the AWL tends to be used in the literature for identifying and teaching targeted vocabulary to L2 students rather than as a measure of lexical difficulty, it was felt that this would provide a useful way to compare the lexical density of items with each other. By identifying difficult questions statistically, comparing the statistics of various subtypes of questions, analysing question readability using several different measures, and hearing what the students themselves say about why particular questions are difficult, the study was able to shed light on multiple aspects of MCQ difficulty. A contribution was also made to the debate and controversy regarding the theoretical validity of Bloom's cognitive levels. It was shown that the various Bloom levels were reflected not only in question facility, as previous research has suggested (and contested), but also in question discrimination and the score gap between L1 and L2 students. This finding provides a new source of empirical support for evaluating the validity of Bloom's taxonomy.

At the descriptive-analytical level, the study contributed to a deeper understanding of the ways in which particular types of MCQ items impact on second language speakers of English. Eight item-writing guidelines were investigated empirically to determine their effect on item quality (facility and discrimination), but also their impact on second language students in particular. The study also provided a more detailed understanding of the item quality of particular kinds of negative, AOTA and NOTA questions, which are not homogeneous in their effects on student performance, as the literature has often implied. The study also provided insight into the strategies that students make use of when answering MCQs. One of the most important contributions of the study was its exploration of the interactions between cognitive difficulty and linguistic difficulty in MCQs and of the way that various kinds of difficulties interact to make questions particularly challenging for L2 speakers. Studies have tended to focus either on readability *or* on cognitive complexity in relation to question facility and few, if any, efforts have been made in the literature up till now to disentangle the interactional effects of these two aspects of MCQ difficulty and provide a clearer view of the issue.

At an applied level the findings of the study contributed to an understanding of what constitutes a fair and comprehensible MCQ in a multilingual university context, to the debate about the necessity and practicability of readability measures for MCQs and to the empirical investigation of item-writing guidelines. Most importantly, the findings have practical implications that can inform guidelines for item design and evaluation and assist MCQ-writers to improve MCQ test validity and attempt to narrow the gap in performance between L1 and L2 students.

6.4 Limitations of the study and suggestions for further research

The study confined itself to investigating 160 MCQs over two separate assessment events in first-year Linguistics at Unisa. Although there was close agreement in most respects between the findings of the two examinations in 2006 and 2007, replicating the study on MCQs in other content subjects and at other levels of study would make for interesting comparison.

Eight language-based MCQ guidelines were systematically investigated in the quantitative portion of the study, although the qualitative interviews allowed more open-ended exploration of linguistic issues contributing to MCQ difficulty. In a few cases there were very few questions of a particular type, e.g. only nine double negatives, only four AOTA-as-distractor questions and only four NOTA-as-key questions, and the statistical findings for these item types would therefore benefit from further corroboration. It would also be interesting to further explore the unexpected finding that two similar answer choices were more difficult and caused more problems for L2 speakers than 3, 4, or 5 similar answer choices, and to see if lexically similar and syntactically similar answer choices both followed this same pattern. Further research of an experimental nature could also be conducted to compare the item quality of ‘improved’ versions of questions based on guidelines supported in this study.

While the difficulty differential measure gave insight into MCQs and MCQ-types that were particularly difficult for L2 students, differential item functioning (DIF) was not measured. Further research using statistical methods of identifying DIF will give more insight into items that function differently for L1 and L2 students of matched ability.

As regards the readability measures used, the Dale-Chall readability score is regarded as one of the most reliable (DuBay 2004), but due to the fact that most MCQs are less than 100 words long, readability was calculated by prorating from short questions to the required 100-word sample. Researchers disagree as to whether this is advisable or not (Homan, Hewitt & Linder 1994, Chall & Dale 1995, Allen, McGhee & van Krieken 2005), but the point remains that the scores were calculated in the same way for each question and therefore provide a valid means of comparing question readability, though probably not of establishing reliable reading grade levels for each question. Other readability measures (see Harrison & Bakker 1998 or Homan, Hewitt & Linder 1994) could be used in future research to compare MCQs, and relative answering times of lower-level and higher-level questions could be more systematically investigated. The usefulness of using AWL density as a practical measure of MCQ readability (see section 6.6) could also be explored further.

The findings relating to the effect of Bloom's cognitive levels on the difficulty, discrimination and L1-L2 difficulty differential of MCQs were interesting and could be followed up and explored in more depth. Further research in this area would however benefit from improved inter-rater reliability with regards to the Bloom ratings, which would result from a more homogeneous group of raters, familiar with Bloom's taxonomy, with the study material on which the assessment was based and with previous assessment events to which students had been exposed. The findings regarding the interactional effects of readability and cognitive difficulty (Bloom levels) could also be investigated in alternative ways, including using factor analysis to determine the relative importance of these and other aspects of difficulty.

As regards the qualitative aspects of the study, only 13 average to good students were interviewed. This limits the insight provided by the study into the particular problems experienced by weak readers and at-risk students answering MCQs. Further research into the difficulties experienced by at-risk L2 students would complement this study, although the think-aloud methodology requires considerable metacognitive awareness on the part of students and may not be suitable as a way of identifying and exploring the difficulties experienced by weaker readers, especially when the language used for the think-aloud protocol is not their primary language.

6.5 Revisiting the research questions, triangulated findings and conclusions

The conclusions of the study can best be summarised in the form of answers to the three research questions. In sections 6.5.1 to 6.5.3 below the qualitative and quantitative findings with respect to the three research questions are compared and discussed in order to draw together the findings from Chapters 4 and 5:

6.5.1 Which kinds of multiple-choice questions (MCQs) are 'difficult'?

'Difficulty' was measured quantitatively in this study in terms of the facility (percentage of students answering correctly) of each question and the average facility of various subtypes of questions in comparison to the overall average of approximately 70%. In the think-aloud interviews, difficult questions were identified either by the students themselves as being 'tricky' or by observed difficulties such as misreading, a long answering time or a wrong answer.

Most of the students interviewed reported finding MCQ tests less demanding than written tests and generally experienced MCQs as neutral or easy, but acknowledged that MCQs can be tricky at times and require careful deliberation. The difficulties experienced by students were of a varied nature. They included both linguistic difficulties, such as terms whose definitions had been forgotten and MCQs that were ambiguous or unclear, and non-linguistic difficulties such as lack of interest in the subject matter, layout difficulties and MCQs that were cognitively demanding, such as those at the application and analysis levels of Bloom's hierarchy.

The quantitative study showed that the most difficult items (at an average facility of 51,7%) were those with NOTA as key, probably because students did not expect NOTA to be the key and often ignored it in their deliberations. Items with AOTA as distractor were almost as difficult for the converse reason, namely that students expected AOTA to be the key and picked it without paying much attention to the other answer choices.

The following linguistic characteristics also resulted in above-average difficulty: MCQs with a high density of general academic terms, long items of over 50 words (the average question length

was 33 words), MCQs with negative stems such as ‘Which of the following is false?’, double negatives with a negative stem and at least one negative answer choice, and grammatically non-parallel options with answer choices consisting of different phrase types. The most difficult of these were double negatives, closely followed by items with negative stems. The interviews corroborated these quantitative findings, indicating that students experienced difficulties when the abovementioned MCQ guidelines were not adhered to. These difficulties included losing focus while answering long items, experiencing confusion and frustration while attempting to disentangle similar answer choices, missing a critical word in the middle of a sentence or selecting the first right answer and failing to read on to the end of the question. This resulted in wrong answers by some students to items with AOTA as the key.

The difficulty with the negative items seemed to lie in the fact that the negatives took somewhat longer to process and comprehend, and that students occasionally forgot that they were looking for the false statement among the answer choices or missed the negative word altogether. For this reason, negative stems such as ‘Which of the following is false?’ emerged clearly as a cause of wrong answers when students were not paying sufficient attention.

The slow answering time that tended to accompany long items, negative stems, very similar answer choices, context-dependent items and high-level questions contributed to the difficulty of these items, with students sometimes forgetting the question by the time that they had read to the end of these items. The interviews also showed that on the odd occasion, context-dependent items can confuse students when students fail to realise that they have to refer back to the passage, or when students come back accidentally to the wrong item after a reference to a passage that required a page turn.

The study paid particular attention to the readability of questions and the effect of readability scores on question difficulty. This was quantified by calculating the Dale-Chall readability score of individual questions, which is a function of sentence length and the number of unfamiliar words not on Dale and O’Rourke’s 3000-word list (1981). More than a third of the items had the highest Dale-Chall reading level of 16 as the MCQs contained many technical linguistic terms and because MCQs often consist of single long sentences that offer four or five possible endings

for an incomplete stem. Question difficulty did not appear to correlate with the Dale-Chall readability score or the density of unfamiliar words, as mixed positive and negative results were obtained for these correlations in the 2006 and 2007 examinations. Questions at readability level 16 were not more difficult than average, which raises doubts as to the usefulness of Dale-Chall readability analyses for individual items. The density of AWL (Academic Word List) words in each question (Coxhead 2000) was used as an additional measure of vocabulary load and readability. AWL density did correlate weakly with facility and did result in more difficult questions on average. This indicates that vocabulary load does have an effect on facility.

Support for the last-mentioned finding also came in the form of the many difficulties relating to readability in the think-aloud protocol. These included instances where students misread words, overlooked words in the middle of sentences, failed to understand words or phrases altogether or were confused by ambiguities, overgeneralisations or hedges in the questions. Some questions proved difficult due to a lack of interest on the part of the student in certain sections of the course material, as the level of reader interest is also a contributor to readability. However, by far the majority of wrong answers were a result of 'unfamiliar' vocabulary in the form of forgotten technical terms rather than of misunderstanding or misreading the question. The large number of forgotten word pairs (such as *immersion* and *submersion*, *convergence* and *divergence*, etc.) suggests that extra attention needs to be given to these word pairs in assignments and in the study material to assist students to unravel and memorise the morphological and meaning contrasts inherent in pairs (or sets) of similar technical terms.

Almost all of the 17 very difficult questions (those with a facility of below 50%) violated one or more of the MCQ item-writing guidelines investigated in the study, including eight negative items (of which four were double negatives) and six long items (50 words or more). Five items had readability scores above Grade 13 and nine had at least two similar answer choices, although the latter on its own did not seem to affect difficulty. Almost half (eight) of the most difficult items were context-dependent questions that required students to refer back to the passage before answering, adding a further layer of complication. The data seemed to suggest that a piling up of factors that make comprehension more difficult (reference back to a text, similar answer choices to compare and untangle, negatives, high readability scores and general wordiness) leads to very

difficult questions. Multiple guideline violations therefore need to be avoided in the interests of fair and comprehensible MCQ assessment.

The cognitive difficulty associated with each question was also ranked using levels 1-4 of Bloom's (1956) taxonomy. The difficulty of the level 1 (recall) and level 2 (comprehension) questions (which made up three-quarters of the sample) was fairly similar. However there was a drop in facility at level 3 (application), and a further drop at level 4 (analysis). This evidence suggested that application questions were harder than recall and comprehension questions, and that analysis questions were the hardest of all. This is as predicted by Bloom's taxonomy and therefore provides empirical evidence in support of the model.

The study also investigated the interactional effects of cognitive difficulty (Bloom level) and readability (Dale-Chall readability score and AWL density). The cognitive level of the question was shown to be more important than readability in affecting question facility. However, an interaction effect was evident whereby a high readability score had little effect on the facility of lower order questions (recall and comprehension), but a larger effect on higher-order questions (application and analysis). A high readability score coupled with the cognitive demands of a Bloom level 3-4 question resulted in reduced facility. Similar results were obtained when AWL density was used as a measure of readability. Recall questions were shown to be unaffected by AWL density, but comprehension questions proved much more difficult on average when coupled with a high density of academic words. This suggests that in the interests of fairness, attention needs to be paid to ensuring that cognitively more demanding questions are readable by the weaker readers in the group, while for recall questions, student results are less sensitive to difficult wording and long sentences.

Content-related difficulties in the MCQs included statements that students had not encountered before in the study guide and questions for which they felt there was insufficient information provided or where there seemed to be more than one right answer. Questions where the expected answer was not there, questions based on hypothetical situations and questions about which students had opinions or general knowledge also proved problematic. These content-related difficulties described here are an unavoidable part of the nature of MCQs in my view and can be

addressed only by considerate test design, careful editing, adequate preparation by students and by 'thoughtful consideration of answer choices' (Farr, Pritchard & Smitten 1990), including eliminating implausible options and guessing between the remaining options where necessary.

The question of what kinds of MCQs are difficult cannot be answered fully without making reference to items that were *not* more difficult than average. These included items with incomplete statement stems, items at readability level 16, items with negatives in the answer choices, and items with AOTA as key or NOTA as distractor. None of the students interviewed identified these features as a source of difficulty. While some students did comment on the difficulty of MCQs with similar answer choices and of context-dependent items with ambiguous referents or page-turns, these two categories of items showed mixed results in the 2006 and 2007 examinations and would need to be followed up before more definite conclusions can be reached as to their difficulty.

6.5.2 What kinds of MCQ items present particular problems for L2 speakers?

In the interests of assessment fairness, the study aimed to identify particular types of MCQs that were more difficult for L2 students than for L1 students. The quantitative measure that was used for this purpose was the difficulty differential, the average difference between the percentage of correct answers by L1 and L2 speakers for each item type. Think-aloud protocols with 10 L2 speakers provided additional insights into the nature of these difficulties.

As a result of difficulties with the medium of instruction, coupled in many cases with inferior schooling opportunities, the L2 speakers scored an average of 15% less in 2006 and 13,8% less in 2007 than the L1 speakers for the MCQ Linguistics examination. There was a correlation between difficulty and difficulty differential, indicating that the more difficult questions tended to have a higher difficulty differential, and were therefore particularly hard for L2 students.

The worst offender in terms of widening the difficulty differential between L1 and L2 students was using AOTA as a distractor. Linguistic characteristics of questions that caused more problems for L2 students than for L1 students included, in decreasing order of magnitude, negative stems and double negatives, long items, items with high AWL density (0,12 and above),

and items with 2 similar answer choices (but not 3, 4, or 5 similar answer choices). These same characteristics not only made MCQs difficult for all students, as we saw in 6.5.1 above, but also disadvantaged the L2 students with respect to the L1 students.

The difficulties faced by L2 speakers when answering MCQs was illustrated by the fact that the L2 students in the think-aloud protocol identified almost twice as many difficult questions as the L1 speakers. While many difficulties experienced by the 3 English-speaking and 10 L2 students were the same, e.g. problems with layout, content, accidentally skipping important words in the middle of a question or forgetting technical terms, there were some areas in which the L2 students appeared to struggle more than the L1 students. Only L2 students mistook words for other words during decoding or failed entirely to understand an academic word or a phrase in a question. The L2 students also misread more questions, forgot more terms and definitions than the L1 students and reported more difficulties disentangling similar answer choices (for more detail see 6.5.3 below).

