AN INVESTIGATION INTO THE SIGNIFICANCE OF LISTENING PROFICIENCY IN THE ASSESSMENT OF ACADEMIC LITERACY LEVELS AT STELLENBOSCH UNIVERSITY

FIONA C MARAIS

Submitted in partial fulfilment of the requirements

for the degree of

Master of Philosophy in Hypermedia for Language Learning

Department of Modern Foreign Languages

at

Stellenbosch University

SUPERVISOR: Ms E K Bergman

CO-SUPERVISOR: Mr T J van Dyk

March 2009

DECLARATION

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the owner of the copyright thereof (unless to the extent explicitly otherwise stated) and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Date: 3 March 2009

ACKNOWLEDGEMENTS

Firstly, I want to thank my mentor and friend, Dr Febe de Wet, without whose practical and moral support this thesis could not have been completed. Furthermore, I am extremely grateful to my supervisors, Ms Lesley Bergman and Mr Tobie van Dyk, for their knowledgeable guidance and assistance in this project.

The technical skills of Mr Charles Carstens and Mr Thys Murray provided invaluable aid in the operationalisation of ALT. Since statistical analysis was such an important aspect of this study, Professor Martin Kidd's assistance in this area is also greatly appreciated.

In addition, I want to extend my gratitude to the students and staff who kindly volunteered to participate in this experiment. Last but by no means least, thank you to my friends, colleagues and family particularly my mother, for her unfailing faith in me and my daughters, for their patience and encouragement.

ABSTRACT

Concern surrounding the low levels of academic literacy amongst incoming first year students has prompted universities and other tertiary education institutions in South Africa to implement tests of academic literacy. At Stellenbosch University, the English version of this test is known as TALL (Test of Academic Literacy Levels) and was developed to assess reading and writing abilities in an academic context. The results are used to 'stream' students into programmes which assist them in acquiring the various skills deemed necessary for their academic success. This study examines, on the one hand, the significance of listening in the assessment of academic literacy levels; on the other hand, it explores the potential for an academic listening test (ALT) to assist TALL in more accurate screening of students, particularly the borderline cases. The design and operationalisation of ALT is based on the theories and approaches of several researchers in the field. The study began with the compilation of the test specification and design of ALT. This was followed by empirically piloting a project where qualitative data concerning the validity of ALT was collected by means of a questionnaire. The next phase involved assessing the academic listening competency of a sample of first year university students. This assessment comprised an initial test administration followed by a second administration of the same test a month later in order to ascertain consistency of measurement over a period of time. The quantitative results obtained from both administrations were then statistically analysed to determine the reliability and validity of ALT. The final phase of the study involved the correlation of these results with those of TALL to determine the level of criterion-related validity as well as to establish whether ALT could be a useful added dimension to TALL.

OPSOMMING

Kommer oor die lae vlak van akademiese geletterdheid onder eerstejaarstudente het daartoe gelei dat universiteite en ander tersiêre instellings in Suid-Afrika begin het om akademiese geletterdheidstoetse te implementeer. Die Engelse weergawe van die toets wat by die Universiteit Stellenbosch gebruik word, staan bekend as TALL ("Test of Academic Literacy Levels") en is ontwikkel om lees- en skryfvaardighede binne ’n akademiese konteks te evalueer. Die resultate van die toets word gebruik om studente in programme te plaas waar hulle ondersteuning kry vir die verwerwing van noodsaaklike akademiese vaardighede. Hierdie studie ondersoek, enersyds, die belangrikheid van luister as ’n vaardigheid tydens die evaluering van akdemiese geletterdheidsvlakke; andersyds ondersoek dit die moontlike waarde van ’n akademiese luistertoets ("Academic Listening Test", ALT) as aanvulling tot TALL om ’n meer akkurate evaluering van studente se akademiese geletterdheid te bied, veral by grensgevalle. Die ontwerp en uitvoering van ALT is gebaseer op die teorieë en benaderings van verskeie kenners in die veld. Die empiriese navorsing het begin met die ontwerp en toepassing van ALT gevolg deur ’n loodsprojek waar kwalitatiewe data oor die geldigheid van ALT ingesamel is by wyse van ’n vraelys. In die volgende fase is die akademiese luistervaardighede van ’n groep eerstejaarstudente geassesseer aan die hand van ’n toets; om betroubaarheid te verseker, moes hulle dieselfde toets een maand later weer voltooi. Die resultate van albei toetsgeleenthede is statisties ontleed om die betroubaarheid en geldigheid van ALT te bepaal. In die laaste fase van die studie is die ALT-resultate vergelyk met dié van TALL om die vlak van kriterium-gebaseerde geldigheid vas te stel asook om te bepaal of ALT ’n bruikbare aanvulling tot TALL kan wees.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

THE CONTEXT OF THE STUDY

1.1 INTRODUCTION

Universities and other higher education establishments throughout the world, including in South Africa, have become concerned about the academic literacy levels of the students they enrol. The problem at most South African tertiary education institutions is certainly considerable where almost a third of the students are identified as being at risk. A lack of ability in academic discourse is seen as a major cause of students' failure to complete their courses within the given period (Weideman, 2003b:56). According to Van Dyk and Weideman (2004a) and Van Schalkwyk (2008:2) an under-preparedness for the 'intellectual demands of higher education programmes' has often been cited as a contributory factor to the current problem.

In this chapter the rationale for and specific research aims of this study will be presented. A brief discussion of the literature reviewed in order to provide the theoretical framework of the study as well as the methodology I followed to address the research problem, is also included. In addition, this chapter provides an outline of the structure of the study.

1.2 RATIONALE AND BACKGROUND

In 2006, as part of an attempt to remedy the academic literacy crisis, the University of Stellenbosch, in collaboration with academics from other universities, officially decided to implement a test of academic literacy in both English and Afrikaans. The English test, known by the acronym TALL (Test of Academic Literacy Levels) is a paper-based test and focuses on reading and writing tasks. The results are used to place students into programmes that assist them to acquire some of the skills needed for academic success. Since students are sorted according to their TALL results into 'high risk' and 'low to no risk' categories, the need for a more refined method of screening the students whose performance on the test falls between these two groups has been identified (Van Dyk, 2007). This objective forms the rationale and constitutes the relevance of this study whereby an academic listening test was designed, operationalised and examined for usefulness as an added dimension to TALL. In order to avoid confusion, the academic listening test designed and developed for this study will henceforth be referred to as ALT (academic listening test).

Administrative and logistical limitations have thus far prevented listening skills from being included in the construct of TALL but there is general consensus that listening is an important skill, particularly at university level. The initial focus of my research was, therefore, to design and implement an appropriate test (ALT) to determine the level of listening proficiency among a sample of first year students at the University of Stellenbosch. The second phase of this research involved a correlation of the results of ALT with those of TALL to determine whether a listening component would provide a more accurate screening of incoming students' academic literacy levels.

1.3 THE SCOPE OF THE STUDY

This study encompasses the design of an academic listening test primarily to, examine the significance of listening in the assessment of academic literacy levels. A secondary aim was to explore the potential for an academic listening test to assist TALL in making more informed decisions on the levels of academic literacy displayed by the candidates. Unfortunately, one of the main limitations of the study was the small sample size of the candidates. However, this situation can be remedied in further research emanating to this study.

An understanding of the concept of academic literacy, although important for this study, is not the main focus. A more comprehensive explanation of the term is given by Boughey (2000), Gee (1990) and van Schalkwyk (2008). For the purposes of this study, however, the construct of TALL given in 2.7, is used as a basic list of skills deemed necessary to be declared ‘academically literate’ at the University of Stellenbosch. The term ‘assessment’ in this context is used as a synonym for ‘testing’. The term ‘listening’ which is of course the central theme of the study, is discussed at length in 2.8. and 2.9.

1.4 THE RESEARCH TOPIC: A STUDY OF THE LITERATURE

In an attempt to refine and focus my research topic as well as develop a theoretical framework for the assessment instrument (ALT), I decided to begin my study of the literature with material on language testing in general. It was also necessary to examine the abilities that are required of a student in order to be considered academically literate. Since the design of the test I had in mind for the study was to be delivered by computer for practical reasons, the technological aspect of language testing also had to be considered. The field was then narrowed to listening assessment with particular emphasis on testing listening proficiency in an academic setting. Preliminary research into the literature will be touched upon in this chapter and reviewed in greater depth in Chapter 2.

1.4.1 General language testing

It appears to be widely acknowledged in the literature that language tests are a means of measuring general or specific language abilities through the execution of tasks (Bachman, 1990; Bachman & Palmer, 1996; Weir, 1993). These should be carefully designed so as to be able to predict, as accurately as possible, an individual's ability in a real life context (McNamara, 2000:11; Douglas, 2000:42). It seems logical therefore, to assume that a test of language ability which is limited to a particular setting would be more useful to a test-taker than a more general approach (Rost, 1990:180). However, defining a target language use (TLU) domain is 'extremely difficult' because of the abundance of variables which need to be considered (Fulcher, 1999:224). This has occasioned Buck (2001:106) to suggest that test designers have to be content with an 'approximation' of the TLU situation. Nonetheless, once a future language use situation has been identified, there are two aspects which are essential for defining the construct. The first is the specific abilities that test-takers should possess to be successful in the TLU domain and the second, the kind of tasks they should be able to perform (Buck, 2001:102). The content of the language test should thus be 'relevant to the knowledge, skills or abilities important in the domain' (Fulcher, 1999:227).

1.4.2 Academic literacy testing

A test of academic literacy would, thus, set out to predict the future ability of entry level students to meet the linguistic requirements of academic learning. Therefore, the purpose of such testing would be to assess the language abilities and thinking skills of students so as to determine their preparedness and odds of success as well as determine the type of support that may be required to facilitate this. Some examples of these skills which are included in the construct of TALL are:

- understanding academic vocabulary in context;
- making distinctions between important and less important information;
- being able to infer meaning from implicit rather than explicit information (Weideman, 2003a:xi).

It is also possible to measure these abilities in a listening test since many cognitive language theorists agree that listeners employ the same schemata to process auditory input as they would in other sensory processing, such as reading (Alderson, 2005:138; Anderson & Lynch, 1988:18; Lynch, 1998:10).

1.4.3 Test validation

A recurring theme present in the literature concerning language testing is the necessity for test developers to ensure, as far as possible, that tests are both reliable as well as focused on relevant validity. These two concepts are so enmeshed that a test cannot be reviewed for one without the other being taken into consideration. According to Alderson, Clapham and Wall (1995:187), a test cannot be valid if it is not reliable. In other words, a test has to be consistent in its measurement or it cannot be considered accurate. On the other hand, a test can be reliable without being valid if consistent results are recorded on a test but the test does not measure what it was designed to assess. However, Alderson et al. (1995:188), maintain that when conducting validation studies, the most important consideration is 'whether the test yields a score which can be shown to be a fair and accurate reflection of the candidate's ability'.

For the purposes of this study, the question of bias, as a result of computerised language testing, has also had to be addressed for reasons of validity. Since the delivery mode of ALT is through the computer, careful consideration had to be given to ensure that test performance would not be significantly affected by the level of a candidate's computer skills. In computerised testing test designers need to be mindful of the test method effects such as the quality of the recordings and the layout of the test (Douglas, 2000:277). Since reading on screen is known to be more difficult than on paper, font size and spacing are important considerations. According to Buck (2001:255), the overriding issue is whether the computer can deliver tests which are more true to life and therefore have a more realistic listening construct than conventional tests.

1.4.4 Listening comprehension

According to the literature, researchers have yet to agree on a widely-accepted definition of listening comprehension. This could be as a result of the numerous different processes and variables which are involved making it almost impossible to provide a single comprehensive definition (Wagner, 2002:1). However, an accord seems to exist among researchers regarding the characteristics which make up the listening process (Brindley, 1998:172; Dunkel, Henning & Chaudron, 1993:180; Lynch, 1998:3). Most academics in the field agree that listening comprehension comprises linguistic as well as non-linguistic knowledge. The importance of linguistic knowledge which involves the structure of the language at word, sentence, paragraph and even whole text level should not be underestimated but it is insufficient without extra-linguistic knowledge. The latter applies to what an individual knows about a topic and its context as well as their general knowledge of the world (Buck, 2001:2; Lynch, 1998:3).

Buck (1991:67), Rost (2002:59) and Brindley (1998:181) are, furthermore, all of the opinion that listening comprehension entails far more than merely applying one's knowledge of the language in order to interpret a text. They also agree that listening is a process whereby listeners extract meaning based on their own previously stored knowledge and experience. Shohamy and Inbar (1991:26) in accordance with the ideas of Buck, Rost and Brindley (above), also maintain that there has to be an interaction between the listener's background and the spoken input. Because of differences in knowledge, memory capacity and general mental ability, the results will vary substantially from individual to individual (Bejar, Douglas, Jamieson, Nissan & Turner, 2000:4). Anderson and Lynch (1988:11) add to this theory by suggesting that listener performance in understanding an utterance is affected by the listener's purpose in addition to their background knowledge and ability to store information. Given the common denominators in the opinions of the researchers mentioned above, it seems that there is some agreement that the intricate process of listening comprehension involves linguistic, situational and background knowledge which have to be synthesized in order to achieve meaning (Bejar et al., 2000:4).

1.4.5 The two stages of listening

The view that listening is a two stage process regularly emerges from the literature (Buck, 2001:51; Chaudron & Richards, 1986:113; Rost, 1990:33; Shohamy & Inbar, 1991:29; Weir, 1993:98). These two stages consist of bottom-up processing, involving the more 'local' skills such as the identification of details and extraction of facts, and top-down processing which requires interpreting the more implicit information such as inferencing or listening for gist. However, there seems to be no particular sequence in which the two processes take place and they often occur simultaneously in a so-called parallel process (Rubin, 1994:211). This makes it very difficult to attribute task responses to any one particular skill or construct (Brindley, 1998:173; Buck, 2001:106).

1.4.6 Academic listening

As is repeatedly mentioned in the literature, language skills cannot be separated and a good example of this is in a lecture situation where listening, reading and writing are fully integrated. Students listen to a lecture, take notes and then use the notes for study or assignment purposes. Flowerdew (1994:11) has found that the processing required for effective academic listening is far more complex than, for example, listening to a conversation. Rost (2002:162) maintains that this is because academic listening is mostly a non-collaborative or one-way listening process of which a lecture is the most typical example.

Hansen and Jensen (1994:249) suggest that questions on a lecture comprehension task should have two aims. Firstly, they should be designed to assess the candidate's *grasp* of the content and secondly to evaluate their employment of effective auditory skills. 'Global' or 'top-down' questions serve to measure the former using such sub-skills as the identification of main themes as well as the aim and topic of the lecture. 'Detail' questions based on 'bottom-up' processing are designed to evaluate the candidate's ability to recognize the most important key concepts or terms. Hansen and Jensen (1994:254) also emphasise the importance of using authentic rather than scripted or staged discourse. Authentic lecture recordings, for example, will include pauses, fillers and other general characteristics of natural speech, important for assessing performance in real life settings. The use of authentic material and tasks reflective of those that will be required at university, gives the test strong content validity.

My research into language testing in general and, more specifically, listening testing has emphasised the difficulty of accurately assessing this complex linguistic ability. However, although daunting, the design of an assessment instrument that could prove be an effective means of testing academic listening proficiency as an added dimension to TALL, seemed to be a worthwhile project.

1.5 RESEARCH OBJECTIVES

The principal aim of this study was to design a computerised test to qualitatively and quantitatively assess the academic listening skills of a selection of first year university students. A retest of ALT was conducted a month after the initial testing to determine the reliability of ALT and the results obtained from both administrations of ALT had to be analysed to present an argument for validation. A question which stemmed from this procedure was whether a significant difference could be determined in the test and retest scores. To address issues of content, face and construct validity, a pilot test was carried out where participants responded to a questionnaire. The pilot testing was also an important check for any technological problems which could threaten the construct validity of ALT.

A subsequent research objective entailed examining the comparison between the results of ALT and those of TALL. The outcome of such a correlation would serve two purposes: firstly, it would be useful for the assessment of criterion-related validity of ALT and secondly it would determine whether a listening component would indeed be a useful added dimension to TALL.

1.6 RESEARCH QUESTIONS

In order to conduct a validation study of the results obtained from the administration of ALT, various qualitative and quantitative analyses had to be carried out and the following research questions investigated:

1.6.1 Questions based on qualitative data

1. How representative and relevant are the tasks included in ALT?
2. Are experts in the field confident that the results of ALT would be an efficient indication of academic literacy levels?
3. Is the level of difficulty, layout, sound and visual quality of the clips, and clarity of instruction included in ALT conducive to good construct validity?

1.6.2 Questions based on quantitative data

1. Does ALT provide internal consistency of measurement for each section?
2. Do the internal correlations of the different sections in ALT provide evidence of construct validity?
3. Is there a significant difference between the scores from the first ALT administration and the retest?
4. Does ALT show concurrent validity by demonstrating a positive correlation with TALL?
5. What is the correlation between the scores of the borderline TALL test-takers and their performance on ALT?

From these research questions, the following hypothesis was formulated:

An academic listening test would be a useful added dimension to TALL, currently implemented at the University of Stellenbosch.

1.7 RESEARCH METHOD

The literature indicates that a test can only be deemed 'useful' if its ultimate purpose is known (Bachman & Palmer, 1996:38). Therefore, without a clear purpose for the assessment, no decisions can be made on the reliability, validity or appropriacy of test specifications or of its underlying theory (Brindley, 1998:183). As has been previously mentioned in this chapter, the

purpose of ALT was to assess the listening skills in an academic context of a sample of first year university students. This purpose, along with a theoretical framework and knowledge of the tasks required by the TLU domain were essential for making decisions concerning the content of ALT. These issues will be explained in greater detail in Chapter 3.

The design process began with test specifications which determined both the method and the content of ALT. This test 'recipe' stipulated the type and length of texts, details of the instructions as well as how the test would be scored (McNamara, 2000:31). The test specifications for ALT are included in section 3.10 in Chapter 3. The framework of ALT was based on the theories and approaches of several researchers in the field such as Buck (2001), Weir (1993), Wagner (2002) and Jordan (1997), as well as the compilers of TOEFL (Bejar et al., 2000). Since listening comprehension is an internal process which cannot be directly observed, researchers have had to resort to assessing the more easily measured skills associated with the listening process (Brindley, 1998:172; Weir, 1993:98; Rost, 1990:33). For the purposes of this study, I decided to use the often cited 'two-stage listening process' as discussed above (Buck, 2001:51; Chaudron & Richards, 1986:113; Rost, 1990:33; Shohamy & Inbar, 1991:29; Weir, 1993:98), as a theoretical framework on which to base the abilities I wanted to test. 'Local' skills such as identifying specific details and facts and recognising supporting ideas represented bottom-up processing while identifying the main theme of a text or inferring meaning from more implicit information involved top-down processing skills.