The Bloom levels were shown to have an effect not only on facility, as described in 6.5.1 above, but also on discrimination and the score gap between L1 and L2 speakers, with analysis questions proving much more difficult for L2 students than for L1 students. Interactional effects of cognitive level and readability level indicated approximately average or below average difficulty differential for all the more readable questions, but a high readability score coupled with the cognitive demands of a Bloom level 3 question resulted in a high difficulty differential of 17,6% and at level 4 this escalated to a very high 31,7%. Making simultaneous high demands on students' cognitive problem-solving abilities and their comprehension abilities (with respect to the sentence length and vocabulary level of items) therefore appeared to be negatively affecting the chances of a correct answer for L2 students even more than for L1 students. Again this points to the double burden of cognitive complexity and low readability combining to heavily disadvantage L2 speakers.

Multiple violations of item-writing guidelines caused particular difficulty for second language speakers. The 25 questions with a very high difficulty differential (over 25% difference between L1 and L2 speakers) tended to have more than one of the following characteristics: more than 50

words per item, a required reference back to the text passage, a readability score of 13-15 or higher and therefore a large number of unfamiliar words and/or long sentences, two or more very similar answer choices, negative stems such as 'Which of the following is false?' or double negatives.

Item types that did *not* discriminate unduly against L2 speakers and displayed average or below-average difficulty differential included incomplete statement stems, 3, 4, or 5 similar answer choices, AOTA as key, NOTA as distractor, context-dependent items and grammatically non-parallel options. Readable questions (those at level 12 and below on the Dale-Chall index or an AWL density of 0,11 and below) were also shown to be fair to all students.

6.5.3 What contribution do linguistic factors make to these difficulties?

As we saw in the discussion of the previous two research questions, linguistic factors such as readability, question length, negatives and similar answer choices do make a difference to the difficulty that both L1 and L2 students experience when answering MCQs.

MCQ guidelines enjoin writers to make their questions as readable as possible, and the study confirmed that this is an important consideration when setting MCQs. Readability did not correlate with question difficulty, but was clearly a contributing factor, especially for L2 students as we saw in 6.5.2 above. The findings suggest that a heavy load of academic vocabulary, unfamiliar words and very long sentences can disadvantage L2 students and increase the L1 – L2 score gap. While none of these linguistic features is entirely avoidable in university-level MCQs, fairness will be compromised when these go beyond a certain level. More readable questions (with reading levels of 7-8 and below or an AWL density of 0,11 and below) resulted in less of a gap between the scores of L1 and L2 students. This supports the guideline to make items as readable as possible.

The most common readability difficulty students experienced was at the lexical level – forgetting technical terms or failing to differentiate pairs of very similar technical terms sufficiently. In a few instances, there were general academic terms or non-technical words the L2 speakers did not know (*working class, idiosyncratic, idiomatic, mimic* and *location)*, even though these had been

used in the study material. Nominalisations like *proficiency*, *study, variation* were unclear to some L2 students, while homonyms like *relative* and *location* caused problems if they were interpreted in their most familiar usage. Since students tended not to choose answer choices that they did not fully understand, these kinds of complexities were particularly unfair when they occurred in the answer key. Questions with a large proportion of academic words were the most discriminating, reinforcing the importance of high levels of vocabulary as an indicator of academic success.

A lack of clarity also affected the readability and comprehensibility of questions. Examples of wording that contributed to the difficulty of MCQs included ambiguous referents like *this evidence, the data above, the pupils in the case study*, overgeneralisations like *impossible, all* or *The most…*, which are bound to have exceptions and tend not to be chosen by students, hedges like *mostly* and *mainly* which lead to vagueness, and passives which left the agent unspecified. All of these should be avoided as far as practically possible.

While readability scores take average sentence length into account, it was also clear that the overall length of questions affected student performance. Long items (50 words or more) were slightly more difficult than average, more discriminating than average and had above average difficulty differential. The high difficulty differential suggests that long items are causing more problems for L2 students than L1 students, probably by overloading working memory. Long items had more unfamiliar words than average (14 as opposed to 9), were often read and answered slowly in the interviews and were sometimes difficult to comprehend because students tended to lose concentration, get disheartened or forget the stem by the time they had read all the answer choices. There is therefore a case to be made for limiting items to under 50 words.

In the negative questions both L1 and L2 students occasionally missed the negative elements while reading or forgot that they were supposed to be looking for false statements. The statistics supported this finding, showing that negative stems and double negatives were on average more difficult, more discriminating and harder for L2 than for L1 speakers. Negative answer choices did not pose particular problems and do not need to be avoided.

Students reported difficulty differentiating between answer choices that were syntactically similar or consisted of very similar technical terms. Difficulty differential statistics showed that questions with two similar answer choices are harder for second language speakers than questions with no similar answer choices. This could have been due to the fact that L2 speakers were confusing paired terms more than the L1 speakers.

Grammatically non-parallel options with answer choices consisting of different phrase types were more difficult than average but did not seem to pose particular problems for L2 students. The statistics do not indicate any cause for avoiding items with grammatically non-parallel options and students did not comment on this issue, but the low facility suggests that students are having to reread the stem several times to recast the sentence and that these might be time-consuming and unnecessarily confusing items.

6.6 Applied implications of the study for MCQ design and assessment

The study reaffirmed the importance of consciously addressing issues of fairness and validity at all stages of MCQ assessment: during teaching, test design, analysis of results and before reusing test questions.

Prior to the assessment it would be worthwhile for educators to pay extra attention to and provide extra practice in differentiating similar word pairs in terms not only of their meaning, but also their morphology and etymology. This is because forgetting technical terms and especially pairs of similar technical terms such as *divergence* and *convergence* was the major cause of wrong answers. As part of their preparation for MCQ assessment, which should of course include practice answering MCQs, students need to be actively reminded to read right to the end of each option and each question before selecting their final answer.

In designing the test, it may be worthwhile to restrict passages to, say, under 250 words, as students complained about the length and difficulty of the Murchison case study (337 words) and not about any of the shorter texts. Restricting context-dependent item sets to 12 items so that they fit on a single page or double-page spread would also assist students (cf Haladyna

2004:125). As we saw above, the only layout difficulty that students experienced related to questions that required reference to a passage on a previous page. While page-turns are difficult to avoid when there are more than seven questions per passage, they can be minimised by starting each item-set (passage and questions) on a new, preferably left-hand page as Haladyna (2004) suggests.

In designing the MCQs themselves, a view of MCQs as a cooperative conversation between lecturer and student and adherence to the spirit of Grice's maxims would go a long way towards making test questions accessible. Test developers should acquaint themselves thoroughly with generally accepted item-writing guidelines (e.g. Haladyna 2004), with the various MCQ review processes recommended by Haladyna (2004) and with writing guidelines such as those of Fairbairn and Fox (2009:14) which focus on making test questions as accessible as possible for L2 students. While item-writing guidelines do not need to be followed slavishly, particular attention should be paid to avoiding multiple guideline violations as these lead to questions that are very difficult and that widen the gap between L1 and L2 students. The findings relating to the eight item-writing guidelines investigated in the study can be summarised as follows:

(a) The guideline to avoid negative stems (including double negatives) was supported. These were often slow to read and proved both difficult and unfair to L2 students. Avoiding negative stems will help to narrow the gap between the scores of L1 and L2 students and help prevent wrong answers caused by students missing negative morphemes in the middle of sentences or forgetting to look for the false answer choice in a *Which of the following is false?* question. Negative answer choices did not pose particular problems and do not need to be avoided.

(b) The guideline to avoid incomplete statement stems was not supported as these were not more difficult than question stems for either L1 or L2 students.

(c) The guideline to simplify vocabulary and aim for maximum readability was supported. Readable questions benefit all students and can reduce the gap between L1 and L2 scores, especially when questions are cognitively demanding. More readable questions can be achieved by avoiding nominalisations, passives, negatives, ambiguity, overgeneralisations and hedges. Keeping academic words to a minimum, for example by

revising the wording of questions containing more than six academic words as discussed in section 4.3.4 above, would be particularly beneficial as items with high AWL density were both difficult and unfair. The study also showed that while the score gap between L1 and L2 speakers was below average for recall questions, the difficulty differential escalated rapidly for comprehension, application and analysis questions, indicating that readability is particularly critical for L2 speakers answering questions at the higher levels of Bloom's taxonomy.

(d) The guideline to keep items brief was supported as long items (50 words or more) proved difficult for all students, probably by overloading working memory. The interviews showed that long items were read and answered slowly, particularly by L2 students, and were sometimes difficult to comprehend because students tended to lose concentration, get disheartened or forget the stem by the time they had read all the answer choices. Practical ways to keep items brief include moving material that is repeated at the beginning of all the answer choices to the stem and limiting answer choices to three as suggested by Rodriguez (2005).

(e) The guideline to avoid similar answer choices was supported to some extent, although the quantitative findings were mixed and inconclusive, with 2 similar answer choices proving more problematic than 3, 4, or 5 similar answer choices. The interviews showed that students struggled with paired technical terms, and while these are an intrinsic part of the test, questions with similar answer choices required close reading and long deliberation. An attempt to reduce the linguistic similarity of a set of answer choices would in my view assist L2 students in particular.

(f) The guideline to keep answer choices grammatically parallel was not supported although non-parallel items were more difficult than average.

(g) The guideline to avoid AOTA was supported to some extent as items with AOTA as key were classified as fair and not difficult, while items with AOTA as distractor proved difficult and unfair. While I believe there is a case to be made for retaining a small number of mixed AOTA items in a test (as argued in section 4.5.5 above), in the interviews it became clear that students often chose the first right answer and missed the fact that all the answer choices were correct, making AOTA the correct answer. Students, especially weaker students, will probably benefit from eliminating this format.

(h) The guideline to keep NOTA to a minimum was supported to some extent as items with NOTA as distractor were classified as fair and not difficult, while items with NOTA as key proved difficult and unfair. In the interviews, students tended to ignore the NOTA answer choice. Again, weaker students will probably benefit from eliminating this format.

As guideline (c) above implies, the AWL density measure has some potential as a pre-test check on MCQ readability and fairness. AWL density emerged as a possible measure that could be used prior to a test to get a quick indication of the questions that could usefully be simplified to reduce the gap between L1 and L2 scores. Both AWL words and number of words are easy to count using the AWLhighlighter website (Haywood n.d.) and Microsoft Word respectively, and if a benchmark maximum level is set (say 0,15 or twice the average AWL density value), only a small number of questions would have to be reviewed. Even more practical would be to look at all questions with more than 6 AWL words (since the average question length is 33 words, questions with 6 or more AWL words are fairly likely to have an AWL density of 0,15 or more). Substituting some of the AWL words in just these questions with more familiar synonyms or paraphrases would probably help lessen the gap between L1 and L2 scores.

After the assessment it is certainly worthwhile to discard questions with low or negative discrimination before calculating the final marks and look more closely at questions with low discrimination and/or low facility before reusing these questions.

6.7 In conclusion

Within the current university context of heightened attention to accountability and fairness, empirical research into tests and test results is an important aspect of our responsibility as educators. Educators need to consider the impact of their examining practices on L1 and L2 students and take action before, during and after the assessment to minimise the risk of different success rates for different language groups.

The study reinforced the notion that the difficulty of an MCQ is affected by several different variables which interact in complex ways. These include the type of MCQ, its cognitive level, its

adherence to MCQ guidelines and its readability. As far as the language of MCQs is concerned, there is a significant amount that test compilers can do to ensure that students do not encounter unnecessarily confusing or time-consuming questions and that L2 students are not unfairly discriminated against. More readable questions benefit all students and can reduce the gap between L1 and L2 scores.

It is hoped that these findings may well be useful and generalisable in other situations where MCQs are used to test L2 students in content subjects and that they will contribute to greater test validity and more considerate testing practices for all our students.

References

Abedi, J., Lord, C., Hofstetter, C. & Baker, E. 2000. Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice* 19 (3): 16-26.

Alderson, J.C. & Urquhart, A.H. 1984. *Reading in a foreign language.* New York: Longman.

Allan, S., McGhee, M. & van Krieken, R. 2005. Using readability formulae for examination questions. Unpublished report commissioned by the Qualifications and Curriculum Authority, London. SQA Research and Information services. Accessed 28 January 2009 from http://www.ofqual.gov.uk/files/allan_et_al_____using_readability_formulae_for_examination_questions_pdf_05_1607.pdf .

Anderson, R.C. & Davison, A. 1988. Conceptual and empirical bases of readability formulas. In Davison, A. & Green, G.M. (eds) 1988. *Linguistic complexity and text comprehension: Readability issues reconsidered.* Hillsdale, New Jersey: Lawrence Erlbaum Associates: 23-53.

Anderson, L.W. & Krathwohl, D.R. (eds) 2001. *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Andres, A.M. & del Castillo, J.D. 1990. Multiple-choice tests: Power, length, and optimal number of choices per item. *British Journal of Mathematical and Statistical Psychology* 45: 57-71.

Angoff, W.H. 1993. Perspectives on Differential Item Functioning methodology. In Holland, P.W. & Wainer, H. (eds) 1993. *Differential item functioning.* Hillsdale, New Jersey: Lawrence Erlbaum Associates: 3-23.

Armbruster, B.B. 1984. The problem of inconsiderate text. In Duffey, G. (ed.) *Comprehension instruction*. New York: Longman: 202-217.

Baker, E.L., Atwood, N.K. & Duffy, T.M. 1988. Cognitive approaches to assessing the readability of text. In Davison, A. & Green, G.M. (eds) 1988. *Linguistic complexity and text comprehension: Readability issues reconsidered.* Hillsdale, New Jersey: Lawrence Erlbaum Associates: 55-83.

Balch, J. 1964. The influence of the evaluating instrument on students' learning. *American Educational Research Journal* 1: 169-182.

Bauer, L. & Nation, I.S.P. 1993. Word families. *International Journal of Lexicography* 6: 253-279.

Bejar, I.I. 1985. Speculations on the future of test design. In Embretson, S. (ed.) *Test design: Developments in psychology and psychometrics.* Orlando: Academic: 279-294.

Bejar, I.I., Embretson, S. & Mayer, R.E. 1987. Cognitive psychology and the SAT: A review of some implications. Research Report. Princeton, New Jersey: Educational Testing Service.

Bejar, I.I., Stabler, E.P. & Camp, R. 1987. Syntactic complexity and psychometric difficulty: A preliminary investigation. Research Report. Princeton, New Jersey: Educational Testing Service.