ALT was adapted to fit into the assessment format of Blackboard, the learning management system (LMS) used at the University of Stellenbosch. The reasoning behind this decision was the ease and accuracy of scoring as well as the computer's ability to instantly calculate statistical data.

ALT is divided into four sections which are placed in an 'easier-to-more-difficult' order and test-takers are advised of the listening purpose for each task. Clear instructions are given at the beginning of each task and, where necessary, additional information is given before some of the questions. The tasks are all designed to assess certain abilities as well as represent the academic TLU domain on which they were based.

1.8 RESEARCH PROCEDURE

After the completion of the design phase of ALT, a pilot project was conducted before the inception of the main study so as to receive qualitative feedback on ALT and to make sure that there were no technical hitches. A group of nine Health Science first year students and

three lecturers from the Unit for Afrikaans and English all from the University of Stellenbosch, volunteered to complete ALT and respond to a questionnaire. A copy of the questionnaire is included as Appendix A. The rationale for the design of the questionnaire was to collect information from the participants relating to the content and face validity of ALT. Feedback on the representativeness and relevancy of the tasks is important for gauging both content and face validity as is the general opinion of peers. In addition, issues pertaining to ALT's construct validity were also included in the questionnaire. These included aspects such as the appropriacy of texts and tasks, clarity of instructions and the sound and visual quality of the media files included in ALT.

The main study involved requesting volunteers from a group of six hundred and twenty seven Bachelor of Science first year students to take ALT. These students had all attended a semester of *Scientific Communication Skills*, either in English or in Afrikaans, which provides assistance in academic literacy skills. For reasons of reliability, the administration of ALT was conducted in two parts comprising an initial testing and a retest of the same test a month later. Ninety seven students completed both administrations of ALT which was administered in a multimedia lab since headphones were a prerequisite for the test. All the test-taking sessions were monitored by a supervisor to ensure that there were no technological problems or distractions which might affect the performance of the test-takers.

1.9 DATA ANALYSIS

The qualitative data collected from the pilot testing were analysed for feedback on the content, construct and face validity of ALT. The results of the first and second administrations of ALT as well as the correlation with TALL were statistically analysed using STATISTICA. Both the qualitative and quantitative data will be presented and discussed in Chapter 4.

1.10 STRUCTURE OF THE THESIS

This chapter has provided an overview of the framework of the study that will be described in depth in the chapters that follow. It discussed the rationale for and the context of the research aims as well as the methods used to gather and process the data.

In Chapter 2 a review of current and past scholarship in the field of language testing and academic literacy is presented. The particular focus of this chapter is on listening testing with specific emphasis on academic listening testing. The constructs, theories and models of listening comprehension form an important part of this focus.

Chapter 3 includes a description of the research design and methodology of ALT as well as data collection procedures and details of the participants in the study. This chapter forms the foundation for the following chapter where the results of the qualitative and quantitative data are presented and discussed. The findings included in Chapter 4 in turn, provide the focus for Chapter 5 where the implications and interpretations of these results will be further discussed with particular relevance to the theory presented in the literature review. Limitations of the study and opportunities for further research will be reflected upon at the close of this chapter.

CHAPTER 2

A REVIEW OF THE LITERATURE

2.1 INTRODUCTION

As indicated in Chapter 1, the significance of listening as a component of general academic literacy, as well as the assessment of listening competency, forms the nucleus of this study. However, in order to construct a reliable theoretical framework, it was necessary to examine the principles, approaches and models of language testing in general. For this reason, I have begun my review of the scholarship with issues that are common to all language testing. I then proceed to the matter of testing for specific purposes, in this case academic literacy. Listening as a construct had to be investigated at some length before decisions could be made concerning the design of the assessment instrument, ALT for use in this study. Listening skills, strategies and assessment, as well as the theories and approaches which underpin them, will then be examined in some detail.

The test designed for this study, is delivered by computer so relevant aspects of computerised testing will also be discussed in this chapter. The most significant principles and theories, in terms of this particular study, will be revisited in subsequent chapters.

2.2 PRINCIPLES OF LANGUAGE TESTING

According to Bachman and Palmer (1996:9), there are two fundamental principles of language testing. The first dictates the need for close ties between the performance of candidates in a language test and their future language use. The second important consideration is concerned with test usefulness. These principles inform the qualities that are essential to the design and development of a language test.

2.2.1 First principle

This concerns 'a correspondence between language test performance and language use' (Bachman & Palmer, 1996:9). In order to achieve this, there has to be a frame of reference between the language test results and real-life language use. An effective language test, therefore, needs to consider the characteristics of the target language use (TLU) situation, or domain, as well as the characteristics of the test-takers and tasks. However, the distinction between the criterion (relevant communicative behaviour in the target situation) and the test has to be recognized (McNamara, 2000:8).

According to McNamara (2000:8), testing is concerned with making inferences. Test performance is used to deduce criterion performance. Even if a test includes only authentic content, it is still only an *indication* of how someone might perform in reality. Authentic material, although important, can never be considered real because of the artificial or simulated nature of the testing process. The plausibility of inferences gained from language tests and made on the basis of performance in a test is known as test validation. The method used to interpret language test scores depends on the purpose for which the test is intended (Bachman, 1990:226).

2.2.2 Second principle

If a test is to be deemed useful, it has to be developed for a specific purpose. Bachman (1990:226), states that the most important consideration in the design and use of a language test is the purpose and thus the use for which it is intended. According to Bachman and Palmer (1996:38), there are six test qualities which constitute test usefulness. These are: reliability, validity, authenticity, interactiveness, impact and practicality. Ideally, a balance between these qualities should exist which will vary according to the purpose of a particular test. In addition, it is important that they are central to the control of quality throughout the process of designing and developing a particular language test. However, the two most important qualities specifically for testing are *reliability* and *validity* (Bachman & Palmer, 1996:19).

2.3 TEST QUALITIES

All test developers should make sure that tests are reliable as well as focused on relevant validity. Interpretation of scores needs to be done with discernment and intelligence (Spolsky, 1995:356). According to Alderson et al (1995:187), a test cannot be valid if it is not reliable which means that a test has to be consistent in its measurement or it cannot be considered accurate. On the other hand, a test can be reliable without being valid. This means that consistent results are recorded on a test but the test does not evaluate what it was designed to measure.

There has been some debate among language testers that multiple-choice tests have reduced validity because they do not accurately reflect the ability to use language in real life. Neither reliability nor validity is irrefutable and sometimes it is necessary to increase one thereby reducing the other (Alderson et al., 1995:187; Hughes, 2003:50). These two concepts are so interlinked that a test cannot be checked for reliability without the validity being considered and vice versa. Ultimately, whether checking a test for validity or reliability, the most important

consideration is 'whether the test yields a score which can be shown to be a fair and accurate reflection of the candidate's ability' (Alderson et al., 1995:188).

2.3.1 Reliability

Reliability, according to Davies (1990:52) is a 'statistical reassurance of consistency of result'. In other words, the results obtained are dependable. For this to occur, adequate results from sufficient information about a candidate's language abilities must be gathered. The test also has to be an effective measure of whatever it sets out to assess (Davies, 1990:6; Jordan, 1997:88). Hughes (2003:50) agrees that language testing is primarily concerned with consistency; in fact it is often defined as consistency of measurement (Bachman & Palmer, 1996:21).

The reliability coefficient can be calculated either by comparing the test results of two different tests or by administering the same test on two different days. This is also called the test-retest method. Another test for reliability is the coefficient of internal consistency which makes use of the split half method where the results of two halves of the same test are compared (Hughes, 2003:38; Lado, 1961:31; Alderson et al., 1995:87). In this particular study, the test-retest method in order to measure the performance of candidates from one occasion to another will be used as a test of reliability.

The degree to which the test scores are free from measurement error, impacts on the reliability of a test. Therefore, it is important to estimate the effect that various factors may have on test scores. Test scores are interpreted as indicators of specific language ability, so they have to be as reflective as possible of that ability. The generalizability theory enables test developers to identify sources of variance (variables) and differentiate between systematic and random error (Bachman, 1990:226). Systematic or true differences are the degrees of skill being measured. These are mostly as a result of differing proficiency levels. Unsystematic or random variables are due to lack of concentration or distraction. A perfectly reliable test would measure only systematic changes. Although a perfectly reliable test is not possible, test developers can, as far as possible, reduce the variables. Some of the ways of doing this are: by making sure that the rating is consistent, instructions are clear and by removing any ambiguity from test items (Alderson et al., 1995:87). Issues of reliability pertaining to ALT, the specific test used in this project, will be analysed and explained in Chapter 4.

2.3.2 Validity

A valid test must provide consistently accurate measurements (Hughes, 2003:50). Validity is the theoretical framework which gives credibility to the test (Davies, 1990:6; Alderson et al., 1995:180). It is a process involving logical analysis and empirical investigation (Bachman, 1990:289).

The validation of the test rests on the evidence emerging from test scores. This, in turn, supports the credibility of the interpretation of the construct or trait (Weir, 2005:1). It must, however, be remembered that when examining validity, factors, other than the language abilities being measured, will affect the test results (Bachman, 1990:289). The purpose of validation in language testing is to make sure that the inferences drawn from the results of the test are both fair and reliable (McNamara, 2000:48). The better the definition of purpose or the more precise the ability to be measured, the more likely the test is to be valid (Weir, 1993:19). This concurs with Lado's (1961:30) statement that validity is not general but specific.

Most authors agree that a good test should measure specifically what the developer wants it to measure (Alderson et al., 1995:170; Davies, 1990:21; Jordan, 1997:88; Weir, 1993:19). Ideally, a test should be limited to testing only what it means to test and not *incidental* abilities. However, it is almost impossible to divorce one skill from another; for example, if one is testing listening ability, a candidate’s score might be affected by the quality of his/her reading skills. This makes it more likely that listening competency in a listening test is being assessed as a component of an integrated set of language abilities (Rost, 2002:172).

Since the delivery mode of ALT is through the computer, careful consideration also had to be given to ensuring that test performance would not be significantly affected by the level of a candidate's computer skills. The question of bias, as a result of computerised language testing, also had to be addressed for reasons of validity. In the past, this has been measured by comparing the performances of candidates with a range of tests on computer experience and ability. Results showed that there was no significant difference between those who were familiar with computers and those who were unfamiliar since all the candidates had taken a course in basic computer skills (Chapelle, 2001:123). Likewise, the students participating in this study would also have completed an elementary course in computer skills soon after the start of their first semester. Furthermore, as more and more test-takers become increasingly at ease with computers, validity issues based on computerisation-bias is fast losing relevance (Chapelle & Douglas, 2006:17).

Chapelle and Douglas (2006:106) complain that most public discussion surrounding the validity of computerised testing addresses the issue of how technology could undermine conventional methods of testing rather than how these could be improved and innovated. It is, however, a fact that in computerised language testing, different validity problems could arise by possible breaches of test security or technological failure. It is important that only authorized individuals are able to access the test. Test security also extends to protecting the item pool, administrative system and scoring records of candidates. Pilot testing is essential in computerised testing because of the necessity of identifying possible technological problems (Dunkel, 1999:89).

There are several types of validity which are frequently referred to within the context of testing. These all involve the relationship between the test instrument and the domain to be measured (Davies, 1990:6). The most common, according to Davies, Brown, Elder, Hill, Lumley and McNamara (1999:221) are content, construct, concurrent and predictive. One of the most important types of validity is construct validity. The construct validity of a language test 'involves an investigation of the qualities that a test measures, thus providing a basis for the rationale of a test' (Davies et al., 1999:33). However, Shohamy and Inbar (1991:23) suggest that some controversy surrounds the definition of construct validity. While some researchers consider content, criterion and construct validity to be different types of validity, Messick (1989:20, 1996:248) and Bachman (1990:290) see content and criterion validity as building blocks for construct validity.

Davies et al. (1999:221) in their *Dictionary of Language Testing* include face validity as an 'also ran', possibly because it is not considered to be a scientific concept and cannot be presented as evidence of construct validity (Hughes, 2003:33; Clapham, 1993:267). However, there are many who feel it still has a role to play in predicting how candidates will react to a test and that this could affect their scores. For this reason, I decided to include it in my criteria and will revisit the notion in Chapter 4 when the feedback from my qualitative study is discussed. I also include concurrent and predictive validity as two components of criterion-related validity, a concept mentioned by Hughes (2003:27). Criterion-related, content and construct validity will all be revisited in Chapter 4 with particular reference to ALT.

2.3.2.1 Content validity

The content of a test needs to be selected according to the specification of task domain and the abilities required to carry out tasks in this context. If this is done successfully, the test will be deemed to have content validity (Bachman, 1990:289; McNamara, 2000:73). The content of a test also informs the candidate of what is considered important and less important in that

domain (Fulcher, 1999:233). The greater the test's content validation, the more likely it is to be an accurate assessment of what it is supposed to measure (Hughes, 2003:27).

In order to assess this, testers have to rely on their own and their colleagues' professional judgement of whether the test includes adequate and relevant samples of target situation language abilities (Alderson et al., 1995:176; Davies, 1990:23; Messick, 1989:36). Here, content validity merges into construct validity (Davies, 1990:23). Fulcher (1999:234) agrees by saying that test-content validity should not be an end in itself but should instead be approached from an angle where content relevance is considered against the backdrop of construct validity. He reasons that score meaning will be established in the light of construct validity studies rather than just from the content of a test.

2.3.2.2 Face validity

This kind of validity is closely connected to content validity and concerns whether the test looks valid to a non-expert, which could represent how a candidate would view the test. It gives a test developer an idea of whether the test has superficial credibility in the eyes of the general public. With the advent of CLT or Communicative Language Testing, face validity has become more important and the emphasis has shifted towards authenticity where tests emulate the real world (Alderson et al., 1995:172).

Stevenson (1985:111) mentions two problems that could arise from face-validity judgements. The first is that he considers it a mistake to assume that *tests* can be judged to be either valid or invalid when it is actually the *inferences*, based on the test scores, which can be deemed valid or invalid. Messick (1989:13) is in complete agreement and suggests that looking to 'instruments rather than measurements', to test forms rather than test scores could obscure decisions on content and face validity. The second problem is the difficulty in accurately defining or describing the TLU domain because of the multitude of variables. Even experts disagree on what constitutes academic English and the levels of performance required to succeed in this domain (Fulcher, 1999:224). Messick (1989:36) is of the opinion that this has led to a 'simplistic view of validity' because some, so-called authentic, tests have often been declared valid purely on their appearance. However, Bachman and Palmer (1996:23-4) argue that if a test is perceived to be authentic by test-takers, they are more likely to perform the tasks efficiently. This, in turn, provides an accurate reflection of their ability which positively affects score meaning and thus improves validity.

2.3.2.3 Criterion-related validity

A criterion is 'an external variable such as a syllabus, teacher's judgement, performance in the real world, or another test' (Davies et al., 1999:37). Evidence for this type of validity is provided by identifying appropriate criterion behaviour, perhaps from another language test, and then comparing the results of the test to this criterion. The criterion itself can be an indication of validity of the abilities measured in the test (Bachman, 1990:290). Test results can also be compared with another credible and dependable assessment of the candidates' ability. For the purposes of this study, ALT results will be correlated with those of TALL (Test of Academic Literacy Levels), currently used at the University of Stellenbosch and which has proven reliability and validity (Van der Slik & Weideman, 2005; Van der Walt & Steyn, 2007). TALL, and its use as a criterion test, will be discussed in more detail in the next chapter.

There are two kinds of criterion-related validity, namely, concurrent validity and predictive validity. According to Hughes (2003:27), concurrent validity is often measured when the test and the criterion are administered in approximately the same time frame. A high correlation coefficient might be expected since both are testing language ability, but if they are testing different aspects of language ability, a low correlation coefficient could result (Alderson et al., 1995:177). Predictive validity is common in proficiency testing because predictions are being made about how well an individual will perform in the future (Hughes, 2003:29; Alderson et al., 1995:180). Results of past tests could also be included in making these forecasts (Alderson et al., 1995:177).

2.3.2.4 Construct validity

This type of validity is concerned with how well the tasks used in a test, reflect real life requirements (Davies, 1990:23). Bachman (1990:6) states that construct validation is central to language testing research because it involves the examination of the relationship between performance in language tests and the abilities on which the performance is based. It is essential to determine whether a test score is a true predictor of future performance (Bachman, 1990:290; McNamara, 2000:104).

According to Bachman and Palmer (1996:21) and Messick (1996:243), enough empirical evidence has to be supplied to justify the interpretation of a test score. Construct validity is thus the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure, bearing in mind that the interpretations we make of test scores can never be considered absolutely valid (Bachman & Palmer, 1996:22). According to Bachman and Palmer (1996:22), this is because even though there is evidence to support an

interpretation, it has to be viewed as 'tenuous'. Furthermore, both the definition of the construct and the features of the test tasks need to be considered when interpreting test scores, in order for the test to be termed construct valid. There are two reasons why test tasks should be closely examined: the first is to observe how closely the task corresponds to actual practice in the TLU domain, the second concerns the extent to which the task engages the candidate's sphere of language ability (Bachman & Palmer, 1996:21).

Messick (1996:245) mentions two major threats to construct validity: the first is construct under-representation where important 'dimensions or facets of focal constructs' are omitted, causing the test to be too narrow. The second, he calls construct-irrelevant variance, which means that irrelevant factors are added which cloud the interpretation of the construct assessment. In other words, the test is too broad (Messick, 1996:244). Since computers are used as the mode of test delivery in this study, there is also the concern that construct-irrelevant factors such as familiarity with computers or computer anxiety might form part of the assessment and impact negatively on the performance of candidates. However, since all the test-takers were exposed to the same computer training course, it was assumed that they all had some fundamental degree of computer literacy.

Construct validation goes beyond content and criterion-related validity since it 'empirically verifies (or falsifies) hypotheses derived from a theory of factors that affect performance on tests – constructs, or abilities, and characteristics of the test method' (Bachman, 1990:290). According to Messick (1996:248), both content validity, which is achieved through the opinion of experts, and criterion validity, which comes from correlations with other tests, combine to provide proof of construct validity. Messick (1989), in his well-known article on validity, also insisted that the social aspect of testing and its impact on the test-takers had to form part of any inferences drawn from test results. His rationale was that performance assessment by its very nature will reflect an individual's value system. He further averred that these values would have strong links with the test-takers' culture and the society to which he/she belongs. Messick introduced the concept of these aspects of validity as a single theory of validity which he displayed in the following much cited matrix.

TABLE 2.1: FACETS OF VALIDITY (MESSICK, 1989:20)

	TEST INTERPRETATION	**TEST USE**
EVIDENTIAL BASIS	Construct validity	Construct validity + Relevance validity
CONSEQUENTIAL BASIS	Value implications	Social consequences

As with all testing, the construct validity of a computerised test also concerns the degree to which the test scores allow accurate interpretation of the abilities being measured. The establishment of a clear link between the 'focus and structure' of the items and the 'purpose' of a test is not easy. However, there has to be correspondence between the test goals and the kind of inferences made about ability based on test performance (Dunkel, 1999:85).

The literature indicates that there is some skepticism that computer-assisted tests may affect test performance. There is also concern that this kind of testing may prove less valid than conventional paper-and-pencil testing (Chapelle, 2001:95), although the same criteria apply to both types of testing. Chapelle (2001:96) gives an example of how a computer could make a test task dissimilar from the TLU domain in that reading on screen is acknowledged to be more difficult than on paper.

Eva Baker (1998:22), however, believes that technology could be the key to solving the ongoing problem of validity which currently exists in the testing system. For example, authenticity could become a focal point for computerised test development and evaluation. Content and construct validity could be greatly enhanced by technology because of the extensive options it offers test developers by way of text and media with improved reliability, sound quality and authenticity (Jordan, 1997:89).

2.3.3 Authenticity

This is the extent to which the demands of a test task correspond with the characteristics of a TLU task (Bachman & Palmer, 1996:24). Authentic tests include tasks in realistic settings which mirror those in the TLU situation (Messick, 1996:243). Authenticity of test tasks allow score interpretations to go beyond the test performance to improve prediction of language use in the real world. It thus has very close ties with construct validity.

Another important feature of authenticity is its effect on the test-taker’s impression of the test, in other words, its face validity. If the language use needs of test-takers are identifiable, for example a university setting, it can be helpful in determining the kinds of authentic tasks needed in the development of a practical test (Bachman, 1990:356). This could have a positive impact on candidates' performance as it is thought that if they perceive the tasks to be relevant, the effect will be beneficial (Bachman & Palmer, 1996:24). It is also important to identify the communicative language abilities as well as the situation which will determine the kind of interaction which will be required of the candidate in the TLU domain (Bachman, 1990:356).

Chapelle (2001:115) raises the point that the construct of a computerised test is very reflective of tasks in real life, given the modern environment where students spend a lot of their time on computers. Computers, as a medium for testing, might be considered more difficult than the paper-based variety by some students. However, with technology becoming a way of life for most school children as well as university students, the use of computer testing is widely thought of as the way forward (Chapelle, 2003:28).

Computers are also able to provide test developers with 'rich multimodal input' in the form of video, text, sound and graphics which can add to the authenticity of a test (Chapelle & Douglas, 2006:9). The possibility of including multimedia in a self-contained application makes it possible for more authentic material to be used (Chapelle, 2003:28; McNamara, 2000:79). For example, the visual input of a video clip of a lecture attempts to realistically reflect a similar activity in the real world (Chapelle, 2001:108).

2.3.4 Interactiveness

According to Bachman and Palmer (1996:25), interactiveness concerns the degree and kind of interaction that occurs between the test-taker and the task. It is important in the assessment of a candidate's language ability, background knowledge of the topic and their thinking strategies. Computerised testing introduces a heightened engagement between the test-taker and the task since everything happens on screen and the use of multimedia provides increased interactiveness (Chapelle & Douglas, 2006:91). According to Chapelle (2001:1), 'the nature of communicative competence has changed in a world where communication occurs with computers and with other people through the use of computers'. Students exist in a world where communicative competence includes electronic as well as academic literacies.

Because of its link to construct validity, in areas such as linguistic knowledge, strategic competence and background knowledge, interactiveness is a significant quality of test tasks. However, this is dependent upon how the construct is defined as well as the individual characteristics of the test candidates. Moreover, interactiveness in the design of tasks, as with authenticity, can never be guaranteed or precisely defined since all individuals' process information in different ways (Bachman & Palmer, 1996:29).

2.3.5 Impact

This refers to the consequences or decisions which derive from an analysis of test results. An important facet of impact is washback, which examines the effect of testing on teaching and learning as well as on the individual candidates. Test-takers are affected firstly by taking the test; secondly, by feedback they may receive about their results; and, thirdly, by decisions that

are made about them based on their test scores (Bachman & Palmer, 1996:30-1). In a placement test of academic literacy, for example, the results are used to place or stream students into programmes which assist them in acquiring the various skills necessary for their academic success.

2.3.6 Practicality

This test quality differs from the others, in that they are all concerned with the use to which test score inference is put. Practicality, as the name suggests, has to do with the realistic implications and methods of implementing tests. Thus, if a test requires resources over and above what is feasible, practical or even available, then it will not be used. Practicality, therefore, compares the resources needed for the design, development and use of the test, with those that are available, affordable and practical. There are three main types of resources:

1. Human resources, such as test designers, scorers, and test supervisors.

2. Material resources, for example, equipment, such as computers or tape recorders; paper and printing as well as available test venues.

3. Time resources, which include both the design and development phase as well as the administration, scoring and analysis once the test is in operation.

All these resources have accompanying financial costs that have to be considered before proceeding with test development (Bachman & Palmer, 1996:36).

Flexibility of technology for test administration and record-keeping in computerised testing is increasingly seen as a more practical and efficient alternative to paper-and-pencil testing (Dunkel, 1999:77). Chapelle and Douglas (2006:116) maintain that computer-based tests do the same work as other forms of tests, only cheaper and faster (Chapelle & Douglas, 2006:116). Features of computerised testing with special relevance to this study will be discussed further in Chapter 3.

2.4 TEST TYPES

There are two main types of test which are based on *purpose*. These are norm-referenced and criterion-referenced or performance-based tests (Davies, 1990:18; Jordan, 1997:89). The first type reveals a particular candidate's performance as compared to his/her fellow test-takers. Criterion-referenced testing, on the other hand, measures a test-taker's performance in relation to clearly defined criteria (Jordan, 1997:89). In other words a student's results are a

reflection of what they are able or unable to do. Since ALT has a task-based approach with various aspects of listening competency being assessed, it seemed more applicable to employ a criterion-referenced rather than a norm-based measurement.

According to Davidson and Lynch (2002:9), the primary difference between norm- and criterion-referenced testing lies in the test purpose and not in the test content. Davies (1990:19) bears this out by maintaining that criterion-referenced tests are just another way of using norm-referenced tests; they are not, as is sometimes mistakenly thought, different methods of constructing tests.

2.4.1 Criterion-referenced tests

An article by Glaser in 1963 inspired a new trend in language testing. Glaser maintained that an individual's score on a criterion-referenced test would impart significant information on the capabilities of the candidate. Furthermore, the degree of competence could be measured independently of the other test-takers (Davidson & Lynch, 2002:6). In this type of testing, the abilities of the candidate are also assessed in relation to how they will perform in real life. The International English Language Testing System (IELTS) is an example of such a test. Simulated real-life tasks, such as listening to a lecture, often form part of the content (Jordan, 1997:89).

Criterion-referencing sets targets or goals and is concerned with the kind of task to be executed (Davies, 1990:18). The purpose of these tests is to classify individuals according to their ability to perform a task or set of tasks successfully. If, for example, a test is designed to assess the listening ability of entry level students at a university, as is the case in this study, then a criterion-referenced test would be based on an analysis of what the students would be called upon to do in that situation. The main advantage to using this type of test is that a yardstick is set, in terms of what a test-taker can or cannot do (Hughes, 2003:21). According to Davidson and Lynch (2002:7), the real contribution to testing that has been made by the criterion-referencing method is the clarity in test method, content and construct. Computer-based testing is also well suited to criterion-referenced testing where a student's performance is noted in relation to how they will perform in real life (Jordan, 1997:89).

2.4.2 Indirect and direct testing

Indirect tests are often used to forecast language ability in the TLU domain. These tests are 'less natural, more contrived' with an increased emphasis on skills and microskills and as a result, they are usually less authentic. This kind of assessment is commonly used to measure general language proficiency.

Direct tests, on the other hand, are widely used to measure language use in more communicative situations. They are characterized by a task-based approach and favour the use of tasks that are needed in the real world. This kind of test is largely concerned with content validity and focuses on skills relevant to candidates (Hansen & Jensen, 1994:241-2). ALT will focus on these principles and can thus be defined as more direct than indirect.

2.5 TEST SCORING

As was the case in this study, any computerised test can be linked to a learning management system (LMS) such as Blackboard, which is the LMS used by the University of Stellenbosch. Learning management systems are also sometimes called course management systems and, as the name suggests, they are used to manage courses and keep records of student learning activities (Hubbard, 2008). The advantage of using an LMS for testing purposes includes the ability to be operated from a control panel with easily retrievable results. These results can then be stored or compared with other data, making it an extremely practical and reliable means of testing (Dunkel, 1999). The score records allow an examiner to see the total score as well as the performance on each question or task. A discrimination index is also available which allows the examiner to see the difference in performance between top and bottom groups of candidates.

Assessing language proficiency is a decidedly complex task since all measures have a degree of inaccuracy which is inevitable when trying to assess human abilities. Some examples of these causes are: distraction or illness on the part of the test-taker, the competence or bias of the tester, inconsistency in item selection and marking error or lack of objectivity on the part of the raters (Spolsky, 1995:356). To compensate for this, possible causes of error are taken into account by statistically calculating a measurement which defines the 'confidence intervals' surrounding a candidate's score (Davies et al., 1999:186). This is known as SEM (standard error of measurement) and the concept will be discussed in more detail in the section dealing with reliability in scoring later on in the chapter.

2.5.1 Rating methods

Rating is based on recognition of performance at a given level. The scales of rating are defined in terms of the language abilities under assessment in relation to the test task responses. The rating scales have different levels of proficiency and need to be precisely determined (Bachman & Palmer, 1996:226). The most significant statistics to record are the mean, the mode, the median, the range and the standard deviation.

1. The mean should represent the level of ability of all the test-takers.

2. The mode is the score attained by the largest number of candidates.

3. The median is a measure of central tendency and shows where the scores are grouped together. It also describes the score obtained by a candidate which falls in the middle of the other test-takers rankings. This is helpful if the mean is not considered to reflect the general ability of the group.

4. The range refers to the distribution of scores obtained by the group and can also describe the differences between the top and bottom score.

5. The standard deviation, unlike the range, takes every score into account. It measures the average amount that each candidates score differs from the mean (Alderson et al., 1995:92-5).

2.5.2 Statistics in language testing

For research purposes, a numerical method is used to analyse test results. This system uses frequencies, results and correlations (Jordan, 1997:90). There are two main types of analysis, according to Davies (1990:16), and these are known as descriptive and inferential statistics.

Descriptive statistics gives a broad overview of the test findings, sometimes by taking the average score as being representative of the abilities of the group as whole. Validity would fall under this branch of statistics. Inferential statistics, on the other hand, relates to reliability and the extent to which the results of test-takers are representative of the wider group to which they belong is taken into account. Inferential statistics also examines the level of significance or degree of reliability which is needed before a test can be considered satisfactory (Davies, 1990:16).

Computers are able to provide statistical data automatically and can be linked to course management systems (Chapelle & Douglas, 2006:12), making the process of score interpretation much easier. This capability could be further capitalized upon in the future so that even more detailed data are provided (Chapelle & Douglas, 2006). For example, Buck (2001:255) believes that more meaningful test scores would be possible if a score was available for each sub-skill.

2.5.3 Score interpretation

The interpretation of a test score has to be justified by evidence which proves that the score is reflective of the language ability, or criteria we are attempting to measure (McNamara, 2000:104; Weir, 2005:1). Test score meanings form the core of the validity of language tests.

This means that careful consideration of the criteria and rating allocation is essential for construct validity. The results have to be significant, relevant and useful for a specific purpose (Bachman, 1990:49). In order to overcome some of the limitations of test score interpretation, according to Bachman (1990:50), there are three basic steps which need to be followed for effective test development. These are:

- to provide explicit descriptions of the abilities to be measured
- to give precise specifications on how this will be done
- to ensure that the scoring reflects the required features

These provisos supply the necessary link between candidates' ability and performance; they are fundamental for both test development and score interpretation.

2.5.4 Reliability in scoring

Objectivity in scoring is an essential quality for test reliability. Scoring done by the computer is not only more cost effective but also ensures objectivity and efficiency (Chapelle & Douglas, 2006:12). ALT was scored electronically which requires no judgement on the part of the scorer and is a good example of objective scoring (Hughes, 2003:22).

As mentioned previously, the degree of uncertainty of test scores can be counteracted by using a SEM (Davies, 1990:5) which gives an indication of reliability. This measurement is calculated by considering the reliability coefficient and an assessment of the spread of all the scores on the test: the smaller the SEM, the higher the probability that a repetition of the test will yield the same results. This system is also useful in predicting the fluctuation of an individual's scores on two sittings of the same test. Indeed, the SEM is essential for making informed decisions based on test scores. The Item Response Theory or IRT [1] is even more accurate because it gives an estimate of each individual test-taker's ability while SEM gives an estimate for the whole group (Hughes, 2003:42). IRT is an excellent measurement tool for estimating both candidate ability and test task features such as degree of difficulty and bias. This is often used in computer-adaptive testing where the ability level can be adjusted to that of the individual (Bachman, 1990:7).

[1] Item Response Theory is a system of measurement which considers both the characteristics of the candidate and the item (Alderson et al, 1995:291).

2.6 USES FOR TESTS

There are five main uses for tests, namely: achievement; diagnostic; placement; proficiency and progress. It is possible for these to overlap; for example, a measure of proficiency can be used in placement decisions (Davidson & Lynch, 2002:132). Each type of test is determined by the outcomes that the test is designed to measure. As was explained in Chapter 1, for the aims of this study, the focus will be on proficiency and placement testing.

2.6.1 Proficiency tests

According to McNamara (2000:8), proficiency tests look to a future situation of language use. Inferences drawn from proficiency tests are used to predict the candidate's ability, or lack thereof, to deal with the various language tasks that will be required of him/her in the future (McNamara, 2000:8). The purpose of such tests is to assess an individual's potential ability in a language. The content is based on tasks, usually for a particular purpose, which a candidate would have to be able to perform in order to be considered proficient. Some proficiency tests are more general, such as the Cambridge Proficiency in English (CPE) test which merely assesses whether a candidate has reached a certain level of ability. A characteristic shared by all proficiency tests is that they are not based on a syllabus or course that candidates may have taken in the past (Hughes, 2003:12).

Since proficiency testing yields information on a student's ability to perform in future target situations, the appropriateness and level of the items are very important when selecting the task content. The more authentic the language used in the test, the better the evaluation is likely to be (Weir, 1993:6). According to Weir (1993:19), a good test has to reflect the *realistic* use of the particular ability to be measured.

There are a number of proficiency tests (with listening components) that are used internationally to assess entry level students at universities and colleges. The practical advantages of computerised technology has had an important impact in this field and many well known language test organizations are moving towards computer based testing. TOEFL (Test of English as a Foreign Language) and the University of Cambridge Local Examinations Syndicate (UCLES) are both now delivered by computer. Web-based language tests abound and tests have been developed for the EU through the DIALANG (Diagnostic Language Assessment) project (Chapelle, 2001:2).

Although the criteria, methodologies and objectives underlying the various tests vary, there are also distinct similarities or common denominators. The two best known and most widely used tests are the International English Language Testing System (IELTS) and TOEFL. IELTS

is produced by Cambridge University in the United Kingdom and is similar to TOEFL, an American test, in both scope and method. The Cambridge Proficiency in English (CPE) test, although not specifically for academic purpose testing, is recognized by British universities for admission purposes.

2.6.2 Placement tests

As can be inferred from the name, the purpose of these tests is to provide information that helps to place students into teaching programmes most suited to their abilities. They are usually used to sort students into classes of differing levels. The most successful placement tests are those which are designed and developed for specific contexts (Hughes, 2003:16).

Like TALL, the Placement Test in English for Educational Purposes developed by the University of Cape Town was designed to assess entry-level students' capacity to cope with the tasks which would be required of them at tertiary education level (Cliff, 2003; Van Dyk & Weideman, 2004). Similarly, the California State University developed a test to assess academic literacy skills in limited English proficient (LEP) students. In this test a video presentation of a lecture was used to simulate the real world as far as possible (Kuehn, 1996:13).

2.7 SPECIFIC PURPOSE TESTING

Effective tests should include content that is relevant to a candidate, combined with a testing practice that is appropriate to their needs and interests. This has given rise to specific purpose language tests whose content is relevant to a particular setting such as the academic environment of a university (McNamara, 2000:49), as is the case in this study. Thus, the aim is to predict the abilities of an individual in a *specific* context. This is based on their performance in a test where the tasks reflect the future language use situation. Clearly, the first step in the process is to identify the TLU domain as well as the characteristics which describe it. Only then can tasks be designed that will accurately mirror it and consequently lead to accurate inferences being made on a candidate’s ability in this setting in real life (Douglas, 2000:42). In the development of ALT, tasks were selected that are representative of those usually required of first year students at a university. As is the case with TALL, these would include abilities such as identifying main themes and making inferences.

Many incoming university students are likely to be under-prepared for what will be expected of them in their first year of study. A low level of language competency could, for example, contribute to their lack of success at university (Cliff, 2003:2; Van Dyk & Weideman: 2004a:9). Van Schalkwyk (2008:39) believes that before students can even begin to interpret or analyse

an academic text, they need to have certain competency levels in the language. Thus, even though language competency is not the only consideration and some might even feel it plays a relatively minor role, according to Elder, Erlam and von Randow (2002:1), there appears to be a threshold of proficiency below which students are unlikely to cope with academic study.

Lecturers at the University of Stellenbosch have noticed an increasing lack of critical and analytical thinking skills amongst their students. The concept of building an argument by providing the necessary evidence as well as distinguishing between fact and opinion seems to pose a problem for some students. This appears to indicate a lack of the necessary academic literacy skills required for success at a tertiary level (Van Schalkwyk, 2008:2).

Tests of academic literacy are, therefore, designed to assess the degree to which students possess the necessary linguistic capabilities to cope with university courses. In a country such as South Africa, with its widely diverse population, range of schooling standards and socio-economic situations, this seems to be of particular importance (Cliff, 2003:2; Van Dyk & Weideman 2004a:2). Since school-leaving results may be inadequate in reflecting the potential of entry-level students to succeed in higher education (Cliff, 2003:2; Van Dyk & Weideman, 2004a:9), test designers have to identify the kinds of tasks students will be called upon to perform in real-life situations and then attempt to replicate them as closely as possible in the test. As mentioned in the previous chapter, at the University of Stellenbosch, academic literacy levels (in English) are presently assessed by means of TALL which is largely a test of academic reading and writing skills. All entry-level students take the test and the results are used to grade students according to their academic literacy competency. The test is also used as a means of sorting students into programmes where communicative language and thinking skills, required for effective study at university, are taught.