Bennett, R.E., Rock, D.A. & Wang, M. 1991. Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement* 28 (1): 77-92.

Bennett, R.E. & Ward, W.C. (eds) 1993. *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing and portfolio assessment.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Benvenuti, S. 2010. Using MCQ-based assessment to achieve validity, reliability and manageability in introductory level large class assessment. *HE Monitor* 10 *Teaching and learning beyond formal access: Assessment through the looking glass:* 21-34.

Berman, R.A. 1984. Syntactic components of the foreign language reading process. In Alderson, J.C. & Urquhart, A.H. *Reading in a foreign language.* New York: Longman. 139-159.

Biber, D. 1988. *Variation across speech and writing.* Cambridge: Cambridge University Press.

Biber, D. 1992. On the complexity of discourse complexity. *Discourse Processes* 15: 133-163.

Bird, E. & Welford, G. 1995. The effect of language on the performance of second-language students in science examinations. *International Journal of Science Education* 17: 389-397.

Blacquiere, A. 1989. Reading for survival: Text and the second language student. *South African Journal of Higher Education* 3: 73-82.

Bloom, B.S. (ed.) 1956. *Taxonomy of educational objectives. Book 1. Cognitive Domain.* London: Longman.

Bodner, G.M. 1980. Statistical analysis of multiple choice exams. *Journal of Chemical Education* 57 (3): 188-190. Accessed 6 November 2008 from http://chemed.chem.purdue.edu/chemed/stats.html.

Bormuth, J.R. 1969. Development of readability analyses. Final Report, Project No. 7-0052, Contract No. 1, OEC-3-7-070052-0326. Washington, DC: U. S. Office of Education.

Bosher, S. & Bowles, M. 2008. The effects of linguistic modification on ESL students' comprehension of nursing course test items. *Nursing Education Research* 29 (3): 165-172.

Bosher, S. 2009. Removing language as a barrier to success on multiple-choice nursing exams. In Pharris, M.D. & Bosher, S. (eds), *Transforming nursing education: The culturally inclusive environment.* New York: Springer Publishing.

Bowman, C.M & Peng, S.S. 1972. A preliminary investigation of recent advanced psychology tests in the GRE program: An application of a cognitive classification system. Unpublished manuscript. Princeton, New Jersey: Educational Testing Service.

Brewer, R.K. 1972. *The effect of syntactic complexity on readability.* Unpublished PhD thesis, University of Wisconsin.

Brown, P.J. 1999. *Findings of the 1999 plain language field test.* University of Delaware, Newark, Delaware. Delaware Education Research and Development Center.

Brown, G., Bull, J. & Pendlebury, M. 1997. *Assessing student learning in higher education.* London: Routledge.

Bruce, B. & Rubin, A. 1988. Readability formulas: matching tool and task. In Davison, A. & Green, G.M. (eds) 1988. *Linguistic complexity and text comprehension: Readability issues reconsidered.* Hillsdale, New Jersey: Lawrence Erlbaum Associates: 5-22.

Bruno, J.E. & Dirkzwager, A. 1995. Determining the optimum number of alternatives to a multiple-choice test item: An information theoretical perspective. *Educational and Psychological Measurement* 55: 959-966.

Burton, R.F. 2005. Multiple-choice and true/false tests: Myths and misapprehensions. *Assessment and Evaluation in Higher Education* 30(1): 65-72.

Bryman, A. & Burgess, R.G. 1994. *Analyzing qualitative data.* London and New York: Routledge.

Camilli, G. 1993. The case against item bias detection techniques based on internal criteria: Do item bias proceducres obscure test fairness issues? In Holland, P.W. & Wainer, H. (eds) 1993. *Differential item functioning.* Hillsdale, New Jersey: Lawrence Erlbaum Associates. 397-413.

Carneson, J., Delpierre, G. & Masters, K. *Designing and managing MCQ's.* (online handbook from UCT). Accessed 11 May 2007 from www.le.ac.uk/castle/resources/mcqman.html.

Carrell, P., Devine, J. & Eskey, D. 1988. *Interactive approaches to second language reading.* New York: Cambridge University Press.

Carstens, A. 2000. The problem-solving potential of text evaluation: Examination papers in the spotlight. *Document Design* 2 (2): 134-151.

Case, S.M. 1994. The use of imprecise terms in examination questions: How frequently is frequently? *Academic Medicine* 69: S4-S6.

Case, S.M. & Swanson, D.B. 2001. *Constructing written test questions for the basic and clinical sciences.* (3 ed.) Philadelphia: National Board of Medical Examiners.

Case, S.M., Swanson, D.B. & Becker, D.F. 1996. Verbosity, window dressing and red herrings: Do they make a better test item? *Academic Medicine* 71 (10): S28-S30.

Cassels, J.R.T. & Johnstone, A.H. 1980. *Understanding of non-technical words in science.* London: Royal Society of Chemistry.

Chall, J.S. & Dale, E. 1995. *Readability revisited: The new Dale-Chall readability formula.* Cambridge, Massachusetts: Brookline Books.

Chall, J.S., Bissex, G.L., Conard, S.S. & Harris-Sharples, S. 1996. *Qualitative assessment of text difficulty: A practical guide for teachers and writers.* Cambridge, Masachusetts: Brookline Books.

Chen, Q. & Ge, G. 2007. A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes* 26: 502-514.

Cobb, T. & Horst, M. 2004. Is there room for an academic word list in French? In Bogaards, P. & Laufer, B. (eds) *Vocabulary in a second language*. Amsterdam & Philadephia: John Benjamins: 15-38.

Coetzee, W.D. 2003. *Design principles for English medium course material for speakers of other languages.* Unpublished PhD thesis, Open University.

Cohen, A. 2007. The coming of age for research on test-taking strategies. In Fox, J., Wesche, M. Bayliss, D., Cheng, L., Turner, C. & Doe, C. (eds) *Language testing reconsidered*. Ottawa: University of Ottawa Press: 89-111.

Coleman, E.B. 1964. The comprehensibility of several grammatical transformations. *Journal of Applied Psychology* 48: 186-190.

Connolly, J.A & Wantman, M.J 1964. An exploration of oral reasoning processes in responding to objective test items. *Journal of Educational Measurement* 1(1): 59-64.

Conrad, S.M. 1996. Investigating academic text with corpus-based techniques: An example from biology. *Linguistics and Education* 8 (3): 299-326.

Cooper, P.A. 1995. In search of sufficient vocabulary: Testing the vocabulary levels of undergraduate students. *South African Journal of Linguistics* 26: 25-37.

Coxhead, A. 2000. A new Academic Word List. *TESOL Quarterly* 34 (2): 213-238.

Crehan, K.D. & Haladyna, T.M. 1991. The validity of two item-writing rules. *Journal of Experimental Education* 59: 183-92.

Crehan, K.D., Haladyna, T.M. & Brewer, B.W. 1993. Use of an inclusive option and the optimal number of options for multiple-choice items. *Educational and Psychological Measurement* 53: 241-47.

Curtis, P.J.D., de Villiers, J.U. & Polonsky, M.J. 1989. Inter-group differences in examination scores resulting from multiple-choice and essay questions. *South African Journal of Education* 9 (2): 260-262.

Dahl, O. 2004. *The growth and maintenance of linguistic complexity.* Amsterdam: John Benjamins.

Dale, E. & O'Rourke, J. 1981. *The living word vocabulary.* Chicago: World Book/Childcraft International.

Dale, E. & Chall, J.S. 1948. A formula for predicting readability. *Educational Research Bulletin* 27: 11-20, 37-54.

Daneman, M., MacKinnon, G.E. & Gary Waller, T. (eds) 1988. *Reading research: Advances in theory and practice.* Vol 6. San Diego: Academic Press.

Davies, A. 1984. Simple, simplified and simplification: what is authentic? In Alderson, J.C. & Urquhart, A.H. *Reading in a foreign language.* New York: Longman: 181-195.

Davison, A. & Green, G.M. 1988. Introduction. In Davison, A. & Green, G.M. (eds) 1988. *Linguistic complexity and text comprehension: Readability issues reconsidered.* Hillsdale, New Jersey: Lawrence Erlbaum Associates: 1-4.

Dempster, E.R. 2006. Strategies for answering multiple choice questions among South African learners. Conference paper, 4th sub-regional conference on assessment in education, hosted by Umalusi, University of Johannesburg, 26-30 June. Accessed 18 February 2009 from http://www.umalusi.org.za/ur/ Conferences/apple/2006.06.14. conference %20papers. pdf.

Dempster, E.R. & Reddy, V. 2007. Item readability and science achievement in TIMSS 2003 in South Africa. *Science Education* 91(6): 906-925.

Devine, J., Carrell, P.L. & Eskey, D.E. (eds) 1987. *Research on reading English as a second language.* Washington, DC: TESOL.

DISA 2010. Online report on student profile generated by Department of Informational and Statistical Analysis, Unisa. Accessed 4 May 2010 from Institutional Information and Analysis Portal at http://heda.unisa.ac.za/heda/fsMain.htm.

Dochy, F., Moerkerke, G., De Corte, E. & Segers, M. 2001. The assessment of quantitative problem-solving with 'none of the above'-items (NOTA items). *European Journal of Psychology of Education* 26 (2):163-177.

Dorans, N.J. & Holland, P.W. 1993. DIF detection and description: Mantel-Haenszel and standardization. In Holland, P.W. & Wainer, H. (eds) 1993. *Differential item functioning.* Hillsdale, New Jersey: Lawrence Erlbaum Associates: 35-66.

Downing, S.M. 2002. Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine* 77 (10): S103-S104.

DuBay, W.H. 2004. The principles of readability. Accessed 26 July 2007 from http://www.impact-information.com/impactinfo/readability02.pdf.

Duffield, K.E. & Spencer. J.A. 2002. A survey of medical students' views about the purposes and fairness of assessment. *Medical Education* 36: 879-886.

Durham, G. 2007. *Teaching test-taking skills: Proven techniques to boost your student's scores.* Lanham, Maryland: Rowman & Littlefield Education.

Eisley, M.E. 1990. *The effect of sentence form and problem scope in multiple-choice item stems on indices of test and item quality.* Unpublished doctoral thesis. Brigham Young University, Provo.

Fairbairn, S.B. & Fox, J. 2009. Inclusive achievement testing for linguistically and culturally diverse test takers: Essential considerations for test developers and decision makers. *Educational Measurement: Issues and Practice* 28 (1): 10-24.

Farr, R., Pritchard, R. and Smitten, B. 1990. A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement* 27: 209-226.

Fellbaum, C. 1987. A preliminary analysis of cognitive-linguistic aspects of sentence completion tasks. In Freedle, R.O. & Duran, R.P. (eds) *Cognitive and linguistic analyses of test performance.* Norwood, New Jersey: Ablex. 193-207.

Fellenz, M.R. 2004. Using assessment to support higher level learning: The multiple choice item development assignment. *Assessment and Evaluation in Higher Education* 29 (6): 703-719.

Fleiss, J.L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76 (5): 378-382.

Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology* 32 (3): 221-223.

Flesch, R. 1950. *How to test readability.* New York: Harper & Brothers.

Flesch, R. 1962. *The art of plain talk.* New York: Collier Books.

Frary, R.B. 1991. The none-of-the-above option: An empirical study. *Applied Measurement in Education* 4: 115-124.

Frederiksen, N. 1984. The real test bias. *American Psychologist* 39: 193-202.

Frederiksen, N., Mislevy, R.J. & Bejar, I.I. 1993. *Test theory for a new generation of tests.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Freebody, P. & Anderson, R.C. 1983. Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Reading Research Quarterly* 18: 277-294.

Fry, E. 1968. A readability formula that saves time. *Journal of Reading* 11 (7): 265-71.

Fry, E.B.1977. Fry's readability graph. *Journal of Reading* 20: 242-52.

Fry, E.B. 1987. The varied used of readability measurement today. *Journal of Reading* 30 (4): 338-43.

Gierl, M.J. 2005. Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement: Issues and Practice.* Spring: 3-14.

Ghorpade, J. & Lackritz, J.R. 1998. Equal opportunity in the classroom: Test construction in a diversity-sensitive environment. *Journal of Management Education* 22 (4): 452-471.

Gleason, J. 2008. An evaluation of mathematics competitions using Item Response Theory. *Notices of the American Mathematical Society* 55 (1): 8-15.

Goh, D.S. 2004. *Assessment accommodations for diverse learners.* Boston: Pearson.

Govender, D. 2004. Computer-based assessment. In Heydenrych, J., & Govender, D. *Distance learning technologies in context: Selected papers on agent, process and product at the University of South Africa.* Vol 1. Pretoria: Unisa Press: 87-110.

Graff, G. 2003. *Clueless in academe: How schooling obscures the life of the mind.* New Haven and London: Yale University Press.

Grice, H.P. 1975. Logic and conversation. In Cole, P., & Morgan, J. (eds) *Syntax and Semantics 3: Speech Acts.* New York: Academic Press: 41–58.

Gross, L.J. 1994. Logical versus empirical guidelines for writing test items. *Evaluation and the Health Professions* 17: 123-126.

Haladyna, T.M. 1994. *Developing and validating multiple-choice items.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Haladyna, T.M. 1997. *Writing test items to evaluate higher order thinking.* Boston: Allyn & Bacon.

Haladyna, T.M. 2004. *Developing and validating multiple-choice test items.* (3 ed.) Mahwah, NJ: Lawrence Erlbaum Associates.

Haladyna, T.M. & Downing, S.M. 1989. A taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education* 2: 37-50.

Haladyna, T.M., Downing, S.M. & Rodriguez, M.C. 2002. A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education* 15 (3): 309-334.

Halliday, M.A.K. 1985. *An introduction to Functional Grammar*. London: Edward Arnold.

Halliday, M.A.K. 1989. *Spoken and written language*. (2 ed.) Oxford: Oxford University Press.

Halliday, M.A.K. 1990. New ways of meaning: The challenges to applied linguistics. *Journal of Applied Linguistics* 6: 7-36.

Halliday, M.A.K. 1994. *An introduction to Functional Grammar*. (2 ed.) London: Edward Arnold.

Halliday, M.A.K. & Hasan, R. 1976. *Cohesion in English.* London: Longman.

Hambleton, R.K. & Jones, R.W. 1993. Comparison of Classical Test Theory and Item Response Theory and their application to test development. *Educational Measurement: Issues and Practice* 12 (3): 535-556.

Hampton, D.R. 1993. Textbook test file multiple-choice questions can measure (a) knowledge, (b) intellectual ability, (c) neither, (d) both. *Journal of Management Education* 17 (4): 454–471.