Cooper and Van Dyk (2003) maintain that vocabulary is a good predictor of academic performance, but, according to Weideman (2003), a test of academic literacy has to test much more than vocabulary and grammar. Students at tertiary level have to be able to perform tasks that go beyond mere comprehension; interpretation and critical evaluation are vital skills if they are to succeed at university.

In order to determine the level of academic literacy of incoming students at the University of Stellenbosch, the following skills were identified as being the most important and as such, are included in the construct of TALL (Weideman, 2003a:xi).

- To understand a range of academic vocabulary in context;
- to interpret and use metaphor and idiom, and perceive connotation, word play and ambiguity;

- to understand relations between different parts of a text, be aware of the logical development of (an academic) text, via introductions to conclusions, and know how to use language that serves to make the different parts of a text hang together;

- to interpret different kinds of text type (genre), and show sensitivity for the meaning that they convey, and the audience that they are aimed at;

- to interpret, use and produce information presented in graphic or visual format;

- to make distinctions between essential and non-essential information, fact and opinion, propositions and arguments; distinguish between cause and effect, classify, categorise and handle data that make comparisons;

- to see sequence and order, do simple numerical estimations and computations that are relevant to academic information, that allow comparisons to be made, and can be applied for the purposes of an argument;

- to know what counts as evidence for an argument, extrapolate from information by making inferences, and apply the information or its implications to other cases than the one at hand;

- to understand the communicative function of various ways of expression in academic language (such as defining, providing examples, arguing); and make meaning (e.g. of an academic text) beyond the level of the sentence.

Although the requirement at the University of Stellenbosch is that students ought to be proficient in both English and Afrikaans, in this study I will be focusing on academic listening abilities where English is the medium of instruction. According to Fulcher (1999:234), it is important to point out that a test is not testing listening skills in academic English but rather those of English in an academic context. The focus is thus not on content validity as such but on the relevance and representativeness of the content against the backdrop of construct validity.

2.8 RESEARCH INTO LISTENING AS A CONSTRUCT

Before embarking on the design of a test of academic listening competency as was the case in this study, it was necessary to do thorough research into the listening process as it is perceived by various scholars. Literature on listening strategies and factors that affect listening comprehension also provided essential information on listening as a construct.

Researchers have only relatively recently begun to recognize the important role listening plays in comprehensible input (Krashen, 1985) as well as its significance in general communication

abilities. Alderson and Buck, in their editors' notes in the book *Assessing Listening* (Buck 2001:x), state that 'the assessment of listening abilities is one of the least understood, least developed and yet one of the most important areas of language testing and assessment'. Buck, (1991:67) blames this neglect on the absence of a generally accepted theory of listening comprehension. Researchers have yet to reach consensus on a general definition of listening comprehension (Dunkel et al., 1993:180). This could be a result of the complexity of the listening process and the difficulty of measuring such an intricate operation. However, there does seem to be some agreement on the *characteristics* which make up the listening process (Lynch, 1998:3; Brindley, 1998:172).

2.8.1 Listening as a two stage process

Test developers agree that there are identifiable listening comprehension skills and that these can be ordered from lower order (literal understanding of a text) to higher order (inference and critical analysis) (Rost, 1990:33; Weir, 1993:98). Buck (2001:2) describes the two approaches to listening comprehension as the bottom-up and top-down views. Both these views involve the order in which a listener applies his/her different kinds of knowledge. The bottom-up view is where the listener starts by deciphering a syllable and ends with a paragraph which he or she then processes so as to elicit meaning from a text. The top-down view is concerned with predicting or inferring what is going to be said by using ones strategic competence. According to Rubin (1994:210), top-down processing focuses on meaning and involves the interaction between the lexical, syntactic and interpretation of knowledge.

To sum up, researchers tend to agree that when processing input in a language there is no definite order and some steps may happen simultaneously, which is referred to as parallel processing (Rubin, 1994:211). Listening comprehension is usually a top-down process since different kinds of knowledge come into play in no particular order. Listeners engage in an interactive process where any available information is used to create meaning. This may, for example, take the form of linguistic knowledge, sound input or content identification (Buck, 2001:3). Factors such as the difficulty of the text, the listener’s background knowledge and language competency level have an important influence on how and when listeners use the various processing strategies (Rubin, 1994:211).

These two stages comprise the framework for the listening test designed for this study and details of the various skills and listening strategies which fall within the bottom-up and top-down processing categories will be revisited in Chapter 3.

2.8.2 Theories and approaches to listening comprehension

As mentioned in 2.8, there are many different approaches to listening comprehension. However, most researchers seem to agree that all listening comprehension involves the use of both linguistic and non-linguistic knowledge. Linguistic knowledge includes phonology, lexis, syntax, semantics and discourse structure as well as the ability to interact with the input in real time (Buck, 2001:3). Non-linguistic knowledge is concerned with such aspects as contextual knowledge (Buck, 2001:2; Lynch, 1998:3).

Scholars in the field, such as Anderson, Buck, Brindley, Inbar, Lynch, Rost and Shohamy to name a few, have all contributed to the research on listening comprehension. In the paragraphs that follow, I have given a brief synopsis of the most commonly cited opinions and theories of some of these researchers.

- According to Buck (1991:67), listening comprehension goes far beyond the mere application of language knowledge in order to interpret a text. It is a process whereby listeners extract meaning based on their own knowledge and experience. He believes that since comprehension takes place in the listener's mind, the setting or context for 'interpretation is the cognitive environment of the listener'. Because of the lack of visible signs of comprehension, task performance has to form the basis of inferences about the extent of understanding (Buck, 2001:99). Buck's (2001:31) definition of listening comprehension is that it is 'an active process of constructing meaning … by applying knowledge to the incoming sound'.

- Like Buck, Rost (2002:59) believes that listening comprehension is a process whereby language is linked to previously stored notions and associations in real life. In other words, it is the understanding of what the language is referring to, based on one's past experience or knowledge. Rost elaborates on this point by saying that the merging of new information received by the listener with what he or she already knows is central to the comprehension process. One needs to retrieve stored information from one's memory in order to understand and process the new data. When this background knowledge is activated, the listener undergoes an affective response which influences his/her reaction to what is being said (Rost, 2002:63). This is of course in line with cognitive/constructivist theories that new meaning is constructed from prior knowledge. When listening in one's mother tongue, the meaning of an utterance is seldom understood on a purely lexical basis. The listener needs to interact with the text by assessing it in terms of personal opinions and interests. A listener usually relates what is heard to what he/she already knows or has experienced and supplements any missing information (Buck, 2001:19; Rost, 2002:150).

- Shohamy and Inbar (1991:26) are of the opinion that listening comprehension entails an ability to receive and interpret input simultaneously since a listener cannot replay what he/she has heard. A competent listener is therefore dependent on instant comprehension as well as an

ability to remember information. Unlike written text, utterances are dependent on the context and are usually unedited. In accordance with the ideas of Buck and Rost, Shohamy and Inbar also believe that listening comprehension necessitates an interaction between the listener's background knowledge and the spoken text.

- Listening is described by Bejar, Douglas, Jamieson, Nissan and Turner (2000:2), as 'the process of receiving an acoustic signal which is then structured'. They go on to say that the reception of the signal occurs in real-time but the structuring is dependent upon cognitive processes. These processes involve linguistic, situational and background knowledge which have to be synthesized in order to achieve meaning. Because of the wide range of listener variables such as differences in knowledge, memory capacity and cognitive abilities, the results will vary from individual to individual (Bejar et al., 2000:4).

- Similarly, Anderson and Lynch (1988:11) maintain that listener performance in comprehending oral input is affected by listening purpose and background knowledge as well as the capacity to store information. This can result in listeners formulating their own interpretation of an utterance which is different from what the speaker intended.

- Listeners often need to transcend the literal meaning of what has been said to infer meaning. Thus, there is a clear distinction between hearing what has been said and understanding the meaning. Widdowson (1983:13) terms these two concepts as systemic or linguistic knowledge and schematic or non-linguistic information. According to Anderson and Lynch (1988:14), both the systemic or perception of speech and schemata or interpretation, are essential for effective listening comprehension.

- Listening has often been thought of as a receptive and thus passive skill. However, Anderson and Lynch (1988:6) insist that a successful listener is definitely active even though the processing strategies are internal and therefore cannot be seen. 'Effective listeners actively engage in the process of comprehension' and 'do not passively receive and record'. Buck (2001:12) too, supports this theory in his description of how the listening process occurs. He suggests that a listener interprets the incoming acoustic signal by using different kinds of knowledge for specific communicative reasons. He/she then uses relevant information to extract meaning from the utterance. Buck (2001:29) believes this would be impossible without actively listening.

- Brown (1995:24) divides listening comprehension into four stages. The first stage is the identification of the information. Then the new information is integrated with what the listener already has stored in his/her memory. Thirdly, the incoming information is filed and finally used as and when appropriate. Brown also emphasises that each listener has his/her own motive for listening. Brown's model seems to point to a sequence which is followed in order to understand spoken discourse. However, as Buck (2001:3) mentions, it is more likely that there is no definite order to the process and that simultaneous or parallel processing is more likely.

- Brindley (1998:181) also defines the listening process as multi-faceted and interactive where 'listeners use verbal and non-verbal cues to interpret messages'.

- Listening comprehension, according to Hansen and Jensen (1994:243), is not so much a single process as the result of a series of processes. Their research has shown that there is usually significant interaction between higher and lower levels of processing (Hansen & Jensen, 1994:244). This concurs with Buck's (2001:2) findings that listening is an interactive process that occurs in no specific order.

Given the different theories and approaches to listening comprehension and in spite of the fact that there is some consensus, it is clear that listening researchers are faced with multiple challenges when trying to assess an invisible and automatic process. To add to the complexity of the process, listeners will have differing levels of linguistic abilities and background knowledge and will therefore use different processing strategies (Lynch, 1998:6). It is for this reason that test designers resort to assessing the more measurable skills associated with listening rather than attempting to test the process itself (Brindley, 1998:172). A discussion of some of these skills and the rationale behind their inclusion in ALT will be addressed in greater detail in the next chapter.

2.8.3 Listening strategies

As has been suggested in 2.8.1, the two listening stages are not directly visible, so they have to be assessed through more measurable skills such as recognizing facts and identifying main and supporting ideas. These may be either bottom-up or top-down abilities. The so-called global skills such as 'listening for gist' (Weir, 1993:98) involve processing the information from general to specific (Shohamy & Inbar, 1991:26). Global coherence strategies are in line with the top-down approach to processing. Mother-tongue listeners are more likely to process text in this knowledge-based fashion. The other method of comprehending spoken text is data-driven or local which features the bottom-up method of processing. However, in order to effectively contextualize and interpret the finer details in a passage the listener has to refer to the framework of the global perspective. According to Shohamy and Inbar (1991:36), this seems to be an approach that is popular with less proficient language users.

Competent listeners employ both these procedures to interpret a text, since both are necessary for effective interpretation. Local strategies provide evidence in support of global speculation. However, if only local strategies are used, the information has no reference and comprehension is impaired (Hansen & Jensen, 1994:244).

Theorists agree that the knowledge-based approach, mentioned above, is the more effective strategy and that a skilful listener will identify the main theme and then analyse the details, rather than the other way round (Shohamy & Inbar, 1991:26). Hansen and Jensen (1994:260) have confirmed this in their research. Their survey showed that more proficient listeners performed better on both local and global questions but the difference in the top-down approach was more marked than on the bottom-up strategy. Hansen and Jensen (1994:265) also maintain that their findings pointed to the use of predominantly bottom-up processing strategies by less competent students. They found that these students depended on extracting the details directly from the text and found it difficult to consolidate assorted pieces of information to arrive at a response. They offer this as an explanation for the marked drop in performance on global questions amongst less proficient candidates. However, according to Buck (2001:106) and Brindley (1998:173), it is extremely difficult to apportion different abilities accurately to set cognitive approaches. They maintain that the two listening processes are so interconnected that it is hard to distinguish between these two processing levels. In the same way, responses in a test are also difficult to attribute to any one skill. Rubin (1994:211) reasons that variance in the characteristics of the candidates makes firm decisions on processing patterns very difficult. A listener's purpose for listening will also alter the way that he/she processes the information, for example, if a listener needs to analyse and evaluate, critical listening needs to take place (Rost, 1990:9).

2.8.4 Inference

World knowledge is applied through the process of inferencing (Buck, 2001:18). Listeners have to infer meaning from what they hear because they do not have access to the speaker’s intended meaning (Rost, 2002:64). Acceptable understanding is reached when both the speaker and the listener are satisfied with the inferences that are made. Misunderstanding can occur when there is conflict between what the speaker intended and the interpretation of the listener (Rost, 1990:62). Lynch (1998:6) supports this idea by saying that the speaker and the listener need to be 'on the same page' or have a mutual cognitive environment.

Inference is a process which requires language ability as well as problem-solving and thinking-skills (Rost, 2002:64). Buck (2001:60) defines inferencing as the ability to deduce information based on background knowledge of the topic and being able to fill in any gaps through knowledge of the context. There are several kinds of inferencing, the most important being: perceiving references with more than one meaning, filling in missing information and predicting the speakers' purpose. Thus, although conventional knowledge of language structures is central to comprehension, some portions of meaning are likely to be implied rather than explicit (Rost, 2002:65). Research has shown that testing a listener's

comprehension of explicitly stated or literal information is far easier than measuring a listener's inferential skills (Buck, 1991:86). Despite the difficulty in assessing implicit comprehension, many language testers agree that such tasks ought to be included in listening tests (Buck, 2001:60; Weir, 1993:98).

2.8.5 Listening compared to reading

A comparison of listening and reading skills comprises a necessary part of this project since the results of the listening test used in the study will be compared with TALL, which is a test of reading (and writing). Since the tests assess different language skills, one might expect a low correlation but the fact that they are both receptive skills and both measure an aspect of language proficiency could raise the correlation.

Buck (2001:4) maintains that there are three important considerations when comparing the comprehension of spoken language and written text. The first is that 'speech is encoded in the form of sound; secondly, it is linear and takes place in real time, with no chance of review; and thirdly, it is linguistically different from written language'.

2.8.5.1 Similarities

Cognitive language theorists agree that listeners employ the same processing schemata as they would in other sensory processing, such as reading (Alderson, 2005:138; Anderson & Lynch, 1988:18; Lynch, 1998:10). This could account for the large number of researchers who believe that testing of listening proficiency has much in common with that of an assessment of effective reading. Firstly, they are both receptive skills and secondly they are both cognitive in nature. This makes assessment much more difficult (Weir, 1993) than in the case of writing and speaking, which are productive skills. In both reading and listening, there is a sequence involved in the processing of information. Both readers and listeners would be able to form an overview or extract the gist from a passage. Both skills require an understanding of individual sentence meaning as well as identifying the overall theme of a text in order to fully comprehend the meaning (Anderson & Lynch, 1988:18). It is possible to determine the opinion of the writer or speaker, as well as listen or read for specific detail. Meaning can also be inferred from both written and spoken texts. Main ideas are explicitly or implicitly stated in both types of text and background information can to be integrated with information given in the text to make pragmatic inferences (Alderson, 2005:138; Buck, 2001:19). In spite of these similarities in text interaction, the skills needed by a listener to decode information are very different from those needed for visual recognition (Rost, 1990:9).

2.8.5.2 Differences

One of the main differences between reading and listening is the effect of speaker-related variables on the process. These variables include factors such as stress, rhythm, intonation, pauses, fillers and redundancy of information (Weir, 1993:109). A listener's comprehension could be significantly hampered by the speaker's pronunciation or prosodic features such as stress and intonation, which are not present in written text (Shohamy & Inbar, 1991:25; Alderson, 2005:138).

Readers can identify key words in a text at leisure, while listeners have to recognize the sounds and words automatically and remember them, since listening occurs in real-time. There is also no opportunity for a listener to go over what has been said, as would be the case in written text (Lynch, 1998:11; Weir, 1993:101; Alderson, 2005:138). Lund (1991:198), in his comparison of reading and listening methods, found that while readers could recall more information and in greater detail, listeners were better at identifying the main themes and inferencing.

Lynch (1998:10) adds that spontaneous speech is not usually clearly chunked as is the case with written text in the form of paragraphs and headings, and is often not fluent.

According to the researchers of the Cambridge Proficiency in English tests, another of the major differences between reading and listening is that with read text, a student can determine his/her own speed, whereas when listening, the speed is determined by the person talking, although the context, the topic and the type of communication also play a role (Developing general listening skills, [s.a.]).

2.8.6 Factors affecting listening comprehension

An analysis of the factors which affect listening comprehension is essential for reliable interpretation of listening assessment results. Careful consideration of the aspects listed and discussed below had to be implemented in the selection of texts and tasks to be included in ALT; these will be revisited in the next chapter. As the literature points out, wherever possible, the effects of factors such as speech rate, accent and text type should be kept to a minimum so as to make the assessment as fair as possible.

2.8.6.1 Speech rate

Research has found, and it seems logical to assume, that the quicker the rate of speech, the harder it is to comprehend. More informal speech, such as conversations and interviews, has an average rate of 210 and 190 words per minute respectively. A radio monologue, on the

other hand, is much slower at 160 words per minute and a lecture is the slowest rate of speech at 140 words per minute. Lecturers are inclined to speak more slowly to assist in comprehension. The speed at which utterances are delivered has also been found to have a more significant impact on comprehension than when the speaker has an accent that varies significantly from the standard (Buck, 2001:40). Rubin (1994:200) believes that when examining the effect that speech rate has on listening comprehension, one has to take into account the type and difficulty of text, as well as the background knowledge required to understand the text. She also maintains that interpretive and short-term listening are not as affected by the rate of speech as is the case in comprehensive listening.

The listening clips included in ALT, feature speakers with varying rates of speech and these will be discussed in the test description in the next chapter.

2.8.6.2 Phonology

Buck (2001:32) states that 'the sounds of a language must be learned in order to understand speech' however, he goes on to say that it is 'not the sounds themselves … that cause the most comprehension problems, but the way they vary in normal speech'. These sounds change depending on the speaker, the register or situation in which they are used and where they are situated in a sentence. In a formal setting, such as a lecture, there would be less modification of sounds and thus clearer pronunciation than in conversational speech. The most important changes are:

- assimilation, when sounds affect one another by being adjacent to each other,
- elision, when sounds are omitted because of fast paced speech and
- intrusion, where a new sound is introduced between two sounds. An example of this is if two vowel sounds meet then an extra sound is often inserted such as in the phrase 'please do it'. Here, a /w/ sound is included between the words 'do' and 'it' (Buck, 2001:32).

2.8.6.3 Accents

Everyone speaks with an accent, usually as a result of geographical factors or, if the speaker's mother-tongue is not English, his/her accent would reflect the elements of his/her first language. Accent is a significant variable in listening comprehension as it can make understanding very difficult, if not impossible. Here, the language ability of the listener plays a role, with non-native speakers being more likely to pronounce words 'in a non-standard manner' (Buck, 2001:34). The texts used in ALT feature South African English and Afrikaans

mother-tongue speakers because these accents are typical of the majority of lecturers and students at the University of Stellenbosch.

2.8.6.4 Prosodic features

According to Rubin (1994:202), listeners, generally, do not pay enough attention to the meaning conveyed by different stress and intonation levels. Both stress and intonation are important word recognition tools and often signify a meaning that either emphasizes or differs from the literal meaning of a word (Buck, 2001:35-6). Recognising stress and intonation is particularly important in spoken English because word boundaries are not easy to identify (Lynch, 1998:4). Being aware of the rhythm of speech, created by patterns of stress and intonation, is fundamental to effective listening comprehension (Buck, 2001:36). Wagner (2002:7) supports Buck's theory and maintains that because of the importance of prosodic features in understanding spoken language, they should form part of the construct definition of all tests of listening comprehension. He emphasises that the prosodic patterns in read texts are completely different from those found in free speech.

Prosodic awareness in spoken discourse, directly affects how listeners organize and interpret what they hear. Evidence has shown that listeners make an effort to identify accented words because they are an indication of which information is important (Lynch, 1998:5).

2.8.6.4.1 Stress

The three main types of stress are syllable, word and sentence stress. Stressed syllables in English are pronounced more clearly and are often louder than the rest of the word. Different syllables in a word may be stressed which change the meaning of the word. In sentence stress, the emphasis is placed on the most important words in the sentence. This signifies the point the speaker is trying to make. The words that are not stressed are often spoken more quickly and in some cases the sounds are run together so rapidly that the words almost disappear completely. This is because English is a stress-timed language (Buck, 2001:35).

2.8.6.4.2 Intonation

This is related to change in pitch in spoken language. It has a number of important functions:

1. Emotional: it shows the speaker's feelings towards, or opinion of, a particular subject.

2. Grammatical: intonation performs the role that punctuation does in written text.

3. Informational: as with sentence stress, important information is often spoken in a higher pitch.

4. Textual: in written language, paragraphs are used to make the text cohesive. In the same way, intonation is used to indicate contrast, agreement or a different idea.

5. Psychological: intonation can be used to chunk information to make it easier to remember.

6. Indexical: in this role, intonation can be used as an identifier, for example, newscasters may have a kind of recognizable signature intonation when introducing themselves.

Rising intonation at the end of a sentence is often associated with an enquiry or question.

2.8.6.5 Hesitation

There are four main types of hesitation in spontaneous speech:

1. Unfilled pauses (silence).

2. Filled pauses such as *um*, *mm* and *anyway*.

3. Repetitions, where a speaker repeats a word or part of a word.

4. False starts, where a speaker stops what he/she is saying and substitutes another word or phrase in place of what he/she has just said (Buck, 2001:41).

These hesitations characterize spoken language that is not read and are essential for the construct validity of a listening test (Buck, 2001:85; Messick, 1996:243). Research has shown that pauses, both where they occur, and their frequency, seem to have more of an impact on listening comprehension than the rate of speech (Lynch, 1998:9). Hesitation can present problems to second language speakers and errors in perception are substantially affected by hesitations. On the other hand, it can also be an asset to comprehension since it slows down the speed of the utterance, making it easier to understand. However, if a listener is unable to recognize the pauses as fillers then comprehension could suffer (Buck, 2001:41).

2.8.6.6 Text type

Shohamy and Inbar's (1991:23) study found that the type of text affected the performance of candidates on listening comprehension tests. The more facts contained in a text, for example, a news bulletin, the more difficult understanding it proved to be. In accordance with Buck's (2001:43) findings, Shohamy and Inbar maintain that more informal speech, which commonly

contains redundancies, was found to be easier to comprehend. Providing students with extra or redundant information can aid comprehension but a surplus of background detail can detract from the main idea and have the opposite effect (Rubin, 1994:203).

A listener uses periods of rephrasing or pausing to analyse and process the information he/she has heard. If there is an intensive rate of factual input it stands to reason that a listener would have more difficulty in comprehending it. The more formal structure of literate texts, such as lectures, which often use the passive voice, may be harder to comprehend than the relatively simpler structure of the more informal or oral texts (Shohamy & Inbar, 1991:34). Dunkel and Davis (1994:56) make the point that understanding lectures involves much more than merely being able to interpret individual words and sentences. They maintain that competent lecture comprehension is dependent on understanding how the sentences inter-relate to form the structure of the whole text. Text characteristics that have been found to aid in lecture comprehension are repetition and clear speech markers. According to Buck (2001:42), academic lectures vary so significantly from informal speech that they can be considered a distinct domain, separate from other language use.

2.8.6.7 Non-verbal communication

Not all important or relevant information is transferred by language. Listeners may not always interpret non-verbal information accurately but seem to decode it more effectively than much of the verbal communication (Buck, 2002:46). Although considered an important aspect of listening, non-verbal communication is often overlooked in selecting the mode of delivery of listening texts (Kellerman, 1992:241).

The role that visual input plays has been the subject of much research over the last decade and according to Lynch (1998:7) visual delivery has been found to increase the levels of motivation and attention as well as 'improve … comprehension of gist'. However, Lynch (1998:7) and Buck (2001:47) state that there is, as yet, no conclusive evidence that visual input is helpful as a long-term memory aid or that it necessarily improves comprehension. Bejar et al. (2000:12), on the other hand, report that there are some who firmly believe that it assists in listening comprehension.

Speakers often express their feelings or opinions with physical gestures and this can affect the interpretation of the listener. Message-related movements can include hand and facial expressions which reinforce certain points. The effect of this body-language, particularly hand signals, can significantly influence the way a listener processes what is heard (Buck, 2002:46; Jordan, 1997:184). A speaker's body-language is also linked to syllables or words which are

stressed. These are important for recognizing new or significant information. The use of video in listening testing, therefore, seems to be a logical solution as it allows listeners to process visual as well as verbal input (Wagner, 2002:7).

2.8.6.8 Listener variables

General language proficiency, memory capacity, background knowledge and processing strategies are a few examples of the factors that seem to play the most significant role in an individual's listening comprehension ability. Various studies mentioned in Rubin's (1994:211) article, seem to indicate that the cognitive processes such as storing and retrieving information, are affected by an individual's linguistic abilities. This research has also shown that generally, the lower the linguistic proficiency of the candidate, the more the reliance on syntactical rather than contextual information. However, Rubin (1994:210) concludes that most listeners, despite their linguistic competency employ text-based or bottom-up strategies on more complicated texts.

Call (1985:768) suggests that an individual's memory capacity for linguistic input is an effective measure of general language proficiency. This seems to indicate that short-term storage of information does not necessarily make for effective listening comprehension but can aid in concept and detail identification. Lynch (1998:4) bemoans the limitations placed on human beings as processors by their memory capacity. He proposes that a second language speaker would be obliged to devote much of his/her processing space to decoding the lexico-grammatical aspect of an utterance which would take precedence over higher level processes such as inference.

Background knowledge has a major impact on listening comprehension since listeners relate incoming information to a general schema. How effectively listeners use their background knowledge to correctly interpret meaning, still needs more research. Studies have shown, however, that background knowledge of a specific topic improves listening comprehension (Rubin, 1994:209).

In conclusion, misunderstanding can occur either through speaker error, or more commonly, a listener's unfamiliarity with the terminology as used in a formal context. However, different accents or dialects of speakers are some of the most common reasons for intelligibility problems. Intelligibility can also be affected by the quality of the speaker's voice and the clarity of articulation. Phonological structures may differ from one environment to another and a good listener needs to understand the rhythm of the language before efficient auditory perception can take place. Identification of the context of the aural input is also essential for

comprehension of spoken discourse (Rost, 1990:54-7). Buck (2001:48) states that although the processes are the same for first and second language speakers, the lack of either linguistic or background knowledge, or both, in second language speakers, could be a major disadvantage in fully comprehending spoken discourse.

Abilities, other than listening comprehension, may be involved which are not included in the listening construct, for example, computer skills or ability to read with comprehension. Here, it is necessary to consider how much variance this is likely to cause. If all the test-takers have similar levels of competency then the variance will be kept to a minimum. However, it is important to prevent a candidate's degree of computer literacy from becoming an additional variable affecting performance in a computerised language test (Bejar et al., 2000:29).

2.9 ACADEMIC LISTENING

The TLU domain of the listening assessment employed in this study is set within the context of a university with its accompanying features such as lectures and tutorials. Since these form the basis of the tasks included in the test, it was essential to investigate the opinions and findings of experts in the field of academic listening.

According to Ferris and Tagg (1996:299) and Kuehn (1996:29), academic listening tasks have proved challenging for all students, even those whose mother-tongue is English. This is corroborated in Flowerdew (1994:11) who found that academic listening is significantly more complex than, for example, conversational listening. Rost (2002:162) maintains that this is because academic listening is mostly a non-collaborative or one-way listening process where a lecture is the most typical example. Since lectures play such an important role in any academic programme, effective listening in lectures is fundamental to any student's success at university (Flowerdew, 1994:7).

Because of the importance of understanding lectures and taking effective notes, a substantial portion of ALT has been devoted to a lecture task. Decisions on the type and length of text, as well as the questions included in the lecture task, were modelled on the findings and suggestions of the researchers mentioned below. These will be discussed in detail in Chapter 3.

Buck (2001:43) termed academic lectures a distinct domain, and as a result, lecture comprehension has characteristics which differ from other types of listening. For example, lectures usually contain formal vocabulary with more complex sentence structures. Global skills necessary for identifying the main points are of major importance. It is also essential that students are able to concentrate on and process long pieces of information (Flowerdew,

1994:11). Chaudron, Loschky and Cook (1994:75) believe that there are some strategies which are paramount for lecture comprehension. These include eliciting the main ideas and supporting details, identifying key phrases and noting words that signal discourse relevance. According to Jordan (1997:181), it is also crucial that note-takers are able to discriminate between important and less important information.

As the literature repeatedly attests, language skills cannot be dissociated. A good example of this is in a lecture situation where listening, reading and writing are completely integrated. Students listen to a lecture, take notes and then use the notes for study or assignment purposes. The development of these skills often presents a problem for students (Jordan, 1997:9) often as a result of unfamiliarity with the terminology associated with a specialist subject. This problem is exacerbated for second language speakers by a general lack of semantic knowledge.

Since the aim of the lecturer is to deliver information to an audience, he/she usually tries to do this in a systematic fashion. Thus, it is usual to present information in a lecture in a logical and sequential order so as to make it easier for the listeners to understand and retain (Hansen & Jensen, 1994:246). Lecturing styles, such as rate of delivery and repetition of important points, contribute to better comprehension of lecture material but it is believed that a conscious mental process related to understanding, storing and using linguistic knowledge is still the most important factor in improving comprehension. Thus, in spite of the non-interactive nature of a lecture, in that the listener is not expected to respond orally, it is still a communicative activity.

A lecturer's purpose, apart from providing information, is to influence with intent. This means that the aim is to enlighten or raise awareness as well as to encourage the students to form their own opinions about the concepts presented. Where listening differs from reading is that visual or aural clues enable the listener to perceive the lecturer's view of the topic. It also makes it easier to perceive the gist of the material. However, by reading a text, it is easier to remember the important facts and specific details, hence, the need to take notes in a lecture (Rost, 2002:162).

Jordan (1997:179) identified three main areas which pose a problem for students in a lecture situation when it comes to listening and note-taking. The first challenge is to decode the incoming material. This involves not only understanding what has been said but also recognizing pauses, fillers, false starts as well as patterns of stress and intonation. As has been suggested, the lecturer's accent and speech rate also have an impact on the level of comprehension. The second requirement is the ability to identify and extract the main and

subsidiary points. Note-takers have to be able to discriminate between the most important information and the supporting detail. The last area which could present a problem links to the second and involves the ability to write down the key information quickly without getting left behind. According to Chaudron et al. (1994:76), note-taking has two main benefits. The first is that it helps in organizing the lecture content while listening. The second advantage is that it assists in the recall of facts for study purposes.

The absence of visual signposts, such as headings and subheadings which would indicate new sections in a written text, forces the listener in a lecture to identify discourse markers which indicate a change of topic. These markers are also helpful in enabling the listener to recognise the main points (Hansen, 1994:134). Most lecturers make use of these and other listening cues to show their audience which ideas are important. The most significant of these, in addition to the lexical discourse markers, are prosodic features and the use of syntactic structures such as relative clauses or noun complements. Pitch or volume variance and speaking more slowly can also be an indication of emphasis (Jordan, 1997:183). Effective lecture listeners, therefore, have to be aware of these non-verbal signals which are so important for interpreting meaning.

Students at university need to adopt appropriate listening strategies depending on the type of lecture they are attending. In lectures where there is a large transfer of information, a listener would have to adopt an information-driven listening strategy which is necessary for absorbing facts. In lectures of a more problem-solving nature where discussion and academic argument are encouraged, students would be more likely to employ a point-driven listening strategy. Here they would need to make inferences about the speaker's purpose and intention (Jordan, 1997:182). The lecture clip included in this study falls into the first category since candidates are required to process and absorb information rather than make inferences.

In conclusion, the literature seems to highlight the highly complex nature of the listening process as well as the inherent difficulties in attempting to measure listening abilities. However, according to Brindley (1998:181), there is still a lot of research to be done which will hopefully lead to a better empirical basis on which to design listening test specifications. In the next section a review of the scholarship in the listening testing field will be discussed so that decisions made regarding the design of ALT, can be justified.

2.10 LISTENING ASSESSMENT

Listening testing has evolved over the past few decades and there has been a major shift away from merely being able to distinguish between phonemes and identifying patterns of

stress and intonation (Weir, 1993:99). Nowadays, the emphasis is on whether the meaning has been successfully communicated rather than just a literal understanding of the structure (Buck 1991:67; Rost, 1990:150). However, the intrinsic difficulties in testing an invisible mental process such as listening may account for the relative lack of 'empirically sound models of listening comprehension' which could be used as a framework for testing (Brindley, 1998:171; Buck, 1991:67; Dunkel et al., 1993:180).

2.10.1 Defining the construct of listening testing

The 'construct' or 'what is to be tested', as with all testing, is central to the assessment of listening abilities (Rost, 2002:170). According to Buck (2001:95), it has to be accurate both in theory and in practice as it will determine the meaning of the scores. Decisions based on scores will be invalid and unreliable if the construct is not accurate. When defining the construct of any test, which of course includes listening tests, the primary consideration has to be the *purpose* of the test. The significance of the test's purpose is the considerable effect it has on the construct that the test developer wants to assess (Buck, 2001:95). The purpose of ALT has already been addressed and the precise criteria used in its operationalization will be discussed in much greater detail in Chapter 3.

Rost (2002:170) believes that when attempting to measure an individual's auditory skills, it is important to isolate the characteristics of comprehension that are peculiar to listening. The more listening-specific features that are included in a test of listening, the better the construct validity of the test will be. The assessment of listening comprehension is complicated and time consuming because of the search for high quality and appropriate recordings. Therefore, unless the intention is to test listening skills in particular, the option to use paper based texts for tests of reading and writing makes much more sense (Buck, 2001:31).

According to Buck (2001:102), there are two methods that can be used to define the construct of a listening test. The first is concerned with the strategic and linguistic competence that a candidate should possess. The second involves the tasks that candidates should be able to perform.

2.10.1.1 Competency-based listening constructs

In this approach, the *ability* of the test-taker is fundamental to defining the construct. Test developers have to make decisions, based on the test purpose, on which kinds of competencies to include, how to incorporate them, and what their weighting should be (Buck, 2001:104).

According to Buck (2001:103), the differences between language competency and strategic competency are indistinct as a result of the difficulty in divorcing linguistic knowledge from general mental ability. However, variations in performance amongst second language speakers will more likely be due to the levels of language knowledge than the degree of strategic ability. He goes on to explain that in the case of adult second language speakers, strategic competence or cognitive abilities will be 'relatively developed and stable, whereas language knowledge is, by definition, only partially developed' (Buck, 2001:105).

Another dilemma is to identify precisely which abilities are required for which specific task. This presents construct definition problems because it is difficult to ensure that test-tasks are actually measuring the competencies they set out to assess. Buck (2001:106) and Brindley (1998:173) agree that items designed to measure a particular skill in a test might, in fact, be testing an entirely different one.

2.10.1.2 Task-based listening constructs

In this approach, the listening construct is defined in terms of what test-takers are capable of. According to Buck (2001:107), there are two steps involved: firstly, to identify the TLU tasks that will be required of the candidate and secondly, to simulate those TLU tasks as closely as possible in the test tasks. Careful consideration has to be given to what the test designer deems to be the most important features of the TLU tasks and which abilities are required for which tasks.

Students at university are likely to find themselves in listening situations where they are not expected to respond verbally, such as lectures and seminars (Buck, 2001:98). This explains why non-collaborative listening tasks are included in most academic proficiency tests. In the next chapter, the various abilities for the specific tasks included in ALT will be discussed in detail.

2.10.1.3 Combination of competence and task-based constructs

The most important consideration, according to Buck (2001:107), is not how closely the task resembles the TLU task but whether the interaction between task and test-taker is similar to what it would be in real life. In other words, similar abilities are required in both the test and the real world. Both ability and tasks form the basis of the construct definition and thus this combined approach seems the most sensible, since effective listening requires linguistic ability tempered by the setting. Therefore, in spite of the differences in the underlying considerations of these two approaches, Buck (2001:109) advocates an integrated approach

based on both ability and task. Buck's advice seems to encompass the best of both worlds and for this reason is the one I chose to implement for the purposes of this study.

2.10.2 Theories and approaches to listening testing

There are several different approaches when it comes to listening testing but for the purposes of this study, three of the main theoretical ideas will be discussed briefly. The first is not relevant to the test used in this particular study but serves as background for the other two.