Hancock, G.R. 1994. Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *Journal of Experimental Education* 62 (2): 143-57.

Haney, W. & Scott, L. 1987. Talking with children about tests: An exploratory study of test item ambiguity. In Freedle, R.O. & Duran, R.P. (eds) *Cognitive and linguistic analyses of test performance*. Norwood, New Jersey: Ablex. 298-368.

Harasym, P.H., Doran, M.L., Brant, R. & Lorscheider, F.L. 1993. Negation in stems of single-response multiple-choice items. *Evaluation and the Health Professions* 16 (3): 342-357.

Harasym, P.H., Leong, E.J., Violato, C. Brant, R., & Lorscheider, F.L. 1998. Cuing effect of 'all of the above' on the reliability and validity of multiple-choice test items. *Evaluation and the Health Professions* 21: 120-133.

Harasym, P.H., Price, P.G., Brant, R., Violato, C. & Lorscheider, F.L. 1992. Evaluation of negation in stems of multiple-choice items. *Evaluation and the Health Professions* 15: 198-220.

Harrison, S. & Bakker, P. 1998. Two new readability predictors for the professional writer: Pilot trials. *Journal of Research in Reading* 21 (2): 121-138.

Hay, H.R. & Marais, F. 2004. Bridging programmes: Gain, pain or all in vain*? South African Journal of Higher Education* 18 (2): 59-75.

Haywood, S. n.d. AWLHighlighter. Accessed 12 February 2007 from http://www.nottingham.ac.uk/%7Ealzsh3/acvocab/awlhighlighter.htm

Hewitt, M.A. & Homan, S.P. 2004. Readability level of standardized test items and student performance: The forgotten validity variable. *Reading Research and Instruction* 43 (2): 1-16.

Holland, P.W. & Wainer, H. (eds) 1993. *Differential item functioning.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Holsgrove, G. & Elzubeir, M. 1998. Imprecise terms in UK medical multiple-choice questions: What examiners think they mean. *Medical Education* 32: 342-350.

Homan, S., Hewitt, M. & Linder, J. 1994. The development and validation of a formula for measuring single-sentence test item readability. *Journal of Educational Measurement* 31 (4): 349-358.

Hornberger, N.H. 2001. Educational linguistics as a field: A view from Penn's program on the occasion of its 25th anniversary. *Working Papers in Educational Linguistics* 17 (1-2): 1-26.

Hubbard, E.H. 1996. Profiling contextual support for technical and general academic terms in writing for students. *South African Journal of Linguistics* Supplement 32: 95-106.

Hubbard, E.H. 2005. Readability and language complexity in old and new law study guides: A corpus-based account of a linguistic shift. *Language Matters* 2: 176-192.

Huckin, T.N. 1983. A cognitive approach to readability. In Anderson, P.V., Brockman, R.J. & Miller, C.R. *New essays in technical and scientific communication: Research, theory, practice.* New York: Baywood: 90-108.

Hult, F.M. 2008. The history and development of educational linguistics. In Spolsky, B. & Hult, F.M. (eds) *The handbook of educational linguistics.* Malden, Massachusetts: Blackwell Publishing: 10-24.

Hunt, K. 1965. Differences in grammatical structures written at three grade levels. NCTE Research Report 3. Urbana, Illinois: National Council of Teachers of English.

Hyland, K. & Tse, P. 2007. Is there an 'academic vocabulary?' *TESOL Quarterly* 41 (2): 235-253.

Independent Online 2008. 'One fifth of students graduate' August 18. Accessed 24 October 2008 from http://www.iol.co.za/index.php?set_id=1&click_id=105&art_id=nw20080818145329748C156847.

Jansen, J.D. 2003. On the state of South African universities: Guest editorial. *South African Journal of Higher Education* 17 (3): 9-12.

Johnstone, A.H. 1983. Training teachers to be aware of student learning difficulties. In Tamir, P., Hofstein, A. & Ben Peretz, M. (eds) *Preservice and inservice education of science teachers.* Rehovot, Israel & Philadelphia: Balaban International Science Services: 109-116.

Just, M.A. & Carpenter, P.A. 1987. *The psychology of reading and language comprehension.* Boston: Allyn & Bacon.

Kehoe, J. 1995a. Writing multiple-choice test items. *Practical Assessment, Research and Evaluation* 4 (9). Accessed 11 May 2007 from http://PAREonline.net/getvn.asp?v=4&n=9.

Kehoe, J. 1995b. Basic item analysis for multiple-choice tests. *Practical Assessment, Research and Evaluation* 4 (10) Accessed 11 May 2007 from http://PAREonline.net/getvn.asp?v=4&n=10.

Kemper, S. 1988. Inferential complexity and the readability of texts. In Davison, A. & Green, G.M. (eds) *Linguistic complexity and text comprehension: Readability issues reconsidered.* Hillsdale, New Jersey: Lawrence Erlbaum Associates. 141-165.

Kilfoil, W.R. 2008. Assessment in higher education. In Dreyer, J.M. (ed.) *The educator as assessor.* Pretoria: Van Schaik: 106-143.

Kilpert, O. 2008. Interpreting Unisa MCQ statistics. Personal communication.

Kintsch, W. & Miller, J.R. 1981. Readability: a view from cognitive psychology. In *Teaching: Research reviews*. Neward, Delaware: International Reading Association.

Klare, G.R. 2000. The measurement of readability: Useful information for communicators. *ACM Journal of Computer Documentation* 24 (3): 11-25.

Kniveton, B.H. 1996. A correlational analysis of multiple-choice and essay assessment measures. *Research in Education* 56: 73-84.

Knowles, S.L., & Welch, C.A. 1992. A meta-analytic review of item discrimination and difficulty in multiple-choice items using none-of-the-above. *Educational and Psychological Measurement* 52: 571-577.

Kolstad, R.K & Kolstad, R.A. 1991. The option ‘none-of-these’ improves multiple-choice test items. *Journal of Dental Education* 55: 161-63.

Kostin, I. 2004. Exploring item characteristics that are related to the difficulty of TOEFL dialogue items (Research Report 79). Princeton Educational Testing Service.

Krathwohl, D.R. 2002. A revision of Bloom's taxonomy: An overview. *Theory into Practice* 41 (4): 212-218.

Kress, G. 1985. *Linguistic processes in sociocultural practice.* Oxford: Oxford University Press.

Kunnan, A.J. 2000. Fairness and justice for all. In Kunnan, A.J (ed.) *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium.* Orlando, Florida: 1-14.

Lampe, S. & Tsaouse, B. 2010. Linguistic bias in multiple-choice test questions. *Creative Nursing* 16 (2): 63-67.

Landis, J.R. & Koch, G.G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33: 159-174.

Langer, J.A. 1987. The construction of meaning and the assessment of comprehension: An analysis of reader performance on standardized test items. In Freedle, R.O. & Duran, R.P. (eds) *Cognitive and linguistic analyses of test performance*. Norwood, New Jersey: Ablex: 225-244.

Lassen, I. 2003. *Accessibility and acceptability in technical manuals: A survey of style and grammatical metaphor*. Amsterdam: John Benjamins.

Leedy, P.D. & Ormrod, J.E. 2001. *Practical research: Planning and design*. (7 ed.) New Jersey: Prentice-Hall.

Lesgold, A.M., Roth, S.F. & Curtis, M.E. 1979. Foregrounding effects in discourse comprehension. *Journal of Verbal Learning and Verbal Behavior* 18: 291-308.

Lesiak, J. & Bradley-Johnson, S. 1983. *Reading assessment for placement and programming.* Springfield: Charles C. Thomas.

Lyman, H.B. 1998. *Test scores and what they mean*. (6 ed.) Boston: Allyn & Bacon.

Martin, J.R. 1989. *Factual writing*. Oxford: Oxford University Press.

Martin, J.R. 1991. Nominalization in science and humanities: Distilling knowledge and scaffolding text. In Ventola, E. (ed.) *Functional and systemic linguistics.* Berlin: Mouton de Gruyter: 307-337.

Martinez, M. 1999. Cognition and the question of test item format. *Educational Psychologist* 34 (4): 207-218.

McCoubrie, P. 2004. Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher* 26 (8): 709-712.

McNaught, C.M. 1994. Learning science at the interface between Zulu and English. Unpublished PhD thesis, University of Natal, Pietermaritzburg.

Messick, S. 1987. Assessment in the schools: Purposes and consequences. Research report 87-51. Princeton, New Jersey. Educational Testing Service.

Messick, S. 1989. Validity. In Linn, R.L. (ed.) *Educational measurement* (3 ed). New York: Macmillan: 13-103.

Miller, R., Bradbury, J. & Lemmon, G. 2000. Justifying means with ends: Assessment and academic performance. *South African Journal of Higher Education* 14(1): 166-173.

Miller, R., Bradbury, J. & Wessels, S.L. 1997. Academic performance of first and second language students: Kinds of assessment. *South African Journal of Higher Education* 11(2): 70-79.

Mislevy, R.J. 1993. Foundations of a new test theory. In Frederiksen, N., Mislevy, R.J. & Bejar, I.I. *Test theory for a new generation of tests.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Mitchell, J. & Roman, R. 2006. Report on research conducted on twenty modules with low pass rates from the Unisa College of Law. Unpublished report by the Unisa Institute for Curriculum and Learning Development. Pretoria.

Mobley, M. 1986. *Evaluating curriculum materials.* York: Longman.

Mouton, J. 2001. *How to succeed in your Masters' and Doctoral studies: A South African guide and resource book.* Pretoria: Van Schaik.

Mudraya, O. 2006. Engineering English: A lexical frequency instructional model. *English for Specific Purposes* 25: 235-256.

Nagy, W., Anderson, R., Schommer, M., Scott, J.A. & Stallman, A. 1989. Morphological families in the internal lexicon. *Reading Research Quarterly* 24: 262-281.

Nel, C., Dreyer, C. & Kopper, M. 2004. An analysis of the reading profiles of first-year students at Potchefstroom University: a cross-sectional study and a case study. *South African Journal of Education* 24 (1): 95-103.

Nel, C., Troskie-de Bruin, C. & Bitzer, E. 2009. Students' transition from school to university: Possibilities for a pre-university intervention. *South African Journal of Higher Education* 23 (5): 974-991.

Nield, A. & Wintre, M. 1986. Multiple-choice questions with an option to correct: Student attitudes and use. *Teaching of Psychology* 13 (4): 196-199.

Norris, S.P. 1990. Effects of eliciting verbal reports of thinking on critical thinking test performance. *Journal of Educational Measurement* 27: 41-58.

O' Neill, K.A. & McPeek, W.M. 1993. Item and test characteristics that are associated with Differential Item Functioning methodology. In Holland, P.W. & Wainer, H. (eds) *Differential item functioning.* Hillsdale, New Jersey: Lawrence Erlbaum Associates: 255-279.

Orey, n.d. website that says synthesis can't be tested with MCQs.

Osterlind, S.J. 1998. *Constructing test items: Multiple-choice, constructed response, performance and other formats.* (2 ed.) Norwell, Massachusetts: Kluwer Academic Publishers.

Paxton, M. 1998. Transforming assessment practices into learning processes: Multiple-choice questions and the essay as tools for learning. In Angelil-Carter, S. (ed.) *Access to success: Literacy in academic contexts.* Cape Town: UCT Press.

Paxton, M. 2000. A linguistic perspective on multiple choice questioning. *Assessment and Evaluation in Higher Education* 25(2): 109-119.

Perfetti, C.A. 1988. Verbal efficiency in reading ability. In Daneman, M., McKinnon, G.E. & Gary Waller, T. (eds) *Reading research: Advances in theory and practice.* Vol 6. San Diego: Academic Press.

Perkins, D.M. 1991. Improvement of reading and vocabulary skills at the University of Transkei. *South African Journal of Education* 11: 231-235.

Personal communication 2010 with Unisa colleague on the difficulty of assigning Bloom ratings to MCQs from an unfamiliar discipline.

PlainTrain Plain Language Online Training Program. Accessed 4 June 2010 from http://www.plainlanguagenetwork.org/plaintrain/.

Pretorius, E.J. 2000a. Inference generation in the reading of expository texts by university students. Unpublished D Litt et Phil thesis. Pretoria: University of South Africa.

Pretorius, E.J. 2000b. Reading and the Unisa student: Is academic performance related to reading ability? *Progressio* 22 (2). Accessed 2 February 2010 from http://www.unisa.ac.za/default.asp?Cmd=ViewContent&ContentID=13398.

Pretorius, E.J. 2002. Reading ability and academic performance in South Africa: Are we fiddling while Rome is burning? *Language Matters* 33 (1): 169-196.

Pretorius, E.J. 2005. The reading process. In *Only study guide for LIN307D: Text Quality: Theories, models and techniques.* Pretoria: Unisa.

Prins, E.D. & Ulijn, J.M. 1998. Linguistic and cultural factors in the readability of mathematics texts: The Whorfian hypothesis revisited with evidence from the South African context. *Journal of Research in Reading* 21 (2): 139-159.

Randolph, J.J. 2008. *Online Kappa Calculator.* Accessed 21 July 2010 from http://justus.randolph.name/kappa.

Readability: How readable are your texts? Accessed 7 May 2010 from http://www.readability.biz/.

Read Educational Trust n.d. Literacy in SA schools. Accessed 10 November 2010 from http://www.read.org.za/index.php?id=48.

Rich, C.E. & Johanson, G.A. 1990. An item-level analysis of 'none-of-the-above'. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Richichi, R.V. 1996. An analysis of test-bank multiple-choice items using item-response theory. Research report. Accessed 11 May 2010 from http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/16/5d/eb.pdf.

Roberts, D.M. 1993. An empirical study on the nature of trick test questions. *Journal of Educational Measurement* 30: 331-344.

Rodriguez, M.C. 2005. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice.* Summer: 3-13.

Rogers, W.T. & Harley, D. 1999. An empirical comparison of three-choice and four-choice items and tests: susceptibility to testwiseness and internal consistency reliability. *Educational and Psychological Measurement* 59 (2): 234-247.

Royer, J.M., Cisero, C.A. & Carlo, M.S. 1993. Techniques and procedures for assessing cognitive skills. *Review of Educational Research* 63: 201-243.

Rupp, A.A., Garcia, P. & Jamieson, J. 2001. Combining multiple regression and CART to understand difficulty in second language reading and listening comprehension test items. *International Journal of Testing* 1 (3): 185-216.

Sambell, K., McDowell, L. & Brown, S. 1997. 'But is it fair?': An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation* 23 (4): 349-371.