2.10.2.1 Discrete-point testing

During the era of the language laboratory and the age of Behaviourism, discrete-point testing was the most widely used method of assessing language ability (Buck, 2001:62). Discrete-point testing is based on two main ideas, firstly that individual elements of language use can be isolated and tested separately and secondly, that spoken language is not too different from written language, it is just presented in another form. Dictation and matching-pairs tasks are typical of this type of testing. In light of the more modern approach to listening comprehension, these ideas were criticized and led to a shift towards more integrative and communicative testing (Buck, 2001:67; Hughes, 2003:19). It is for this reason that I decided to concentrate on tasks that were more communicative, relevant and authentic.

2.10.2.2 Integrative testing

Emphasis in this form of testing is placed on measuring how language is processed as a whole, as opposed to how much is known about the individual components that make up a language (Buck, 2001:67). Integrative testing requires a candidate to use a combination of language elements to perform a task, for example, to make notes while listening to a lecture (Hughes, 2003:19). The lecture task included in ALT advises candidates to make notes, which are not for assessment but to serve as a memory aid for the questions that follow. In addition, note-taking while listening to a lecture is a realistic academic activity and as such, adds authenticity to the task.

Reduced redundancy is a concept widely used in integrative testing. Cloze and gap-fill exercises are commonly used in this form of testing (Buck, 2001:67-8). This kind of task is also included in ALT where candidates are given a gap-fill exercise of a summary, rather than the full transcript, of the text.

Integrative testing has been accused of relating more to the first phase of listening, namely the literal meaning of an utterance. There is little call for inferencing and although the skills that are tested are fundamental to listening comprehension, the movement has been criticized for

assessing a range of language abilities that are too narrow. This is because the communicative function of language seems secondary to testing 'isolated events' where the listener is not required to integrate the information into a context (Buck, 2001:82).

2.10.2.3 Communicative testing

This kind of testing gained popularity as a result of the increased interest in communicative second language teaching and learning. The fundamental principle of the communicative approach is that the focus is on use rather than form: on whether an individual can actually use the language to interact in real life. Likewise in testing, there was less concern about an individual's linguistic knowledge and more about whether he/she could communicate successfully.

The tasks in a communicative test set out to simulate the features of target-language use in real life. Although such a test strives to contain tasks that are as authentic as possible, the mere fact that it is a test diminishes the authenticity (Buck, 2001:92). Communicative tests of listening ability should include tasks that evaluate 'higher-level cognitive skills' (Buck, 1991:69). Bachman (1990:356-7) identifies two types of communicative tasks: the first are situational tasks that are similar to those in the TLU domain and the second are interactional, where test-takers interact with tasks by using the same or similar competencies that they would in the real world. This corresponds with Buck's (2001:92) statement that instead of looking at the test itself, the focus should be on the 'interaction between the test tasks and the test-taker' which allows a more reliable prediction to be made about the communicative ability of the candidate. Buck (1991:69) believes that the listening purpose of the listener also influences comprehension and should therefore be included in any communicative listening ability test.

Communicative tests, however, are often less reliable or generalisable because the tasks have to be fairly context specific. The fact that there are so many different communicative settings means that a few examples have to be selected to represent an overview of what can be expected of the candidate in the TLU setting. There is also the matter of language use which varies from one individual to another. This, coupled with the fact that there is likely to be more than one way of interpreting a text, makes reliability of assessment very difficult, especially when candidates are asked to make inferences about a speaker’s implied meaning (Buck, 2001:84-5). In spite of this, since inferencing is such an important critical thinking skill, I decided to include this ability in my test even though according to Buck (1991:86; 2001:85), 'the linguistic system … is … much more amenable to testing'.

2.10.3 Context in listening testing

A test of language ability which is limited to a particular setting or domain is more relevant to a candidate than a more general approach. Listening assessment, like the testing of any macroskill, is based on an evaluation of sample performances. Thus, it is essential that the kinds of samples are an adequate reflection of a candidate's future performance in a particular context such as at a university (Rost, 1990:180). The TLU situation is an extremely important consideration since it affects the whole construct of the test. The test-taker is being assessed on whether he/she will be successful in a specific situation. Thus, the results generalise to the language-use domain since the TLU setting will be instrumental in the selection of tasks and competencies that need to be assessed (Buck, 2001:111). Bachman (1990) maintains that in order to assess effective communicative language competence, the language used must also reflect a real-life context, thus register and style have to be considered when content decisions are made. Based on the TLU domain, test developers decide on which types of topics, texts and tasks to include in the assessment. Test designers will also strive to simulate as many of the aspects of the TLU setting as possible (Bejar et al., 2000:5; Buck, 2001:111).

2.10.4 Skills central to effective listening

The literature suggests, fairly conclusively, that effective listening performance is partially *language* ability, which is concerned with interpreting the input and questions, and partially *procedural* ability, which enables a candidate to respond to the questions. In other words, the candidate needs to comprehend the text in relation to the task (Rost, 1990:181). Buck (2001:51) also mentions a widely used approach which divides the listening process into two parts. Part one involves absorbing the linguistic information and part two is concerned with communicatively processing it to infer meaning. These two stages interact and merge with one another.

There seems to be some consensus among researchers as to which skills should be assessed in tests of listening competency. Most agree that a communicative listening approach should include identifying the central theme of a passage and deduce a speaker's implied meaning (Wagner, 2002:2). According to Buck (2001:114), a good listening test should require a candidate:

- to process extended samples of realistic spoken language, automatically and in real time.
- to understand the linguistic information that is unequivocally included in the text, and
- to make whatever inferences are unambiguously implicated by the content of the passage.

Shohamy and Inbar (1991:29) divide their list of necessary listening skills into global and local categories. Under the global skills they list the ability to synthesize information, to draw conclusions, to focus on cause and effect and to make inferences. Their local skills include being able to pinpoint details, to recognize facts and paraphrase input. Candidates should be able to use both global and local processing skills and Weir's (1993:98-9) summary given in the table below seems to provide a comprehensive list of both these skills.

TABLE 2.2: A CHECKLIST FOR LISTENING COMPREHENSION ABILITIES

a) *Direct meaning comprehension:*

Listening for gist;
Listening for main idea(s) or important information; includes tracing the development of an argument, distinguishing the main idea(s) from supporting detail, differentiating statement from example, differentiating a proposition from its argument, distinguishing fact from opinion when clearly marked.
Listening for specifics; involves recall of important details.
Determining speaker's attitude/intentions toward listener/topic (persuasion/explanation) where obvious from the text.

b) *Inferred meaning comprehension:*

Making inferences and deductions; evaluating content in terms of information clearly available from the text.
Relating utterances to the social and situational context in which they were made.
Recognising the communicative function of utterances.
Deducing meaning of unfamiliar lexical items from context.

c) *Contributory meaning comprehension (microlinguistic):*

Understanding phonological features (stress, intonation, etc.)
Understanding concepts (grammatical notions) such as comparison, cause, result, degree, purpose.
Understanding discourse markers.
Understanding syntactic structure of the sentence and clause e.g. elements of clause structure, noun and verb modification, negation.
Understanding grammatical cohesion, particularly reference.
Understanding lexical cohesion through lexical set membership and collocation.
Understanding lexis.

d) *Listening and writing (note taking from lecture, telephone conversations, etc.):*

Ability to extract salient points to summarise the whole text, reducing what is heard to an outline of the main points and important detail.
Ability to extract selectively relevant key points from a text on a specific idea or topic, especially involving the coordination of related information.

(Weir, 1993: 98-9)

2.10.5 Content of listening tests

When selecting material for a listening comprehension test, decisions need to be made on a number of issues. These factors can be divided into three main categories.

1. Characteristics of the input, for example, accent, rate of speech and prosodic features.

2. Features of the test task, for example, clear instructions, type of context and amount of information given.

3. Individual factors, for example, memory, background knowledge and interest in the topic (Rost, 2002:175).

These considerations formed the platform for the decisions taken on the design of ALT. The specific type and mode of delivery of texts and tasks included in ALT, as well as the rationale behind their selection, will be presented in detail in the next chapter.

2.10.5.1 Texts

Issues worthy of careful consideration involve such matters as: which topics or themes to include, what kind of texts to select (narrative, descriptive) as well as the type and length of text units (dialogues, monologues). The method of delivery (video or audio), variety of English (academic, colloquial) and whether to include texts that are scripted or unscripted, also need to be determined (Rost, 1990:185; Weir, 1993:102). However, according to Shohamy and Inbar (1991:36), the fundamental consideration in the selection of texts is whether they are reflective of the purpose of the test.

Shohamy and Inbar (1991:36) recommend that tests of listening comprehension include texts with varying degrees of listenability. Although results from their study revealed that the more listenable texts were found to be easier, other aspects such as prior knowledge of the subject would affect the degree of difficulty as well. The difficulty of a text can also be affected by factors such as the number of infrequent words, hedging, vague words or very technical terminology (Bejar et al., 2000:13). Bejar et al. (2000:21) disagree with Shohamy and Inbar's (1991:36) theory that the type of text determines the degree of difficulty. They feel that pragmatic features, such as register, determine the level of difficulty.

Tasks should be based on topics and texts which are fair to all test-takers and they should be directed towards linguistic rather than general knowledge. Buck (2001:123), therefore, suggests selecting texts and designing tasks that depend on knowledge that either everyone has or no-one has. The tasks should thus be based on information supplied in the test.

Decisions on the length of a text should involve achieving a balance between long enough for test reliability but not so long as to overload the candidates memory.

2.10.5.2 Use of video

Using video to deliver listening texts has become popular since improved technology has made this a practical option. The combination of both aural and visual input affords listeners the opportunity to process information on two levels. As in many real life situations, listeners are able to see the speaker, so the use of video could be considered more authentic in its reflection of a TLU domain, particularly if it is applied to a lecture situation in a university context (Wagner, 2007:67).

There seems to be some agreement among researchers that allowing candidates to see the speaker during a listening test can aid them in their comprehension of the aural input (Bejar et al., 2000:12; Buck, 2001:123; Wagner, 2002:8) because the listener is given visual clues regarding the context and roles of the speaker. However, Read (2002:4) believes that even though the use of visual media has become very popular in academic listening tests, there is still a place for tasks which involve listening alone without seeing the speaker. Wagner (2007:67), on the other hand, is of the opinion that if assessment of listening ability takes place within a communicative framework as suggested by Bachman (1990) and Bachman and Palmer (1996), then the visual or non-verbal aspects are important features of spoken communication.

Chapelle and Douglas (2006:13) debate the advantages and disadvantages of video input by maintaining that, on the one hand, it is good for authenticity but, on the other, it could be a threat to the assessment of listening as the primary objective. Chapelle (2001:125) reports on a study conducted to compare the effects of different types of input on an academic listening test. Two different groups of students, with similar language ability levels, were tested. The input was in the form of a video (with audio) for one of the groups, while the same text was delivered to the other group using audio alone. A comparison of the test scores showed that the audio delivery was found to be more difficult. Therefore, if a construct theory of a listening comprehension test predicts that test tasks with visual input makes listening easier than those without seeing the speaker, then a test should include tasks using both types of input (Chapelle, 2001:127).

2.10.5.3 Tasks

Tasks should be based on an adequate comprehension of the text. In order to complete these tasks, the abilities, skills and knowledge specified in the construct are necessary. Only then

can inferences be made about the levels of proficiency of a candidate, in other words, 'how well they have mastered the construct' (Buck, 2001:117). Test tasks should serve a defined purpose for listening and it is often helpful to candidates if they are told what to listen for. It is also essential to ensure that tasks cannot be successfully completed without an understanding of the text (Buck, 2001:127). All tasks have strengths and weaknesses but by using a wide enough range of task types, the test is more likely to be balanced and fair in its assessment (Buck, 2001:153). If test tasks are to reflect real life as closely as possible, then the focus of the questions should be on the more global type of processing skills. Many listening test tasks fail to engage the test-taker because they do not actively require the listener to incorporate their language ability with background knowledge (Wagner, 2002:2).

Test developers also have to decide what response type and form they want to include in the test, as well as whether any restrictions should be placed on a response. For example, responses can either be selected, as in multiple-choice formats, or constructed, as in short answers or gap-fill. Some tests require test-takers to formulate their own answers; some are 'forced-choice items' (Davies et al., 1999:32,64), while others require both types of response (Bachman, 1990:157).

Task validity has a direct link with content and construct validity and, according to Fulcher (1999:227), it is essential that items are not only relevant but also representative of the TLU domain. Thus, the more detailed the domain specifications are, with regard to relevance and representativeness, the better the construct validity of the test is likely to be. The degree of difficulty of items or tasks, the quality of the instructions and the accuracy of the scoring are all factors which affect construct validity. Inappropriate difficulty or inaccuracies serve to create construct irrelevant variance and thus threaten validity.

Some of the criteria which Rost (1990:185) maintains are important for task validity are as follows:

- Tasks should be similar to those experienced in the TLU domain.
- Test-takers should be able to use their background knowledge of the context to respond to the tasks.
- Tasks should focus on meaning rather than form.
- Test-takers should be allowed more than one correct answer on relevant tasks (Rost, 1990:185).

The literature has shown that there is no such phenomenon as a completely valid language test; however, it has been my intention in this study to try and remove as many of the variables which could negatively affect validity as possible. The required response types and task validity criteria for the listening test associated with this study will also be discussed in the next chapter.

2.10.6 Conclusion

In this chapter some important premises were explored as a result of my review of current and past scholarship. The literature pertaining to general language testing, specific purpose testing, computer-based testing and academic literacy was essential for background information and contextualization of my study. However, the material on listening testing and specifically academic listening assessment, exerted the most influence on the empirical aspect of my research.

My inquiry into language testing in general and listening testing in particular has emphasised the complexity of accurately assessing linguistic ability. It seems that to isolate listening ability from other language skills would be an impossibility as they are all integrated in general language use. Thus, even if one sets out to test listening proficiency, reading and writing would of necessity form part of the assessment.

An explanation of the specific tasks selected for ALT along with an explanation of their relevance to the assessment will be included in the test description in the next chapter. Chapter 3 will also return to some of the theories and approaches of the researchers mentioned in this chapter since much of the test content was based on their findings and insights.

CHAPTER 3

RESEARCH DESIGN AND METHODOLOGY

3.1 INTRODUCTION

Chapter 2 explored the scholarship pertaining to the underlying theories and approaches to language testing in general and more specifically academic listening testing. This research constituted the basis for the design of ALT, the listening test developed as the assessment instrument for this study. According to Mouton (1996:107), 'a research design … follows logically from the research problem' and consequently, the hypothesis and research questions mentioned in Chapter 1 will be revisited in this chapter. A description of ALT as well as the rationale behind the theories and approaches on which it was grounded, will also be included in this chapter. Details of the participants in the study as well as an explanation of the data collection process, both qualitative and quantitative, will be described in some depth. Finally, this chapter will address the types of analyses used to measure the reliability and validity of ALT as well as its correlation with TALL.

3.2 RESEARCH QUESTIONS

The aim of my research was to design an academic listening test which could be used to quantitatively assess the academic listening skills of a selection of first year university students. The results of ALT were then compared with those of TALL to establish the feasibility of including a listening component as an added dimension to the assessment of reading and writing abilities as used currently. A number of qualitative and quantitative research questions, as outlined in 1.6, emerged from this procedure, the answers to which served to guide and direct my study.

The qualitative research questions formed the basis of a questionnaire circulated during a pilot project which will be described in 3.11. These questions concerned the construct, content and face validity of ALT by inviting comment on such issues as the relevancy, representativeness, degree of difficulty and quality of media clips included in the test. The quantitative research questions stemmed from an investigation into the levels of reliability and validity present in ALT. Issues of reliability such as the degree of internal and external consistency as well as evidence of construct and concurrent validity are addressed in these research questions.

3.3 HYPOTHESIS – A BRIEF DEFINITION AND RATIONALE

A primary and secondary hypothesis could be formulated once the above research questions had been addressed:

ALT is a valid and reliable test of academic listening skills.

An academic listening test would be a useful added dimension to TALL, currently implemented at the University of Stellenbosch.

As has been mentioned in the preceding chapters, the results of TALL form the current assessment of academic literacy abilities of students at the University of Stellenbosch. This paper-based test focuses on reading and writing tasks and it is compulsory for all incoming first year students to take the test. The results are used to 'stream' students into programmes which assist them in acquiring the various skills deemed necessary for their academic success.

TALL results take the form of a coding system which ranges from 1 – 5. About a third of the students tested fall into categories 1 and 2 which constitute the 'high risk' students. Categories 4 and 5 are those of 'low to no risk' with the 'borderline' students being those in category 3. Students who fall within category 3 are placed in either the top or the bottom group and there is a perceived need for more refined 'screening' so as to improve the accuracy of the placement of these students. It is, therefore, this group who will be of particular interest to this study.

3.4 THE ASSESSMENT INSTRUMENT

As was mentioned in the previous chapter, test design begins by making decisions concerning the content of a test; in other words, the test construct (McNamara, 2000:25). The purpose of the test should be stated as fully as possible since it is this *purpose* that will influence the content of the test (Weir, 1993:19). The theoretical framework will also to a large extent determine what will be included in a test. Knowing the 'domain' or set of tasks required by the criterion setting is essential for establishing the content. These may consist of practical, real-world tasks or be more abstract in construct. The required setting will then dictate the relevance and authenticity and as most researchers agree, a well designed test will always be content relevant (McNamara, 2000:25). Bachman and Palmer (1996:178-9) maintain that it is not enough merely to select tasks *relevant* to the construct domain; they must also *represent* the target setting. Therefore, the content aspect of construct validity has to be addressed by considering both the relevance of the content as well as how representative the tasks are.

Moreover, the way the test is structured as well as how the test-takers interact with it, must be fair to all candidates in achieving a scorable performance (McNamara, 2000:25). When developing language tests it is important to consider the cultural background, mother tongue, gender, age and background knowledge as some of the characteristics which could cause the test to be biased for, or against, various test-takers (Bachman, 1990:291; Shohamy & Inbar, 1991). Since ALT was designed to test the listening competency of first year students at the University of Stellenbosch, the speakers included in the clips are either English or Afrikaans mother tongue speakers as this represents the majority of lecturers at this institution.

Much has been written on how a test method may influence the performance of a test-taker. According to Bachman and Palmer (1996:46), 'the characteristics of the tasks used are always likely to affect test scores to some degree'. Bachman (1990:9) suggests that the biggest challenge facing language testers is to make sure that the test methods will reflect performance that is characteristic of ability in a non-test situation. In addition to the abilities that the test designer wishes to measure, the methods used to test these abilities also have a significant impact on test performance (Bachman, 1990:156; Alderson et al., 1995:44). This is known as the 'method effect' and test developers generally strive to reduce its influence. Douglas (2000:41) states that 'changing any of the situational characteristics of the input has the potential to change the performance they are meant to elicit and consequently, the interpretations we might make of the performance'.