Sandoval, J. & Duran, R.P. 1998. Language. In Sandoval, J., Frisby, C.L., Geisinger, K.F., Scheuneman, J.D. & Grenier, J.R. (eds) *Test interpretation and diversity: Achieving equity in assessment.* Washington DC: American Psychological Association.

Schleppegrell, M.J. 2001. Linguistic features of the language of schooling. *Linguistics and Education* 12 (4): 431-459.

Schmitt, A.P. 1988. Language and cultural characteristics that explain differential item functioning for Hispanic examinees on the Scholastic Aptitude Test. *Journal of Educational Measurement* 25: 1-13.

Schmitt, A.P., Holland, P.W. & Dorans, N.J. 1993. Evaluating hypotheses about Differential Item Functioning. In Holland, P.W. & Wainer, H. (eds) 1993. *Differential item functioning.* Hillsdale, New Jersey: Lawrence Erlbaum Associates. 281-315.

Schmitt, D. & Schmitt, N. 2005. *Focus on vocabulary: Mastering the academic word list.* New York: Pearson ESL.

Schmitt, N. & Zimmerman, C.B. 2002. Derivative word forms: What do learners know? *TESOL Quarterly* 36 (2): 145-167.

Scholfield, P. 1994. *Quantifying language.* Clevedon, Avon: Multilingual Matters.

Schriver, K.A. 2000. Readability formulas in the new millennium: What's the use? *ACM Journal of Computer Documentation* 24 (3):138-140.

Scott, I., Yeld, N., & Hendry, J. 2007. *Higher Education Monitor: A case for improving teaching and learning in South African Higher Education.* Pretoria: Council for Higher Education. Accessed 29 April 2010 from http://www.che.ac.za/documents/d000155.

Seddon, G.M. 1978. The properties of Bloom's taxonomy of education objectives for the cognitive domain. *Review of Educational Research* 48 (2): 303-323.

Seliger, H.W. & Shohamy, E. 1989. *Second language research methods.* Oxford: Oxford University Press.

Sireci, S.G., Wiley, A. & Keller, L.A. 1998. An empirical evaluation of selected multiple-choice item writing guidelines. Paper presented at the Annual meeting of the Northeastern Educational Research Association, Ellenville, New York.

Sperber, D. & Wilson, D. 1986. *Relevance: Communication and cognition.* Oxford: Basil Blackwell.

Spolsky, B. 1978. *Educational linguistics: An introduction.* Rowley, Massachusetts: Newbury House.

Spolsky, B. 2008. Introduction: What is educational linguistics? In Spolsky, B. & Hult, F.M. (eds) *The handbook of educational linguistics.* Malden, Massachusetts: Blackwell Publishing: 1-9.

Statistics South Africa Census 2001. Accessed 4 February 2010 from http://www.statssa.gov.za/census01/html/RSAPrimary.pdf.

Steiner, G. 1978. *On difficulty and other essays*. Oxford: Oxford University Press.

Stiggins, R.J. 2005. *Student-involved assessment **for** learning.* (4 ed.) New Jersey: Pearson Prentice Hall.

Strauss, P.R. 1995. Procedural knowledge of ESL readers in decoding expository text. Unpublished D.Ed thesis: Rand Afrikaans University.

Strother, J. & Ulijn, J. 1987. Does syntactic rewriting affect English for Science and Technology (EST) text comprehension? In Devine, J., Carrell, P. & Eskey, D. (eds) *Research in reading in English as a second language.* New York: Cambridge University Press.

Struyyven, K., Dochy, F. & Janssens, S. 2005. Students' perceptions about evaluation and assessment in higher education: A review. *Assessment and Evaluation in Higher Education* 30 (4): 325-341.

Swales, J.M. 1990. *Genre analysis: English in academic and research settings.* Cambridge: Cambridge University Press.

Tamir, P. 1993. Positive and negative multiple-choice items: How different are they? *Studies in Educational Evaluation* 19: 311-325.

Taylor, I. & Taylor, M.M. 1990. *Psycholinguistics: Learning and using language.* Englewood Cliffs, New Jersey: Prentice Hall.

Terblanche, L. 2009. A comparative study of nominalization in L1 and L2 writing and speech. *Southern African Linguistics and Applied Language Studies* 27 (1): 39-52.

Thompson, S.J., Johnstone, C.J., & Thurlow, M.L. 2002. *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National

Center on Educational Outcomes. Accessed 30 January 2009 from http://education.umn.edu/NCEO/OnlinePubs/ Synthesis44. html

Thorndike, R.L. 1973. *Reading comprehension education in fifteen countries.* New York: Wiley.

Tuckman, B.W. 1999. *Conducting educational research.* (5 ed.) Fort Worth, Texas: Harcourt Brace.

Urquhart, A.H. 1984. The effect of rhetorical ordering on readability. In Alderson, J.C. and Urquhart, A.H. *Reading in a foreign language.* New York: Longman: 160-180.

Van Rooyen, D. & Jordaan, H. 2009. An aspect of language for academic purposes in secondary education: Complex sentence comprehension by learners in an integrated Gauteng school. *South African Journal of Education* 29 (2): 271-287.

Varughese, K.V. & Glencross, M.J. 1997. The effect of positive and negative modes of multiple choice items on students' performance in Biology. *South African Journal of Higher Education* 11 (1): 177-179.

Vongpumivitch, V., Huang, J. & Chang, Y. 2009. Frequency analysis of the words in the Academic Word List (AWL) and non-AWL content words in applied linguistics research papers. *English for Specific Purposes* 28: 33-41.

Welch, T. & Reed, Y. 2005. *Designing and delivering distance education: Quality criteria and case studies from South Africa.* Johannesburg: NADEOSA.

Wells, L.D. 2007. *Vocabulary mastery 1: Using and learning the academic word list.* Ann Arbor: University of Michigan Press.

West, M. 1953. *A general service list of English words.* London: Longman.

Williams, R. & Dallas, D. 1984. Aspects of vocabulary in the readability of content area textbooks: A case study. In Alderson, J.C. & Urquhart, A.H. *Reading in a foreign language.* New York: Longman: 199-210.

Williams, J.B. 2006. Assertion-reason multiple-choice testing as a tool for deep learning: A qualitative analysis. *Assessment and Evaluation in Higher Education* 31(3): 287-301.

Wolcott, H.F. 2009. *Writing up qualitative research.* (3 ed.) Thousand Oaks, California: Sage.

Xue, G. & Nation, I.S.P. 1984. A University Word List. *Language, Learning and Communication* 3 (2): 215-229.

Young, D.J. 1999. Linguistic simplification of SL reading material: Effective instructional practice. *Modern Language Journal* 83 (3): 350-362.

Appendix A

UNIVERSITY EXAMINATIONS UNIVERSITEITSEKSAMENS

UNISA university of south africa

LIN103-Y (469971) October/November 2006

LINGUISTICS 103

Duration : 2 Hours 80 Marks

EXAMINERS :
FIRST : PROF LM SWANEPOEL
SECOND : MRS BE ZAWADA

This paper consists of 19 pages plus instructions for the completion of a mark reading sheet and 4 pages for rough work.

This paper remains the property of the University of South Africa and may not be removed from the examination hall.

Please complete the attendance register on the back page, tear off and hand to the invigilator.

Instructions
Read these instructions as well as the instructions on the mark reading sheet **very carefully** before you start.

1. All the questions in this paper (80 in total) are multiple choice questions and are compulsory.
2. Use **only** the mark reading sheet issued to you by the invigilator to answer the questions.
3. Read each question and **all** the options before you answer the question.
4. Choose the **most appropriate** (i.e., the most correct or the most complete) option.
5. Mark **one and only one** option for each question on the mark reading sheet.
6. Read the instructions on the mark reading sheet **very carefully** before you start marking the sheet.
7. You have to hand in the exam paper together with the mark reading sheet to the invigilator.

File produced at level 10 using AWL Highlighter http://www.nottingham.ac.uk/%7Ealzsh3/acvocab/awlhighlighter.htm on 12 February 2007

Read the following case study and then answer Questions 1 to 9. Some of the questions relate directly to the case study, while others test your knowledge of the **relevant concepts** in a more general way.

The following two sentences were produced by an **adult.** Four children were then asked to repeat these sentences. The children are all aged between 25 and 32 months. Their utterances are given below:

Adult:	*It goes in a big box.*
Child 1:	*Big box.*
Child 2:	*Big box.*
Child 3:	*In big box.*
Child 4:	*It goes in box.*
Adult:	*Is it a car?*
Child 1:	*'t car?*
Child 2:	*Is it car?*
Child 3:	*Car?*
Child 4:	*That a car?*

1. Child 4 in the **data** above is at

   [1] the cooing stage
   [2] the babbling stage
   [3] the one word stage
   [4] the two-word stage
   [5] the telegraphic speech stage.

2. Most children reach the stage of language development of Child 4 above at **approximately**

   [1] 3 months old
   [2] 6 months old
   [3] 1 year old
   [4] 2 years old
   [5] 2 ½ years old.

3. Which of the questions below can be classified as a holophrase?

   [1] Child 1: *'t car?*
   [2] Child 2: *Is it car?*
   [3] Child 3: *Car?*
   [4] Child 4: *That a car?*

4. In the **data** above, the children are producing

   [1] mainly three-word sentences
   [2] incomprehensible sentences
   [3] egocentric speech
   [4] short sentences containing mostly content words
   [5] short sentences containing mostly **function** words.

5. The **data** above **indicates** that children cannot **accurately** reproduce sentences that are above their current level of language development. This kind of **evidence** suggests that

   [1] children are born with innate knowledge of the vocabulary of their language
   [2] children learn language by imitation
   [3] children do not learn language simply by imitation
   [4] children imitate **adults'** language **errors.**

6. The unconscious, informal **process** of 'picking up' a language in the pre-adolescent years is known as

   [1] developmental psycholinguistics
   [2] language **acquisition**
   [3] language learning
   [4] the critical **period**
   [5] additive bilingualism.

7. A child starts to use utterances for **communication** in a meaningful and intentional way during

   [1] the cooing stage
   [2] the babbling stage
   [3] the one-word stage
   [4] the two-word stage
   [5] the multiple-word stage.

8. The first words children **acquire** tend to be

   [1] words referring to objects and actions
   [2] **function** words
   [3] words referring to **abstract** things
   [4] [1] and [2]

9. Which of the following statements is false?

   [1] Language **acquisition** follows **similar** developmental stages in all children irrespective of the language they **acquire.**
   [2] All children **acquire** the language to which they have been **exposed.**
   [3] A child has the **potential** to **acquire** any language.
   [4] **Exposure** to language is not necessary for language **acquisition.**

Read the following case study and then answer Questions 10 to 15.

Say byebye to the aeroplane. Byebye aeroplane. Byebye. Look at the aeroplane. What's it doing? What's the aeroplane doing? It's flying. Up up into the sky. Byebye aeroplane. That's right. Wave goodbye to the aeroplane.

10. The language sample above **illustrates**

    [1] a second language teacher talking to her class
    [2] a small child talking to his or her mother
    [3] a small child engaging in egocentric speech
    [4] a father talking to a small child.

11. Which of the following is a typical characteristic of caretaker speech?

    [1] flat, unchanging intonation
    [2] short, ungrammatical sentences
    [3] frequent repetition
    [4] long sentences

12. An example of a babytalk word in the **data** above is

    [1] *aeroplane*
    [2] *byebye*
    [3] *look*
    [4] *sky*.

13. An example of a command in the **data** above is

[1] Say byebye to the aeroplane.
[2] What's it doing?
[3] Up up into the sky.
[4] That's right.

14. A child who uses the word aeroplane to refer to aeroplanes and helicopters is

[1] overextending the word aeroplane
[2] overextending the word helicopter
[3] underextending the word aeroplane
[4] underextending the word helicopter.

15. The term ‘here and now' means that caretakers tend to talk mainly about

[1] things that happened in the past
[2] objects and activities in the immediate **environment**
[3] activities that will take place in the future
[4] **abstract concepts.**

Read the following case study and then answer Questions 16 to 24. Some of the questions relate directly to the case study, while others test your knowledge of the **relevant concepts** in a more general way.

Jane and Einar are a **couple** who live in England with their two children Marianne (12) and Erik (9). Jane speaks English as L1 and Einar is fluent in both Norwegian (his L1) and English (his L2), which he learnt while at university in England. Jane tried to learn Norwegian after they met, but could not carry out a discussion satisfactorily, even after a six-month stay in Norway. In fact, with Einar's parents, Jane got into the habit of speaking English while they addressed her in Norwegian.

Both Jane and Einar speak English to their children and the children are thus monolingual English speakers. One of the reasons for their decision was that there was little advantage in teaching their children Norwegian as it was not a **major** world language. Since most people in Norway are bilingual in English and Norwegian, the children were still able to get to know their relatives in Norway, except for one grandfather who spoke only Norwegian. Einar and Jane make all sorts of efforts to encourage the children to keep in touch with their Norwegian heritage. They read translated Norwegian books, go on holiday to Norway every year and **maintain** strong **cultural links** with the country.

(based on Harding E & Riley P 1986. *The Bilingual Family: A handbook for parents.* Cambridge: Cambridge University Press.)

16. Which of the following statements is correct?

[1] The family above can be described as bilingual.
[2] The family above can be described as semilingual.
[3] The family above can be decribed as semispeakers.
[4] The family above can be described as rememberers.

17. Jane's **motivation** for trying to learn Norwegian when she met Einar was

[1] **individual**
[2] instrumental
[3] independent
[4] **interactive**
[5] integrative.

18. By speaking English when her parents-in-law address her in Norwegian, Jane is adopting a

[1] convergent **strategy**
[2] divergent **strategy**
[3] bilingual **strategy**
[4] one-person-one-language **strategy.**

19. Which of the following is the best way for Jane to improve her proficiency in Norwegian?

[1] Read Norwegian books that have been translated into English.
[2] Make more effort to speak the language when she visits Norway.
[3] Make more effort to listen to the language when she visits Norway.
[4] Make more effort to read the language when she visits Norway.

20. In South Africa, Norwegian would be considered a

[1] national language
[2] **dominant** language
[3] dying language
[4] foreign language.

21. The ‘language **policy'** of the family can be described as

[1] a one-person-one language **policy**
[2] a two-person-two-languages **policy**
[3] a monolingual **policy**
[4] diglossia.

22. Which family member is a subtractive bilingual?

[1] Einar
[2] Jane
[3] Marianne
[4] the grandfather
[5] None of the above.

23. Which family member is/was an early bilingual?

[1] Einar
[2] Jane
[3] Marianne
[4] the grandfather
[5] None of the above.