According to Alderson et al. (1995:44), not enough research has been done on the effect of various testing methods, so using more than one method is often advised. According to Bachman (1990:156-7), the factors given below should be included in method effect considerations; these are discussed within the framework of my study.

- *The test environment which comprises the equipment used and the conditions under which the test is taken.*

  Of particular relevance to this study, is the effect that the computer might have as the mode of test delivery (Bachman & Palmer, 1996:141). Douglas (2000:277) emphasises the need for caution in controlling the test method effects which are associated with the technological resources available to test designers. The recordings need to be of good quality and the layout of the test clear. Since reading on screen is known to be more difficult than on paper, the font size and spacing are important considerations. The fact that the audio and visual input is delivered through headphones in ALT improves the sound quality and reduces background noise (Buck, 2001:118).

Chapelle (2001:123) maintains that the extent to which the computerised format affects test-takers' performance can only be addressed when the same constructs are measured using computer and non-computer delivery. Chapelle (2001:121) reports on a study where the computer-based TOEFL and the paper-and-pencil version of the same test were compared. The hypothesis was that there would be a high correlation since both tests measured the same abilities. This proved to be correct, resulting in an agreement among the study's researchers that the mode of delivery does not have a significant impact on the inferences deduced from the scores.

- *The test rubric which involves the organization of the test, the time allocation as well as the test instructions*.

  These should be based on clearly defined specifications and explicit criteria (Bachman, 1990:118). In the case of a computerised test, such as the one used in this study, additional instructions on how to navigate through the test form a very important part of the general test instructions. The rubric for the test associated with this study will be described in detail in the test description (3.10) later on in this chapter.

- *The input, for example, audio or visual or both; the nature of the language used and the degree of accuracy with which the problem or task is presented.*

  The inclusion of some form of multimedia is a prerequisite in any test of listening ability. However, before deciding to include visual input in the form of a video clip, I consulted a number of sources which yielded some conflicting findings. For example, Wagner (2007:67) suggests that visual aspects are significant features of oral communication and video and, therefore, lends authenticity to a task. Buck (2001:172), on the other hand, is not convinced of the importance of visual information and feels it would be better to place the emphasis on understanding the aural rather than the visual input. Bejar et al. (2000:28) have mixed opinions on the matter since they agree that while the use of video provides face validity and authenticity; it can also be a potential source of distraction. Further discussions on the input used in ALT will be included in the description of the test (see section 3.10).

- *The expected response, for example, selected or constructed, the use of a linguistic or non-linguistic form of the response, or both, as well as the restrictions that are placed on the response.*

  Blackboard, the learning management system used at the University of Stellenbosch, mentioned in the previous chapter (see section 2.5), includes a quiz template which

allows a broad spectrum of machine scorable formats for tasks. These testing mechanisms are built into the management system and offer test developers a wide range of features which they can include in their assessments (Godwin-Jones, 2001:10). These may include true/false, jumbled sentences, matching pairs, multiple-choice, gap-fill or short answers which use response options such as radio buttons and drop-down menus, amongst others.

Apart from providing scoring objectivity and thus a high measure of internal consistency (Bejar et al., 2000:178), the advantage of using a multiple-choice format in a test is that a wide range of abilities can be assessed in a short time and it makes the test easily scorable (Lado, 1961:34). These features together with considerations for facilitating computer usage were behind the decision to include multiple-choice, jumbled sentences and gap-fill type responses in the listening test specifications for ALT (see section 3.10).

There are, of course, those who believe that language cannot be tested by this method (Van Dyk & Weideman, 2004:2), perhaps because multiple-choice has been accused of enabling a test-taker to guess the correct answer and failing to resemble normal language use. However, high correlations have sometimes been recorded between multiple-choice tests and assessments of productive skills. This provides evidence that there is a fundamental ability which different kinds of language use have in common (Davies et al., 1999:125). Effective score interpretation in a multiple-choice test often depends entirely on the task type and the distractors used. Bejar et al. (2000:24) believe that distractors in a multiple-choice format should differ only slightly from the correct answer. Moreover, according to Alderson et al. (1995:47), candidates should not be able to answer the questions by using only their background knowledge. According to Lado (1961:34), a properly designed multiple-choice test can be an efficient measure of ability.

In a computerised test, the designer designates the acceptable response options. Spelling has to be correct for computer word recognition in electronic marking unless the test developer can anticipate all the possible misspellings and include them in a list of possible answers. A more practical solution is to provide a list of words to choose from (Chapelle & Douglas, 2006:70) this advice was used in ALT.

All these aspects had to be carefully considered and the specific question formats and responses required for each task will be addressed in detail in the description of ALT in 3.10.

3.5 TEST PURPOSE

As I have already indicated, the primary purpose of ALT was to quantitatively assess the academic listening skills of a selection of entry-level university students. ALT attempts to measure listening competency in understanding both explicit and implicit information (Bejar et al., 2000:5). The various texts, tasks and questions were designed to measure a wide range of listening skills and these will be discussed in detail later on in the chapter.

The secondary purpose was to determine whether ALT could make a contribution as an added dimension to TALL, another assessment of academic literacy. The combined aim of both these tests is to 'place' or 'stream' students into support programmes to assist them in their studies.

3.6 IDENTIFYING THE TARGET LANGUAGE USE (TLU) DOMAIN

Rost's (1990:180) statement that a test of language ability which is limited to a particular setting is more relevant to a candidate than a general approach was mentioned in the previous chapter with relevance to the context of listening testing. Within this context, there are two aspects which are essential for defining the construct. The first is the specific knowledge or abilities that candidates should possess to be successful in the TLU domain and the second, the sort of tasks they should be able to perform (Buck, 2001:102; Bejar et al., 2000:5).

Identification of the TLU situation, even if it is confined to an academic setting, will vary according to the different disciplines or faculties so test designers have to make do with an approximation (Buck, 2001:106). ALT focuses on listening as it occurs in a university context and the tasks attempt to reflect those that would be expected of a first year university student in real life. The students who took part in ALT were from different cultural and language backgrounds but all were enrolled as first year Bachelor of Science students and would thus be expected to have similar academic literacy abilities.

3.7 TEST CONSTRUCT – ABILITIES RELEVANT TO THE TLU DOMAIN

Listening comprehension as a process, because it cannot be directly observed, has to be assessed by means of the more measurable skills that constitute the academic TLU domain (Brindley, 1998:172; Weir, 1993:98; Rost, 1990:33). Of the many listening abilities listed in the literature, there are some that stand out as being particularly important in a university setting. These will be mentioned briefly below and discussed in greater detail in section 3.10 where ALT is described in depth.

Decisions on tasks to be included in ALT were based on the definitions of listening and taxonomies of a number of researchers. These include: Buck (2001), Weir (1993), Wagner (2002) and Jordan (1997), as well as the compilers of TOEFL (Bejar et al., 2000) and the Cambridge Proficiency in English (CPE) listening test (Developing general listening skills [s.a.]). Details regarding the various theories and approaches were included in Chapter 2.

Several researchers (Bejar et al., 2000:10, Wagner, 2002:11; Brindley, 1998:172; Shohamy & Inbar, 1991:29) maintain that a listener in an academic context needs to listen for specific details and facts in a text. The importance of a listener's ability to process extensive pieces of information and identify the main theme as well as the supporting ideas found in a text are also emphasised in the literature (Buck, 2001:43; Flowerdew, 1994:11; Jordan, 1997:179). Since listening to lectures comprises such a large part of university study, this capacity seems essential for effective listening in an academic setting.

Researchers have also identified an ability to *infer* meaning from a spoken utterance as an important ability at tertiary level (Buck, 2001:60; Brindley, 1998:172; Lynch, 1998:6; Rost, 2002:64; Shohamy & Inbar, 1991:29). In order to understand spoken discourse, the listening process has to include inference which is based on implication rather than just an understanding of the literal meaning of the words used in the discourse. There are several types of inference which range from the lower-level type such as supplementing information, to higher-level reasoning which involves background knowledge. Making inferences about a speaker's attitude or opinion might not seem especially relevant in an academic TLU domain but a speaker's attitude can nonetheless be indicative of the importance of a piece of information (Wagner, 2002:11; Weir, 1993:99) as can prosodic features such as stress and intonation.

The ability to deduce the meaning of a word from the context has also been described as an important skill by researchers (Buck, 2001:22; Weir, 1993:98; Jordan, 1997:180; Weideman, 2003a:xi). This ability, however, is very difficult to assess accurately since it is almost impossible to gauge whether meaning was inferred through semantic knowledge or based solely on the context. Nonetheless, since this seems a relevant academic listening skill and is often cited as a significant component of listening ability (Buck, 2001:60; Weir, 1993:98), it has been included in ALT.

Since much of the literature reflects the view that listening is a two stage process (Buck, 2001:51; Chaudron & Richards, 1986:113; Rost, 1990:33; Shohamy & Inbar, 1991:29; Weir, 1993:98), the abilities mentioned above have been loosely divided into these two stages. Bottom-up processing involves the more 'local' skills such as identification of details and extracting facts, whereas top-down processing requires interpreting the more implicit

information such as inferencing or listening for gist. However, as has been mentioned earlier, the boundaries between the two processes are blurred and often occur simultaneously in a so-called parallel process (Rubin, 1994:211). This makes it very difficult to attribute task responses to any one particular skill or construct (Brindley, 1998:173; Buck, 2001:106) but since there is a need for some construct definition or theory to serve as a framework, the model represented below has been used for the operationalisation of ALT.

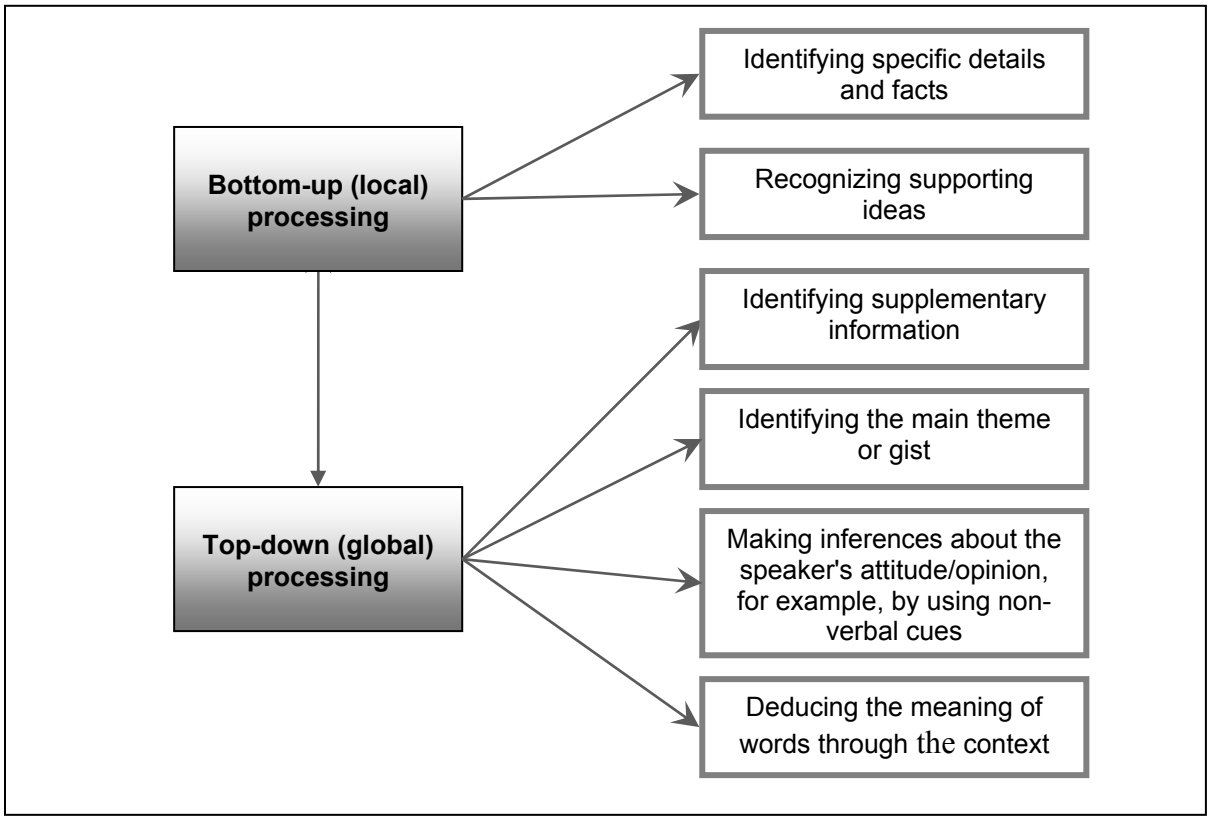

FIGURE 3.1: OPERATIONALISATION OF A MODEL OF ACADEMIC LISTENING ABILITIES (ADAPTED FROM WAGNER, 2002:12)

3.8 TEST SPECIFICATIONS

Test specifications are the result of the design process and determine both the method and the content of the test. This 'recipe' for the test includes features such as the structure and length of each section, the sort of response that is required, the instructions to the candidates as well as the scoring system (McNamara, 2000:31). The specifications for ALT will be addressed later on in this chapter along with the description of the test itself.

Test specifications often change as a result of pre-testing and feedback from colleagues and students. Versions of test specifications serve as a useful record of evidence gathered to address validity issues (Davidson & Lynch, 2002:8). Bachman and Palmer (1996) use the term 'blueprint' for a test specification and divide it into two parts. The first part involves the

'structure of the test', for example, how many sections, items or tasks are presented as well as how they are organized. The second part has to do with 'test task specifications' and contains, for example, such details as test purpose, construct definition, setting, instructions and scoring method.

ALT was pre-tested by colleagues as well as students so as to gain qualitative feedback in order to determine face, content and construct validity (see 2.2.2.1 - 2.2.2.3). The test specifications were not significantly changed as a result of the pilot testing but useful opinions on the relevance and representativeness of the test items were given in response to a questionnaire (included as Appendix A). A description of the pilot testing, both administrations of ALT as well as details regarding participants and procedures will be given towards the end of this chapter and the results of this research will be addressed in the next chapter.

3.9 OPERATIONALISATION OF THE TEST CONSTRUCT

In order to assess relevant listening skills as accurately as possible, careful consideration was required in the selection of the material used in the tasks. Appropriacy of context, accent and rate of speech, length of text as well as task format were some of the factors that needed to be taken into account when compiling the test tasks (Weir, 1993:102).

The rationale behind the decision to use Blackboard as the mode of delivery for ALT was based on the ease of scoring and the computer's ability to instantly calculate statistical data. The flexibility and convenience of students being able to take the test in their own time without the need for test venues or invigilators was another attractive feature of this mode of test delivery. However, because of the inclusion of a lengthy video clip which could take some time to download and in order to manage any technical hitches, it was decided to administer ALT under controlled conditions. The procedures used for the administration of ALT will be described in section 3.12.2.

Within the Blackboard quiz format, a way of limiting the number of times candidates could listen to the audio and video files initially posed a problem. The only solution was to remove the media player controls and since it would be impractical for the files to 'play' when the screen opened, it meant that the media files had to be linked to 'Save and View Next' buttons in order to activate them. This, in turn, prevented candidates from being able to return to previous pages to review instructions or check their work. However, in the interests of making the tasks as authentic as possible (in reality, a student would only listen to a lecture once), two of the three audio files as well as the video could only be accessed once and candidates could not return to a previous screen.

3.10 TEST DESCRIPTION

The opening screen of ALT (a transcript of which is included as Appendix B), informs the students that they have an hour and a half in which to complete the test and that headphones will be required as well as material for note-taking. A brief description of the test is then given in which candidates are told that the assessment consists of four sections. They are also informed that instructions are provided at the beginning of each task regarding the type of response that is expected of them and what the procedure is for test navigation.

Once a test-taker has pressed the 'Begin Assessment' button, he/she is given the opportunity to check their computer's media playback configuration and volume settings. The next screen is in accordance with the requirements of the Ethics Committee of the University of Stellenbosch. ALT was required to include a disclaimer which provides prospective candidates with information on the rationale behind the test and stipulates that no student will be disadvantaged by taking the test. In addition, candidates are made aware of the fact that data will be referred to collectively and not individually, so no names, student numbers or other personal information will be included in the thesis. There is also an opportunity for every candidate to give their consent to the results being used for research purposes.

The four tasks are placed in an 'easier-to-more-difficult' order and candidates are advised of the listening purpose for each task. In addition to the instructions given at the beginning of each task, where necessary, extra information is also given before some of the questions. Using an adaptation of Popham's 'five-component test specification mode' (Davidson & Lynch, 2002:14), details of these four tasks are given below with their criteria, format and response type.

3.10.1 Task 1 – Instructions

(a transcript of the audio clip is included as Appendix C)

3.10.1.1 General description and purpose

This task measures three main abilities:

1. To recognise and remember specific instructions which include warnings, suggestions, recommendations and advice.

2. To recognise the function of non-verbal cues such as stress and intonation, as indications of emphasis.

3. To deduce the meaning of words from the context.

3.10.1.2 Prompt attributes

Candidates listen to an audio clip of about 90 seconds, of a lecturer giving his class various instructions. They are advised to take notes as the clip can only be accessed once and told that the notes do not form part of the assessment. They are also informed that they will be expected to respond to multiple-choice questions after listening to the recording.

3.10.1.3 Response attributes

Selected response: Test-takers are required to select one of four multiple-choice options in response to Question 1 and Questions 3 to 7. Correct responses are worth 2 marks each. Question 2 requires candidates to place four steps necessary for accessing course information into the correct order. These are presented by asking them to select each step from a drop-down menu. Two marks are awarded for each step placed in the correct order and partial credit is given even if all four responses are not correct.

3.10.1.4 Sample item

This task reflects the TLU domain in that all students are given instructions by their tutors or lecturers during the course of their studies and it is important that they are able to focus on specific details and thus retain them.

3.10.1.5 Specification supplement

The following criteria were taken into consideration in selecting the material for the audio clip (Weir, 1993; Hughes, 2003):

1. Selecting a text that was not biased in terms of subject matter or culture.
2. Selecting a text that was of the appropriate length. According to Weir (1993:109), approximately a minute is the optimal length.
3. Sufficient information on which to base the questions had to be included in the clip.
4. The sound quality in an 'authentic' recording, as is the case in this task, is never quite as good as when it is 'staged', so the audibility is adequate rather than perfect.
5. The lecturer featured in this task has a standard South African English accent that is not too pronounced and he enunciates his words clearly.