24. In parts of Norway, people use the standard language Bokmål in formal situations such as education and official transactions, and Ranamål for informal conversations relating to family affairs and everyday events. This is an example of

[1] territorial monolingualism
[2] territorial multilingualism
[3] diglossia
[4] **minority** group bilingualism.

Read the following case study and then answer Questions 25 to 35:

An English **couple** moved to France with their children, Sam (11) and Pam (9). None of the family knew any French before they moved. Within a few weeks of their arrival, the two children had picked up an **enormous** amount of French at school, while the parents still had a lot of trouble making themselves understood. The **couple** told the following anecdote: 'There was another child in the same age group living in the same block of flats and our children started playing together. A few days later, this child's mother appeared at the door to complain that our children were using certain unacceptable words. We **found** ourselves in the awful position of having to ask her what these words were. Gros mots (swearwords), she replied. 'Yes but what are the gros mots of your otherwise so beautiful a language?' By now, faces were scarlet on both sides. But then, taking her courage in both hands, the lady stepped inside our front door, which she closed carefully behind her, and told us.'

(based on Harding E & Riley P 1986. The Bilingual Family: A handbook for parents. Cambridge: Cambridge University Press.)

25. For the children in the case study above, their main **motivation** for learning French is

[1] instrumental
[2] integrative
[3] interference
[4] **internal**.

26. Which of the following statements is false?

[1] L2 proficiency is directly related to the number of years of L2 study.
[2] Learners **benefit** from a silent **period** where they listen to L2 but are not **required** to speak it.
[3] L2 teachers can help students by using gestures, pictures and **contextual** clues to **clarify aspects** of the language.
[4] The younger a person is when he or she starts to learn L2, the more likely it is that he or she will **attain** near-native competence.

27. Which of the following statements is true?

[1] A good L2 learner is shy and introverted.
[2] A good L2 learner has a **negative attitude** to the language and it speakers.
[3] A good L2 learner **focuses** on the meaning of utterances rather than the form.
[4] A good L2 learner never makes mistakes.

28. The critical age **hypothesis** states that

[1] children are better second-language learners than **adults**
[2] the younger a person is when he or she starts to learn L2, the more likely it is that he or she will **attain** near-native competence.
[3] there is a **specific period** for first language **acquisition,** after this **period** it is difficult and maybe even impossible to **acquire** language
[4] in order to become bilingual, children need special **instruction** during a **specific ‘**critical' **period**
[5] there is a **period** between birth and puberty when a second language is usually **acquired.**

29. Sam and Pam's parents can be classified as
[1] monolinguals
[2] semilinguals
[3] early bilinguals
[4] late bilinguals.

30. Sam and Pam can be classified as

[1] monolinguals
[2] semilinguals
[3] additive bilinguals
[4] subtractive bilinguals.

31. By using certain unacceptable words, Sam and Pam were breaking the unspoken conversational rules of the society. What happens when a L2 speaker unintentionally breaks these ‘rules'?

[1] Other members of the society may be upset, offended or embarrassed.
[2] Other members of the society may be surprised or amused.
[3] Other members of the society may **perceive** the speaker **negatively.**
[4] Other members of the society may realise that the person is an L2 speaker of the language.
[5] Any of the above.

32. The avoidance of certain words and/or conversational **topics** is known as

[1] swearing
[2] euphemism
[3] hlonipha
[4] taboo.

33. Which of the following exclamations is not an example of euphemism?

[1] It's boiling hot today
[2] It is blooming hot today
[3] It's frigging hot today
[4] It's flipping hot today.

34. Sam and Pam now speak their L1, English, at home but speak French at school and when playing with friends. This is an example of

[1] partial language **shift**
[2] total language **shift**
[3] gradual death
[4] sudden death.

35. Since French is the **medium** of **instruction** at school in all classes except the English class, Sam and Pam will now experience

[1] mother-tongue education
[2] a dual-language programme
[3] an immersion programme
[4] a submersion programme.

Read the following sociolinguistic profile of Tanzania and then answer Questions 36 to 44:

Sociolinguistic profile of Tanzania

Population: 29 million
Languages spoken: 135-150 different languages
Kisukuma is spoken as a first language by 12.5% of the population
Kiswahili is spoken as a first language by 10% but known by 90% of the population
Kinyambwezi is spoken as a first language by 4.2% of the population
English is known (mostly as L2 or L3) by 20% of the population
Official language: Kiswahili
Language of learning & teaching: Kiswahili at **primary** school, English at secondary and tertiary level
Literacy level: 68%
Media: Printed and radio **media** are **available** in both English and Kiswahili

36. Which language has the most L1 speakers in Tanzania?

[1] Kisukuma
[2] Kiswahili
[3] Kinyambwezi
[4] English
[5] None of the above

37. Which language has the most L2 speakers in Tanzania?

[1] Kisukuma
[2] Kiswahili
[3] Kinyambwezi
[4] English
[5] None of the above

38. Which language is a foreign language in Tanzania?

[1] Kisukuma
[2] Kiswahili
[3] Kinyambwezi
[4] English
[5] None of the above

39. Which language would fulfil the **function** of a national language in Tanzania?

[1] Kisukuma
[2] Kiswahili
[3] Kinyambwezi
[4] English
[5] None of the above.

40. Monolingual Kiswahili speakers cannot understand monolingual Kisukuma speakers. These can be considered

[1] varieties of the same language
[2] two different ethnolects
[3] dialects of the same language
[4] **mutually** intelligible varieties
[5] two different languages.

41. The term used for the linguistic situation in a country like Tanzania where most **individuals** are bilingual is

[1] territorial monolingualism
[2] territorial multilingualism
[3] total bilingualism
[4] **unstable** bilingualism
[5] diglossia.

42. What **percentage** of Tanzanians receive mother-tongue education at **primary** school?

[1] 4.2%
[2] 10%
[3] 12.5%
[4] 20%
[5] 90%

43. Government **policy** relating to the use of the various languages in a country is known as

[1] diglossia
[2] language **shift**
[3] language planning
[4] governmental linguistics
[5] None of the above.

44. In Tanzania, one disadvantage of choosing English as a language of learning and teaching at higher levels would be that

[1] English is an international language
[2] most **sectors** of the population have a **positive attitude** to English
[3] it would **benefit** L1 English speakers at the expense of other language groups
[4] English textbooks are readily **available**

Read the following case study and then answer Questions 45 to 55:

Xhosa is an African language that is spoken all over South Africa, but particularly in the Eastern Cape and Western Cape **regions.** Xhosa belongs to the Nguni language family and is thus related to Zulu. Xhosa and Zulu are **similar** enough to be **mutually** intelligible. However, the history of Southern Africa has **emphasised** the **distinctions** between the two **communities** - Xhosa and Zulu have different writing and spelling systems, developed by different missionary societies in the two **communities,** and the **policies** of the colonial and apartheid governments **emphasised** political and **cultural** differences by setting up different geographical **areas** for different language groups. Different geographical **regions** use slightly different varieties of Xhosa, for example, Xhosa speakers from Middelburg use the terms ipiringi ‘saucer’ and ikamire ‘room’, while Xhosa speakers from Cradock use isosara ‘saucer' and irum ‘room'. **Despite** these differences in vocabulary and accent, Xhosa speakers have no difficulty understanding each other.

45. Zulu and Xhosa are considered separate languages because

[1] they are **mutually** intelligible
[2] they are not **mutually** intelligible
[3] their speakers are **culturally** and politically **distinct**
[4] they are completely unrelated.

46. A language variety that is associated with a particular **section** of society such as the middle class or the working class is known as

[1] a sociolect
[2] a dialect
[3] an idiolect
[4] jargon.

47. Which of the following offers the best **definition** of the sociolinguistic term dialect?

[1] Dialects are **mutually** intelligible forms of different languages.
[2] A dialect is a substandard, low **status,** often rustic form of language.
[3] Dialects are language varieties associated with particular geographical **areas.**
[4] Dialects are language varieties associated with particular social classes.
[5] The term ‘dialect' refers to languages that have no written form.

48. Which of the following statements is true of accents?

[1] Speakers' accents may **reveal** where they grew up.
[2] Speakers' accents may **reveal** that they are speaking a language other than their native language.
[3] Both [1] and [2] are true.
[4] None of the above.

49. Which of the following statements is false?

[1] Some dialects are better than others.
[2] Some dialects have a higher **status** than others.
[3] Any dialect can be raised to language **status** if its speakers have **sufficient economic** and/or **military** power.
[4] The standard language of a nation can be spread **via** the education system.

50. A language variety associated with a particular race group is known as

[1] an idiolect
[2] a dialect
[3] a sociolect
[4] an ethnolect.

51. The Xhosa used in Middelburg and Cradock can be classified as two different

[1] idiolects
[2] dialects
[3] sociolects
[4] ethnolects
[5] languages.

52. The Xhosa terms *isosara* ‘saucer' and *irum* ‘room' are examples of

[1] borrowing
[2] codeswitching
[3] interference
[4] convergence.

53. One advantage of choosing Xhosa as South Africa's only official language would be that

[1] Xhosa is an international language
[2] it could lead to **tension** between **ethnic** groups
[3] it would **benefit** mother-tongue Nguni speakers at the expense of other language groups
[4] it would be cheaper than an 11-language **policy.**

54. Many Xhosa children learn through the **medium** of Xhosa for the first four years of school and switch to English as **medium** of **instruction** from **Grade** 5, with Xhosa as a school subject. This is an example of

[1] mother-tongue education
[2] an immersion programme
[3] a submersion programme
[4] a dual-language programme
[5] a **transitional** programme.

55. According to the custom of hlonipha, a married woman in **traditional** Xhosa **culture** must avoid saying

[1] her husband's name
[2] her husband's name and the names of his family
[3] her husband's name, the names of his family and all other words that start with the same letters
[4] her husband's name, the names of his family and all other words that contain these syllables.

Read the following example and then answer Questions 56 to 58:

Nomsa is talking to her friend Bulelwa. Both girls speak Xhosa as L1 and English as L2.

Whenever sisiye etown you like ukuthetha nabantu.
'Whenever we go to town you like to talk to people.'

56. The linguistic **phenomenon illustrated** above is known as

[1] borrowing
[2] codeswitching
[3] **positive transfer**
[4] **negative transfer**
[5] a learner variety.

57. Which of the following statements is false?

[1] The linguistic **phenomenon illustrated** above **involves** the use of two or more languages in the same conversation.
[2] The linguistic **phenomenon illustrated** above is a way for speakers to express their common **identity** as bilinguals.
[3] The linguistic **phenomenon illustrated** above is a sign of laziness and linguistic decay.
[4] The linguistic **phenomenon illustrated** above is often **linked** to a change in **topic.**

58. The girls in the example above are probably

[1] balanced bilinguals
[2] semilinguals
[3] monolinguals
[4] using a divergent **strategy** to **stress** the differences between them.

Read the following case study and then answer Questions 59 to 68:

I
n the early 1930s, the small **community** of Yaaku people in Kenya decided not to teach their children the Yaaku language. This decision was taken because the **migration** of Maasai to the **area** had caused a change from a hunter-gatherer **economy** to one based on livestock such as goats and cattle. The Maasai language, Mukogodo, was a prestigious language and a language of wider **communication,** with a large number of speakers, many of whom were wealthy cattle owners. Yaaku could not be used for **economic** and social **contacts** with others and so bilingualism and interethnic marriage became common. By 1971 there were only 51 Yaaku speakers, all of whom were bilingual in Mukogodo. Only 10% of the people under 40 had a reasonable command of Yaaku. By 1989, only 10 Yaaku were left. All were over 70 years old and had very limited competence in Yaaku.

(adapted from Brenzinger M 1992 'Patterns of language **shift** in East Africa.' In Herbert RK (ed.) Language and Society in Africa)

59. Which of the following is the most **appropriate** description of Yaaku in 1989?
[1] a **dominant** language
[2] a dead language
[3] a dying language
[4] a replacing language
[5] an ethnolect.

60. The 10 Yaaku speakers remaining in 1989 could be classified as

[1] rememberers
[2] forgetters
[3] semi-speakers
[4] semilinguals
[5] monolinguals.

61. Which of the following does not generally **contribute** to language **shift?**

[1] the **structure** of the language
[2] **attitudes** towards the language
[3] increased **contact** between languages
[4] persecution of speakers of the language
[5] **migration** of speakers to urban **areas.**

62. What is a semi speaker?

[1] Someone with a speech disability.
[2] Someone who knows a language imperfectly.
[3] Someone who was once fluent but now remembers very little of the language.
[4] Someone who is learning a new language.

63. Urbanisation can lead to language **shift** by

[1] causing close **contact** with other language groups
[2] separating speakers from their L1 speech **community**
[3] lessening the usefulness of **minority** languages
[4] All of the above.

64. A replacing language is usually

[1] a **minority** language
[2] a grammatically simple language
[3] an endangered language
[4] a socially and **economically** useful language.

65. Language **shift occurs** when

[1] monolinguals become bilingual
[2] bilingual speakers codeswitch
[3] speakers use their L2 in situations where they used to use the L1
[4] a **minority** language becomes used more frequently.

66. Which of the following does not **contribute** to language **shift?**

[1] A failure to pass a language on to one's children
[2] Industrialisation and urbanisation
[3] **Isolation** from other language groups
[4] When speakers **cease** to **evaluate** their own language **positively.**

67. Halting the **process** of language death is possible if

[1] the **community** value and continue to use their **traditional** language
[2] the language is used in the mass **media**
[3] the government is **committed** to preserving **minority** languages
[4] All of the above.

68. Which of the following is not characteristic of a dying language?

[1] Its lexicon shrinks.
[2] It becomes grammatically more **complex.**
[3] It becomes used in fewer **domains** than before.
[4] It borrows heavily from the **dominant** language.
[5] All of the above.

Read the following case study and then answer Questions 69 to 74:

Simon is a 10-year-old Deaf child of hearing parents. At age 6 he started attending a school for Deaf children. The school discourages the children from using sign language, but Simon **nevertheless** learnt South African Sign Language (SASL) from **interacting** with the other children in the playground. The school uses English as the **medium** of **instruction.** Children are taught to lipread with **supplementary** hand signals used in the class to signal differences between sounds like /t/, /d/ and /n/ which look **similar** when they are pronounced.

69. Simon's school has adopted _____ as the language **policy** of the classroom.

[1] total **communication**
[2] oralism
[3] South African Sign Language
[4] pidgin sign language

70. The **supplementary** hand signals described above are known as

[1] finger spelling
[2] cued speech
[3] pidgin sign language
[4] a **manual** sign **code.**

71. Simon learnt sign language

[1] spontaneously
[2] through guided learning
[3] after the critical **period**
[4] as a foreign language.
[5] Both [1] and [3] are correct.

72. Which of the following statements is false?

[1] Delayed language **acquisition** may have **negative** effects on the linguistic and social development of children.
[2] The language **policy** at Simon's school is the most **beneficial** one for Deaf children.
[3] Deaf children's babbling tends to be non-repetitive and **random.**
[4] Early **exposure** to sign language is **beneficial** to Deaf children.

73. Which of the following is an articulatory **feature** of sign language signs?

[1] facial expression
[2] palm **orientation**
[3] **location** of the sign relative to the body
[4] [2] and [3] are correct
[5] [1], [2] and [3] are correct

74. SASL is

[1] based on South African English
[2] a pidgin sign language
[3] a natural sign language
[4] a **manual** sign **code**
[5] related to Zulu.

Read the following conversation and then answer Questions 75 to 80:

A (from New Zealand): *That's a really gorgeous necklace you're wearing. Those stones are the most beautiful deep indigo blue.*
B (from Samoa): *Here, take it.*

75. Judging by the conversational **style,** what **gender** is speaker A?

[1] Male, because of the choice of adjectives and colour terms.
[2] Male, because direct commands are used.
[3] Female, because of the choice of adjectives and colour terms.
[4] Female, because euphemisms are used.

76. Which of the following statements is true?

[1] Conversations between women tend to **focus** on a few **topics** but in more detail than men's conversations.
[2] Conversations between men tend to **focus** on a few **topics** but in more detail than women's conversations.
[3] In all-female conversations the **topics** tend to have an **external focus.**
[4] Men see conversation **primarily** as a way of building connections and relationships.

77. Which of the following is not typically influenced by **gender?**

[1] the pitch or deepness of the voice
[2] word choice
[3] sentence length
[4] the frequency with which euphemisms are used
[5] conversational **style**

78. Which of the following statements is false?

[1] Conversational rules and patterns are universal.
[2] **Communicative** competence can be **defined** as our knowledge of socially **appropriate** speech behaviour.
[3] Misunderstandings can result when speakers from different **cultures interact.**
[4] Different languages reflect different **cultural** worldviews.

79. Which of the following **aspects** of complimenting behaviour differs from society to society?

[1] the frequency with which compliments are given
[2] the **topics** on which it is considered **appropriate** to compliment
[3] the **appropriate response** to a compliment
[4] who usually compliments whom
[5] All of the above.

80. The most **appropriate response** to a compliment is

[1] a simple ‘thank you'
[2] to give a compliment in return
[3] to **reject** the compliment by **denying** that it is true
[4] to give the person the **item** on which you have been complimented
[5] dependent on the **cultural context.**

TOTAL: 80

Appendix B

UNIVERSITY EXAMINATIONS UNIVERSITEITSEKSAMENS

UNISA university of south africa

LIN103-Y (479349) October/November 2007

LINGUISTICS 103

Duration : 2 Hours 80 Marks

EXAMINERS :
FIRST : MRS PJ SANDERSON
SECOND : MRS DT NKWE

This paper consists of 17 pages plus instructions for the completion of a mark reading sheet and 4 pages for rough work.

This paper remains the property of the University of South Africa and may not be removed from the examination hall.

Please complete the attendance register on the back page, tear off and hand to the invigilator.

Instructions

1. All the questions in this paper (80 in total) are multiple choice questions and are compulsory.
2. Use **only** the mark reading sheet issued to you by the invigilator to answer the questions.
3. Read each question and **all** the options before you answer the question.
4. Choose the **most appropriate** (i.e., the most correct or the most complete) option.
5. Mark **one and only one** option for each question on the mark reading sheet.
6. Read the instructions on the mark reading sheet **very carefully** before you start marking the sheet.
7. You have to hand in the exam paper together with the mark reading sheet to the invigilator.

File produced at level 10 using http://www.nottingham.ac.uk/%7Ealzsh3/acvocab/awlhighlighter.htm on 16 March 2009

Read the following case study and then answer Questions 1 to 5.

> Susan is a six-year-old child who has just moved from South Africa to Botswana. Susan's father speaks Afrikaans (his L1) to her and her mother speaks English (her L1) to her. Susan speaks these two languages equally well. She is now venturing out to play with neighbouring children who only speak Tswana and she is beginning to pick up Tswana phrases in order to make friends.

1. Susan's family have adopted a

   [1] one-person-two-language **strategy**
   [2] two-person-one-language **strategy**
   [3] two-person-two-language **strategy**
   [4] one-person-one-language **strategy**

2. Susan's motivation for learning Tswana is an

   [1] instrumental **motivation**
   [2] integrative **motivation**
   [3] **internal motivation**
   [4] **individual motivation**
   [5] idiosyncratic **motivation**

3. Susan is a

   [1] monolingual
   [2] subtractive bilingual
   [3] balanced bilingual
   [4] late bilingual

4. Susan will learn Tswana by means of

   [1] spontaneous language learning
   [2] guided language learning
   [3] foreign language learning
   [4] third language **acquisition**

5. Which of the following statements is false?

[1] Susan is likely to **attain** greater proficiency in Tswana than either of her parents.
[2] Tswana is a foreign language for Susan, since she is originally South African.
[3] Susan will **acquire** Tswana effortlessly and unconsciously as she is below the critical age.
[4] Susan is likely to experience interference from English and Afrikaans when speaking Tswana.

Read the following case study and then answer Questions 6 to 8.

> Mr Dlamini is a businessman who speaks Zulu as L1 and English as L2. He attended a school in Soweto where Zulu was the **medium** of **instruction**. Now he is learning German through Unisa in order to conduct international business transactions.

6. Mr Dlamini is learning German **primarily** for

[1] instrumental reasons
[2] integrative reasons
[3] interference reasons
[4] **internal** reasons
[5] idiosyncratic reasons.

7. In Mr Dlamini's case, learning German **involves**

[1] first language **acquisition**
[2] second language learning
[3] spontaneous language learning
[4] foreign language learning
[5] third language **acquisition**.

8. Mr Dlamini's schooling can be described as

[1] mother-tongue education
[2] an immersion programme
[3] a submersion programme
[4] a **transitional** programme.

Read the following case study and then answer Questions 9 to 15. Some of the questions relate directly to the case study while others test your knowledge of the **relevant concepts** in a more general way:

> Kim is a English-speaking university student from South Africa who is about to visit some of Tanzania's famous national parks on a two-week holiday. To make it easier to get around, she decides to learn Kiswahili before she goes. Kim buys herself a Teach-yourself-Kiswahili book and studies the vocabulary and grammar of the language in her spare time.

9. Kim's reasons for learning Kiswahili are **primarily**

   [1] instrumental reasons
   [2] integrative reasons
   [3] interference reasons
   [4] **internal** reasons.

10. For Kim, learning Kiswahili would be considered as

   [1] first language **acquisition**
   [2] spontaneous language learning
   [3] guided language learning
   [4] foreign language learning
   [5] [3] and [4].

11. Which of the following statements is true?

   [1] A good L2 learner is shy and introverted.
   [2] A good L2 learner has a **negative attitude** to the language and it speakers.
   [3] A good L2 learner **focuses** on the meaning of utterances rather than the form.
   [4] A good L2 learner is unwilling to try out the language and make mistakes.

12. Which of the following statements is false?

   [1] Kim will use a learner variety of Kiswahili which may differ grammatically from standard Kiswahili in various ways
   [2] Kiswahili is a foreign language for Kim
   [3] Kim will **acquire** Kiswahili effortlessly and unconsciously as she is below the critical age
   [4] Kim is likely to experience interference from English when speaking Kiswahili.

13. Kim's accent will signal that she is an L2 speaker of Kiswahili. The term accent refers to **distinctive**

    [1] idiomatic phrases
    [2] vocabulary **items**
    [3] grammatical differences
    [4] pronunciation
    [5] All of the above.

14. Which of the following statements is false?

    [1] Kim can be considered a semi-speaker of Kiswahili.
    [2] Kim is an additive bilingual.
    [3] **Adult** L2 learners progress faster than children because of their greater **concentration** spans.
    [4] The earlier one starts to learn L2, the more likely it is that one will **attain** near-native competence.

15. Kim is learning standard Kiswahili. A standard language is

    [1] any dialect that has been reduced to writing and has dictionaries and grammar books determining 'correct' usage
    [2] used in more or less the same way by all its speakers
    [3] a prestigious variety within a society
    [4] All of the above.

Read the following case study, then answer Questions 16 to 17:

In the Arab world, **classical** Arabic has coexisted with a **regional**, colloquial variety of Arabic for centuries. Each of these varieties has a different social **function**: **classical** Arabic is learnt at school and used for religious and literary purposes, while colloquial Arabic is the language used for everyday conversation.

16. The situation described above is known as

    [1] **stable** bilingualism
    [2] territorial monolingualsm
    [3] territorial multilingualism
    [4] diglossia
    [5] a dual-language programme.

17. In the Arabic situation above, **classical** Arabic is known as the ----- variety.

[1] **dominant**
[2] **minority**
[3] standard
[4] high
[5] low

Read the following case study and then answer Questions 18 to 33. Some of the questions relate directly to the case study while others test your knowledge of the **relevant concepts** in a more general way:

Case study: Murchison

Murchison is a rural village in Kwa-Zulu Natal, South Africa. With a population of **approximately** 4000, it is characterised by long-standing poverty and unemployment. Although this **community** was formerly monolingual in Zulu, Zulu-English bilingualism is becoming increasingly common, especially among young people. In Murchison, Zulu is the home language in almost all homes and is the **major medium** of **instruction** in the schools. Parents who wish their children to be fluent in English therefore have to enrol them in urban schools in spite of the substantially higher school **fees**. In 1999 **approximately** 570 pupils from Murchison were attending English-medium schools in neighbouring towns. **Research** into the language choices of Murchison pupils attending one of the English-medium schools has **indicated** that male pupils use English more frequently than female pupils as can be seen from the table below:

	Female pupils' use of Zulu	Male pupils' use of Zulu
Home	83%	67%
Neighbour	94%	33%
Parents	83%	40%
Shopkeeper	89%	33%
Doctor	22%	13%

This **gender** difference in the use of Zulu appears to be related to broader social **issues**. English is seen as the high **status** language in South Africa, and proficiency in English is a way for young men to **achieve** higher **status** within the **community**. Proficiency in English also increases job

opportunities and it is the men who are expected to provide for the family by finding a **job**. Males typically move out into urban **areas** in search of **jobs** in mining, agriculture, manufacturing, **transport** and education. Females are expected to continue speaking **predominantly** in Zulu as they remain within the **community** to a far greater extent, as mothers and as subsistence farmers. The **role** of preserving Zulu **culture** is **assigned** to females as they are responsible for the **primary** socialisation of children.

(based on D. Appalraju & E. de Kadt 2002 '**Gender aspects** of bilingualism: language choice patterns of Zulu-speaking rural youth'. *Southern African Linguistics and Applied Language Studies* 2002 20: 135-145.)

18. The L1 of the Murchison pupils discussed in the case study is

[1] English
[2] Zulu
[3] both English and Zulu
[4] the same as their L2.

19. The L2 of the Murchison pupils discussed in the case study is

[1] English
[2] Zulu
[3] both English and Zulu
[4] a foreign language.

20. The syllable-avoidance practised by married women in **traditional** Zulu society is known as

[1] euphemism
[2] taboo
[3] umfazi
[4] hlonipha.

21. Which of the following statements is false?

[1] L2 proficiency is directly related to the number of years of L2 study.
[2] Learners may **benefit** from a silent **period** where they listen to L2 but are not **required** to speak it.
[3] L2 teachers can help students by using gestures, pictures and **contextual** clues to **clarify aspects** of the language.
[4] All of the above.

22. The education of the Murchison pupils discussed in the case study can be described as

[1] foreign language education
[2] mother-tongue education
[3] an immersion programme
[4] a dual-language programme.

23. The Murchison pupils discussed in the case study are

[1] monolinguals
[2] semilinguals
[3] additive bilinguals
[4] subtractive bilinguals.

24. The **motivation** for Murchison pupils to learn English is **primarily** an

[1] instrumental **motivation**
[2] integrative **motivation**
[3] **internal motivation**
[4] **individual motivation**
[5] idiosyncratic **motivation**.

25. Which of the following statements is true?

[1] The boys who were interviewed speak Zulu at home more often than the girls.
[2] The girls who were interviewed are monolingual Zulu speakers.
[3] The boys in Murchison are better language learners than girls.
[4] The girls in Murchison are better language learners than boys
[5] Language choice in Murchison is influenced by **gender**.

26. The pupils in Murchison often answer their Zulu-speaking parents in English. This is an example of

[1] borrowing
[2] codeswitching
[3] a convergent **strategy**
[4] a divergent **strategy**
[5] a one-person-one language policy.

27. Which of the following statements is true?

[1] English is a **minority** language in South Africa because it has fewer L1 speakers than languages such as Zulu.
[2] English is a **minority** language in South Africa because it is not associated with **economic** and political power.
[3] English is a **dominant** language in South Africa because it has more L1 speakers than any other language.
[4] English is a **dominant** language in South Africa because it is associated with **economic** and political power.

28. Younger members of the Murchison **community** tend to be

[1] rememberers
[2] semi-speakers
[3] bilingual
[4] monolingual in the replacing language
[5] Zulu monolinguals.

29. The oldest **generation** in Murchison are mostly

[1] rememberers
[2] semi-speakers
[3] bilingual
[4] monolingual in the replacing language
[5] Zulu monolinguals.

30. The Murchison situation is an example of

[1] sudden death
[2] gradual death
[3] partial language **shift**
[4] total language **shift**.

31. Which of the following is the most **appropriate** description of Zulu?

[1] a dead language
[2] a dying language
[3] a national language
[4] None of the above

32. The Zulu spoken in Murchison differs in several respects from the Zulu spoken in Johannesburg. The language variety as a whole could be characterised as a(n)

[1] learner variety
[2] idiolect
[3] sociolect
[4] dialect
[5] ethnolect

33. The Murchison pupils make frequent use of codeswitching. Which of the following statements is false?

1] Codeswitching **involves** the use of two or more languages in the same conversation.
[2] Codeswitching is a way for speakers to express their common **identity** as bilinguals.
[3] Codeswitching is when a bilingual **community** extends the use of a language to situations where another language was formerly used.
[4] Codeswitching is often **linked** to a change in **topic**.

Read the following case study and then answer Questions 34 to 48. Some of the questions relate directly to the case study while others test your knowledge of the **relevant concepts** in a more general way:

Case study: Genie

The abused 'wild child' Genie was 13 years old in 1970 when she was '**found'** by **welfare** workers. Her spontaneous utterances at that time included only *stopit* and *nomore*. She was **removed** from her parents' care, hospitalised and later placed into foster care. Transcriptions of some of Genie's utterances (marked as G) after two years of **intensive** language teaching and **exposure** are provided below.

1972

M *Where are you going?*
G *Going park.*

M *Where did we get those books?*
G *Library.*

G *I see elephant.*
M *Where did you see elephants?*
G *In the zoo.*

M *What did you do at the gym?*
G *Bounce on trampoline.*

34. ‘Wild children’ are

[1] children who are **physically** abused by their caretakers
[2] children who have grown up with little or no human **contact**
[3] children whose language is delayed due to behavioural problems
[4] children who cannot learn language due to severe **mental** retardation.

35. In 1972 Genie was at

[1] the cooing stage
[2] the babbling stage
[3] the one-word stage
[4] the two-word stage
[5] the multiple-word stage

36. At what age do most children reach this stage of language development?

[1] 6 months old
[2] 1 year old
[3] 2 years old
[4] sometime after 2 years
[5] At puberty.

37. Which of Genie’s utterances above can be considered a holophrase?

[1] *Going park*
[2] *Library.*
[3] *I see elephant.*
[4] *In the zoo.*
[5] *Bounce on trampoline.*

38. In the conversation about elephants Genie is **demonstrating**

[1] egocentric speech
[2] reduplicative babbling
[3] turntaking
[4] overextension
[5] underextension.

39. In the language samples above Genie’s utterances are all

[1] wh-questions
[2] yes-no questions
[3] statements
[4] requests.

40. Which of the following statements is true of the **data** given above?

[1] Genie has mastered the English past **tense** rule.
[2] Genie is overgeneralising the past **tense** morpheme *-ed*
[3] Genie is underextending the past **tense** morpheme *-ed*.
[4] Genie is omitting the past **tense** morpheme *-ed*

41. Genie's case was important for psycholinguistics because it shed light on the **hypothesis** known as the

[1] critical age **hypothesis**
[2] innateness **hypothesis**
[3] wild child **hypothesis**.

42. Psycholinguists who support the innateness hypothesis claim that

[1] babies do not have to be taught to sit up, crawl or walk
[2] humans have an inborn ability to **acquire** language
[3] language has to be formally taught to children
[4] children have an innate knowledge of the vocabulary of a language.

43. Which of the following statements is true?

[1] Language **acquisition** follows **similar** developmental stages in **normal** children and children who grow up without human **contact**.
[2] Children are born with a language **acquisition device** that allows them to **acquire** language without any language **input** from the **environment**.
[3] **Exposure** to language is necessary for language **acquisition**.
[4] **Explicit** language teaching by caretakers is necessary for language **acquisition**.

44. Genie's case shows that

[1] a child's first language is **acquired** easily after puberty
[2] it is impossible to **achieve** proficiency in a first language after puberty
[3] a second language cannot be successfully learned before puberty
[4] a second language cannot be successfully learned after puberty
[5] None of the above.

45. The unconscious, informal **process** by which **normal** children 'pick up' their mother tongues is known as

[1] developmental psycholinguistics
[2] language **acquisition**
[3] language learning
[4] language **shift**
[5] additive bilingualism.

46. After being reintroduced to society, Genie's language improved in the following way:

[1] She became a competent speaker within 2 years.
[2] Her pronunciation improved rapidly but her vocabulary and grammar remained limited.
[3] Her grammar improved rapidly but her vocabulary remained limited.
[4] Her vocabulary improved rapidly but her grammar remained limited.
[5] Her language showed no improvement.

47. Which of the following statements is false?

[1] **Adults** understand more words than they actually use.
[2] A child often understands a word several months before he or she starts to use the word.
[3] Children can usually produce more words than they can understand.
[4] Children can usually understand more words than they can produce.

48. During the telegraphic stage children

[1] produce mainly three-word sentences
[2] cannot be easily understood by **adults**
[3] produce short sentences containing no content words
[4] produce short sentences containing no **function** words.

Read the following conversation between Brenda (aged 20 months) and her mother and then answer Questions 49 to 55. Some of the questions relate directly to the case study while others test your knowledge of the **relevant concepts** in a more general way:

Brenda:	*Mommy*
Mom:	*What's that? Hmm? Sausage. Yum-yum. You want? That's French fries. You want? Oh, can't eat from that. Hmm? Oh yes. That's for Daddy, yes.*
Brenda:	*hiding*
Mom:	*Hide? What's hiding?*
Brenda:	*balloon*
Mom:	*Oh, the balloon? Where? Where is it? Where is it?*
Brenda:	*hiding*.
Mom:	*Where? Can't you find it?*
Brenda:	*find.*

Scollon R. 1976. *Conversations with a one year old: A case study of the developmental **foundations** of syntax.* Honolulu: University Press of Hawaii.

49. Brenda's mom uses the word *yum-yum*. This is

[1] an example of a diminutive
[2] a baby-talk word
[3] an example of babbling
[4] [2] and [3] are correct.

50. Which of the following statements is false?

[1] Brenda's mother makes use of caretaker speech.
[2] Brenda's mother is talking about the 'here and now'.
[3] Brenda's mother uses frequent repetition.
[4] Brenda's mother uses frequent questions.
[5] Brenda's mother uses frequent commands.

51. Which of the following statements is false?

[1] Brenda is engaging in egocentric speech.
[2] Brenda has learnt the rule of turn-taking in conversation.
[3] Brenda's comprehension is more advanced than her production.
[4] Brenda is using holophrases.
[5] Brenda is in the linguistic stage of development.

52. Brenda is at

[1] the cooing stage
[2] the babbling stage
[3] the one-word stage
[4] the two-word stage
[5] the telegraphic speech stage.

53. Most children reach Brenda's stage of language development at **approximately**

[1] 6 months old
[2] 1 year old
[3] 2 years old
[4] 3 years old.

54. A child starts to use utterances for **communication** in a meaningful and intentional way during

[1] the cooing stage
[2] the babbling stage
[3] the one-word stage
[4] the two-word stage
[5] the multiple-word stage.

55. Brenda uses the word *balloon* to refer to balloons and kites. This is an example of

[1] underextension
[2] overextension
[3] babbling
[4] the one-word stage

Read the following case study and then answer Questions 56 to 64. Some of the questions relate directly to the case study while others test your knowledge of the **relevant concepts** in a more general way:

Xhosa is an African language that is spoken all over South Africa, but particularly in the Eastern Cape and Western Cape **regions.** Xhosa belongs to the Nguni language family and is thus related to Zulu. Xhosa and Zulu are **similar** enough to be **mutually** intelligible. However, the history of Southern Africa has **emphasised** the **distinctions** between the two **communities** - Xhosa and Zulu have different writing and spelling systems, developed by different missionary societies in the two **communities,** and the **policies** of the colonial and apartheid governments **emphasised** political and **cultural** differences by setting up different geographical **areas** for different language groups.

Different geographical **regions** use slightly different varieties of Xhosa, for example, Xhosa speakers from Middelburg use the terms *ipiringi* 'saucer' and *ikamire* 'room', while Xhosa speakers from Cradock use *isosara* 'saucer' and *irum* 'room'. **Despite** these differences in vocabulary and accent, Xhosa speakers have no difficulty understanding each other.

56. Zulu and Xhosa are considered separate languages because

[1] they are **mutually** intelligible
[2] they are not **mutually** intelligible
[3] their speakers are **culturally** and politically **distinct**
[4] they are completely unrelated.

57. A language variety that is associated with a particular **section** of society such as the middle class or the working class is known as

[1] a sociolect
[2] a dialect
[3] an idiolect
[4] jargon.

58. Which of the following statements is true of **accents**?

[1] Speakers' accents may **reveal** where they grew up.
[2] Speakers' accents may **reveal** that they are speaking a language other than their native language.
[3] Both [1] and [2] are true.
[4] None of the above.

59. Which of the following statements is false?

[1] Some dialects are better than others.
[2] Some dialects have a higher **status** than others.
[3] Any dialect can be raised to language **status** if its speakers have **sufficient economic** and/or **military** power.
[4] The standard language of a nation can be spread **via** the education system.

60. The Xhosa used in Middelburg and Cradock can be classified as two different

[1] idiolects
[2] dialects
[3] sociolects
[4] ethnolects
[5] languages.

61. The Xhosa terms *isosara* 'saucer' and *irum* 'room' are examples of

[1] borrowing
[2] codeswitching
[3] interference
[4] **accommodation**.

62. One advantage of choosing Xhosa as South Africa's only official language would be that

[1] Xhosa is an international language
[2] it could lead to **tension** between **ethnic** groups
[3] it would **benefit** mother-tongue Nguni speakers at the expense of other language groups
[4] it would be cheaper than an 11-language **policy**.

63. Many Xhosa children learn through the **medium** of Xhosa for the first four years of school and switch to English as **medium** of **instruction** from **Grade** 5, with Xhosa as a school subject. This is an example of

[1] mother-tongue education
[2] an immersion programme
[3] a submersion programme
[4] a dual-language programme
[5] a **transitional** programme.

64. Xhosa is one of the 11 official languages in South Africa. Top-down **policy** relating to the use of the various languages in a country is known as

[1] diglossia
[2] language **shift**
[3] language planning
[4] sociolinguistics.

Read the following case study and then answer Questions 65 to 72. Some of the questions relate directly to the case study while others test your knowledge of the **relevant concepts** in a more general way:

> In the early 1930s, the small **community** of Yaaku people in Kenya decided not to teach their children the Yaaku language. This decision was taken because the **migration** of Maasai to the **area** had caused a change from a hunter-gatherer **economy** to one based on livestock such as goats and cattle. The Maasai language, Mukogodo, was a prestigious language and a language of wider **communication**, with a large number of speakers, many of whom were wealthy cattle owners. Yaaku could not be used for **economic** and social **contacts** with others and so bilingualism and interethnic marriage became common. By 1971 there were only 51 Yaaku speakers, all of whom were bilingual in Mukogodo. Only 10% of the people under 40 had a reasonable command of Yaaku. By 1989, only 10 Yaaku were left. All were over 70 years old and had very limited competence in Yaaku.

65. Which of the following is the most **appropriate** description of Yaaku in 1989?

[1] a **dominant** language
[2] a dead language
[3] a dying language
[4] a replacing language
[5] an ethnolect.

66. The 10 Yaaku speakers remaining in 1989 could be classified as

[1] rememberers
[2] semispeakers
[3] semilinguals
[4] monolinguals.

67. Which of the following does not generally **contribute** to language **shift**?

[1] the **structure** of the language
[2] **attitudes** towards the language
[3] increased **contact** between languages
[4] persecution of speakers of the language
[5] **migration** of speakers to urban **areas**.

68. What is a semi-speaker?

[1] Someone with a speech disability.
[2] Someone who knows a language imperfectly.
[3] Someone who was once fluent but now remembers very little of the language.
[4] Someone who is learning a new language.

69. Urbanisation can lead to language **shift** by

[1] causing close **contact** with other language groups
[2] separating speakers from their L1 speech **community**
[3] lessening the usefulness of **minority** languages
[4] All of the above.

70. A replacing language is usually

[1] a **minority** language
[2] a grammatically simple language
[3] an endangered language
[4] a socially and **economically** useful language.

71. Language **shift** occurs when

[1] monolinguals become bilingual
[2] bilingual speakers codeswitch
[3] speakers use their L2 in situations where they used to use the L1
[4] a **minority** language becomes used more frequently.

72. Which of the following does not **contribute** to language **shift**?

[1] A failure to pass a language on to one's children
[2] Industrialisation and urbanisation
[3] **Isolation** from other language groups
[4] When speakers **cease** to **evaluate** their own language **positively**.

73. Which of the following is not characteristic of a dying language?

[1] Its lexicon shrinks.
[2] It becomes grammatically more **complex**.
[3] It becomes used in fewer **domains** than before.
[4] It borrows heavily from the **dominant** language.
[5] All of the above.

74. Which of the following is a good example of euphemism?

[1] *John is dead.*
[2] *John has kicked the bucket.*
[3] *John has gone to meet his Maker.*
[4] *John has pegged.*

75. Which of the following **methods** used in schools for Deaf children **involves** spoken language together with hand signals to show which consonants are being articulated by the speaker?

[1] natural sign language
[2] fingerspelling
[3] the oral **method**
[4] cued speech

76. Deaf babies who are **exposed** to sign language produce their first signs at about 6 months.

[1] This is earlier than hearing children's first words because the muscular development of the hands is faster than that of the vocal organs.
[2] This is earlier than hearing children's first words because sign language is grammatically simpler than spoken language.
[3] This is later than hearing children's first words because the muscular development of the hands is slower than that of the vocal organs.
[4] This is later than hearing children's first words because they have no spoken language **input**.

77. Twelve **mutually** intelligible varieties of South African Sign Language have been **identified**. These can be considered separate

[1] languages
[2] dialects
[3] idiolects.

78. The **function** of facial expression during signing is to

[1] express the speakers' mood or **attitude**
[2] replace **manual** signs
[3] express grammatical and linguistic **contrasts**
[4] mimic what the vocal organs are doing during speech.

79. Which of the following statements is true?
[1] SASL is a sociolect of South African English.
[2] SASL has no grammar.
[3] SASL signs are iconic.
[4] SASL is a visual-gestural language.

80. **Manual** sign **codes** and finger spelling are
[1] based on spoken language
[2] used when Deaf people **communicate** with each other
[3] used when hearing people **communicate** with each other
[4] natural sign languages.

TOTAL: 80

Appendix C

Consent form for think-aloud interview

I am willing to participate in research on Unisa students' experiences of multiple-choice assessment and for my views to be recorded on video or audiotape. All information in the final report will be anonymous and I will not be identified by name.

Name

Signature

Date

Appendix D

Questionnaire and think-aloud protocol

Student number **Name**

Date and time of interview Started MCQs at about and ended at

Course LIN103Y
L1
L2
L3
L4
Language of instruction at school

How long have you been studying at Unisa?

Other studies completed

What is your general opinion of MCQ as an assessment method?

1	2	3	4	5
Really dislike	Dislike	Neutral	Like	Really like

Why?

How difficult do you find MCQ as an assessment method?

1	2	3	4	5
Very difficult	Difficult	Neutral	Easy	Very easy

Why?

What was your opinion about this particular question paper (written X days ago?)

1	2	3	4	5
Very difficult	Difficult	Neutral	Easy	Very easy

Why?
Did you have enough time to finish the paper?
I want you to go through the paper as you would in an exam, talking me through the paper as you complete each question and explaining the thought processes that help you come to a decision about the right answer. For each question please mark the answer you think is correct and mention any problems you have with the question.