6. The speaker's speech rate is normal for the context and there are enough pauses to facilitate note-taking.

3.10.2 Task 2 – Lecture extract

(a transcript of the video clip is included as Appendix D)

3.10.2.1 General description and purpose

This task measures the following abilities:

1. To identify the main theme of a lecture.
2. To recognise and recall important details and specific information presented in the text.
3. To recognise and recall the stated opinion of the lecturer.
4. To distinguish between the most important information and the supporting detail.
5. To concentrate on, and to process, a long piece of text.

3.10.2.2 Prompt attributes

Candidates listen and watch a 13 minute video clip taken from a first year Psychology lecture. They are once again advised to take notes as the video can only be viewed once. They are reminded that their notes are to help them remember important information and will not form part of the assessment. They are also informed that they will be expected to respond to multiple-choice questions after listening to the recording.

3.10.2.3 Response attributes

Selected response: Test-takers are required to select one of four multiple-choice options in response to 12 questions. Correct responses are worth 2 marks each.

3.10.2.4 Sample item

This task is extremely relevant to the TLU domain in that listening to lectures and taking notes, so as to recall information, constitutes a large part of any academic course (Chaudron et al., 1994:76). The ability to extract the main ideas and important details from a lecture is an essential part of successful study at university (Jordan, 1997:181).

3.10.2.5 Specification supplement

I based my selection of material for the video on criteria recommended by experienced test researchers (Weir, 1993; Hughes, 2003; Jordan, 2006; Rost, 2002).

1. Selection of a text that was completely authentic and recorded in a real-life lecture hall so that it would simulate the TLU domain as closely as possible.

2. Selection of a text that was not biased in terms of subject matter or culture. The content of the lecture was selected on the basis that none of the candidates would be familiar with the topic. Any subjects which could have been taken at school, such as History or Science, were thus avoided.

3. The length of the lecture extract had to be at least ten minutes (Weir, 1993:109).

4. The organisational structure of the text had to have a clear outline with enough information in terms of main ideas and supporting details.

5. The sound and visual quality had to be good.

6. The lecturer's speech characteristics had to be carefully considered. The accent of the lecturer featured in the video as well as his clarity of articulation and lecturing style make him easily understood. In spite of the fact that Afrikaans is his mother-tongue, he speaks English without discernible first language influence.

7. The speaker's rate of speech is very important in a lecture situation. In this extract, the rate is normal for the context and there are enough pauses, hesitations and restarts to facilitate note-taking.

8. The degree of repetition of main ideas in the video is an important factor. This provides 'padding' which allows the listener to make a note of important points without being left behind. The repetition of new terms or concepts, present in this particular clip, helps to familiarise students with new and sometimes technical subject matter.

3.10.3 Task 3 – Discussion

(a transcript of the audio clip is included as Appendix E)

3.10.3.1 General description and purpose

This task measures the following abilities:

1. To deduce information based on background knowledge of the topic and to fill in any gaps through awareness of the context.

2. To make inferences about the speakers' attitudes or opinions.

3. To distinguish between the most important information and the supporting detail.

4. To identify the main theme as well as the supporting arguments presented in the discussion.

5. To identify reformulation as a means of agreement in a dialogue.

6. To listen for attitudes and opinions expressed both explicitly and implicitly.

3.10.3.2 Prompt attributes

Candidates listen to a 5 minute audio clip of a discussion on euthanasia between two law students. The candidates are advised to listen closely to the opinions and supporting arguments given by the two speakers. It is once again suggested that notes would be a useful way of remembering what was said. The test-takers are also informed that the audio input will only play once and that they will be expected to respond to a multiple-choice exercise after listening to the recording.

3.10.3.3 Response attributes

Selected response: Test-takers are required to select one of four multiple-choice options in response to Questions 1 and 2 and 4 to 9. Correct responses are worth 2 marks each. In Question 3, candidates are given four statements and are asked to decide which speaker agrees with the specific statement or whether they both support it. Each correct response to the four statements is worth 2 marks.

3.10.3.4 Sample item

Since students are encouraged to discuss and critically analyse aspects of their coursework, this task seemed relevant to the TLU domain. The ability to fill in missing information and predict a speaker's purpose is an important critical thinking skill which successful students should possess (Buck, 2001).

3.10.3.5 Specification supplement

The text comprises a typically informal discussion between two fellow students. In selecting the material for the discussion, the following criteria had to be considered (Weir, 1993; Hughes, 2003; Jordan, 1997; Rost, 2002).

1. To make the discussion as authentic as possible. In spite of the fact that the recording was 'staged', the students decided on their own topic and talked spontaneously without 'reading' any of the material.

2. To select a text that was not biased in terms of subject matter or culture. The content of the discussion was limited to the legal aspects of the topic, which affect everyone equally. The religious or moral aspects of euthanasia were purposely avoided.

3. The speakers had to be of different sexes so that it would be easier to identify who is speaking at any particular time.

4. The text had to include the exchange of opinions on a particular topic where agreement and disagreement is expressed. It is intentional that agreement or disagreement in the text is only implied, sometimes quite subtly, rather than explicitly stated.

5. The text had to have a main theme or argument as well as supporting arguments.

6. The speakers' accents and clarity of speech had to be considered. The male speaker has Afrikaans as his first language and the female speaker English, but neither speaker has a very pronounced accent and both are easily understood.

7. The quality of the recording had to be considered. In some parts, the speech is slightly indistinct but it is generally reflective of a conversation one might hear in 'real life'.

8. The speakers' rate of speech is very important when listening to a discussion. In this extract, the rate is fairly slow, as the speakers have to think about the next point they want to make.

3.10.4 Task 4 – Tutorial extract

(a transcript of the audio clip is included as Appendix C)

3.10.4.1 General description and purpose

This task measures the following abilities:

1. To use an understanding of the text to fill in the content words omitted from the summary. According to Buck (2001:71), if test-takers are asked 'to fill in blanks on a summary of the passage … [it] forces them to process the meaning'.

2. To concentrate on, and process, a long piece of text.

3. To listen for details and specific information.

4. To identify reformulation or paraphrasing.

3.10.4.2 Prompt and response attributes

Candidates begin the task by being given a piece of text to read where 16 content words have been omitted. The candidates are told that this preliminary reading is to give them an idea of the central theme and structure of the text. They are then asked to re-read the text while listening to a 5 minute audio clip of an extract from a tutorial on *Foreign Direct Investment* given by a senior lecturer in Business Science. The fact that the written text is a summary and not a transcription of the audio-clip is brought to the attention of the test-takers. Candidates are instructed to make notes of the missing words but are not yet able to start typing them in. The next screen presents candidates with the 'answerable' text and the audio clip is heard again. Candidates are advised to check their notes as they listen to the clip a second time and are told that, since it is not a speed test, they do not have to try and type in the words while listening to the clip. A list of the 16 missing words, along with a few distractors, is provided alongside the text. This is to ensure that responses are spelled correctly, since the computer would not be able to recognize incorrectly spelled answers. In some of the gaps, limited response options have been allowed. Each correct response counts 2 marks.

3.10.4.3 Sample item

This task reflects the TLU domain, in that most students attend tutorials where what they hear and what they read is not exactly the same text. The task could also reflect a situation where a student consults a lecturer and has a concept explained to him/her.

3.10.4.4 Specification supplement

The following criteria had to be considered in the selection of the material for the audio clip (Weir, 1993; Hughes, 2003).

1. Selection of an academic piece of text that was not biased in terms of subject matter or culture.

2. Selection of a text with enough subject-specific vocabulary to ensure that the gaps could not be filled by using general knowledge. If gaps can be filled without listening closely to the clip then the construct validity of the exercise is threatened.

3. Selection of a text that was long enough to provide a challenging gap-fill exercise.

4. The sound quality had to be good.

5. The lecturer featured in this task has Afrikaans as a first language and some mother-tongue influence can be detected in her accent. However, she pronounces her words clearly and is easily understood.

6. The lecturer's speech rate is fairly rapid but is representative of the speed at which some tutors or lecturers speak.

3.11 PILOT TESTING

A pilot test was conducted prior to the main study to qualitatively assess ALT. Since ALT is computerized, it was important to identify possible technological problems and to determine the 'user-friendliness' of the instrument. Peer evaluation as well as feedback from a sample of first year students was also a necessary requirement to determine the level of construct, content and face validity of the test. Each member of the beta group was requested to work through ALT and then complete a questionnaire which is included as Appendix A.

3.11.1 Participants

The participants involved in the pilot test comprised a group of nine first year Health Science students as well as three lecturers.

The group of nine student volunteers consisted of eight females and one male. English is the mother-tongue of three of the students, Afrikaans the first language of four, one who professed to be completely bilingual and one whose first language was Xhosa. The students are aged between eighteen and nineteen and are from different cultural as well as linguistic backgrounds.

The lecturers are all experienced language teachers who teach academic literacy support classes to first year students. All three of the lecturers have had extensive experience in course design and coordination, particularly in the field of academic literacy. They are all familiar with TALL and are, therefore, well equipped to give professional feedback on the content of ALT.

3.11.2 Procedure

The beta group included in the pilot testing section of the project were all granted access to ALT for a week and could thus decide on a time and place which suited them to work through the test. Since ALT was available on Blackboard, both students and lecturers could gain access to the test by using their student numbers or user names in the case of the lecturers, and passwords. They were informed that there were download considerations if they sat ALT at home or on a laptop computer without a broadband connection. The Health Science students were all advised to take the test in the computer centre of their campus where headphones are available.

All the participants were made aware that their results were not important to the evaluation but that in order to answer the questions on the questionnaire supplied to them, they would all have to complete ALT. The staff identified no technological difficulties but the students had a few problems as a result of a self-confessed failure to read and follow instructions.

3.11.3 Questionnaire

The questionnaire used in the research (see Appendix A) was designed primarily to assess whether ALT had content and face validity. For this reason, many of the questions were based on the representativeness and relevancy of the tasks. However, other issues pertaining to construct validity, such as task appropriacy, clarity of instructions, sound and visual quality and the effect of the computerised format of the test were also addressed.

The questionnaire supplied to the lecturers contained an additional page of questions which related to their opinion of ALT as an effective indicator of academic literacy. At the end of the main questionnaire (for students and staff) an opportunity was given to make any additional comments that were not covered in the questionnaire.

The response to the questionnaire was very satisfactory with many well-founded opinions given. These will be discussed in more detail in the next chapter together with the test findings.

3.12 TEST ADMINISTRATION

As I have mentioned previously, concern about technological difficulties prompted my decision to administer ALT under controlled conditions. The computer centre of the Faculty of Science where the test-takers are registered is not equipped with headphones and since these are an

essential accessory, a suitably equipped venue for the administration of the test was provided. The procedures for the administration of ALT will be discussed in 3.12.2.

3.12.1 Participants

A group of six hundred and twenty-seven Bachelor of Science first year students were asked to take ALT. Ninety seven students volunteered and completed both sittings of the test. These students were halfway through their first year when they took the test. They had all attended a semester of academic literacy intervention in either English or Afrikaans. The students in both the English and Afrikaans classes had written TALL at the beginning of the year. However, only the students in the English classes had also written TALL in April and June. The literature indicates that both the test and the criterion, in this case TALL, should be administered in approximately the same time frame for reasons of concurrent validity, as discussed in Chapter 2 (Hughes, 2003:27). Since ALT was administered in July and August, this was my rationale in deciding to use only the 'English class' (not necessarily English mother-tongue students) students' data when comparing the June 2008 results of TALL and ALT to check for criterion-related validity. These students comprised sixty of the ninety seven in the sample group. However, the other analyses are all based on the whole sample group comprising students from both the English and Afrikaans classes.

According to Mouton (1996:110), 'representativeness is the underlying epistemic criterion of a "valid", that is, unbiased sample'. Since the candidates included in the test participant group were volunteers, representative sampling was perforce, random rather than systematic. The criteria I used to check for representativeness as shown in the graphs (below) were: home language (Figure 3.2), gender (Figure 3.3) and academic literacy proficiency (Figure 3.4), which was based on the results of TALL written in January 2008.

I divided home languages into English, Afrikaans and other first languages (L1). The latter include African languages such as Xhosa, Zulu, Tswana, Pedi and Venda as well as European languages like French and German. As can be seen from graph 3.2, the students who speak a language other than English or Afrikaans at home are fairly well represented by the sample group. However, the ratio of English to Afrikaans students is noticeably higher in the sample group than in the group as a whole.

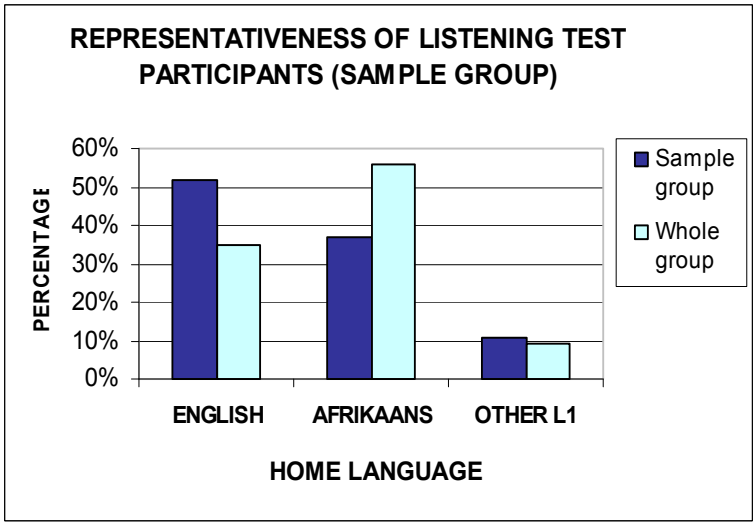

FIGURE 3.2: GRAPH SHOWING THE REPRESENTATIVENESS OF THE LISTENING TEST SAMPLE ACCORDING TO HOME LANGUAGE

The numbers of male and female students in the whole group are evenly divided as is shown in graph 3.3. In the sample group however, there are ten percent more females and ten per cent fewer males.

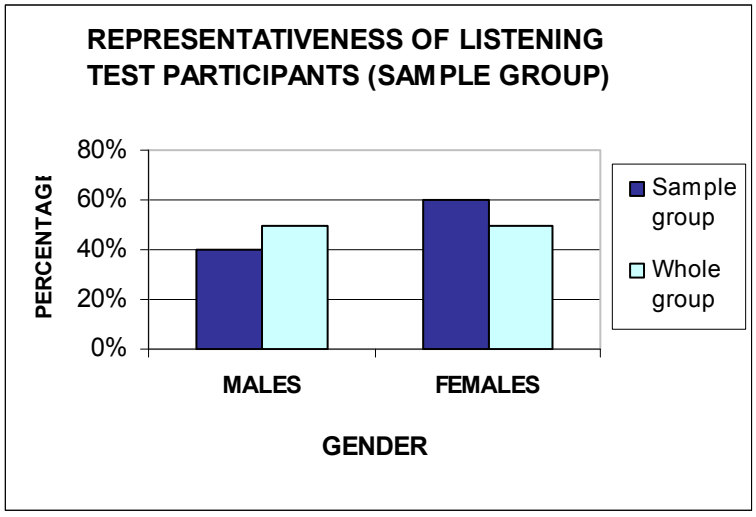

FIGURE 3.3: GRAPH SHOWING THE REPRESENTATIVENESS OF THE LISTENING TEST SAMPLE ACCORDING TO GENDER

For the purposes of this particular investigation, candidates who achieved TALL scores of seventy per cent and above were grouped in the 'high academic literacy level' bracket, TALL scores of between fifty and sixty nine were considered to be of a 'medium academic literacy level' and below fifty constituted the 'low academic literacy level' group. The graph presented below compares the sample and whole groups in terms of academic literacy levels.

Graph 3.4 shows a high level of representativeness between candidates in the medium bracket in both groups. In the top academic literacy category, the sample group had slightly fewer candidates than the whole group but slightly more at the lowest level of proficiency.

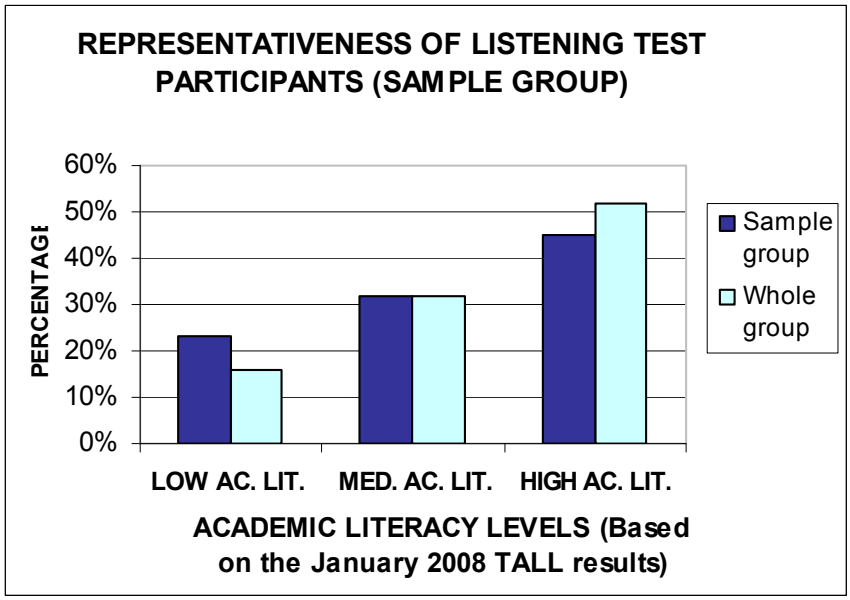

FIGURE 3.4: GRAPH SHOWING THE REPRESENTATIVENESS OF THE LISTENING TEST SAMPLE ACCORDING TO ACADEMIC LITERACY LEVELS

Incoming first year students all take a basic computer-skills course which, to some extent, lessens the method effect, mentioned on page 57, of the use of computers for testing purposes. ALT is also designed to be as user-friendly as possible and makes few demands, in terms of computer literacy, on the candidates. All the students would also have had English as a subject at school and all the candidates were of a similar age, so the question of bias based on these issues was not considered to be significant.

3.12.2 Procedure

All the *Scientific Communication Skills 172* lecturers were given the necessary background information and instructions for the administration of ALT. They were asked to pass this information on to their respective students and instructions were uploaded to the students' portal on Blackboard. ALT appeared on the candidates' module profile on Blackboard and they were given 'sign-up' options to select a time slot which suited them from a list of various dates and times spanning several days. This meant that they could access ALT only in their chosen time slot. The students were also informed that they would be required to take the same test twice with an interval of a month between the two sittings. As an incentive they

were promised five bonus marks in their *Scientific Communication Skills 172* module for completion of both test sessions.

The sessions were all supervised to ensure that the candidates did not experience any technical problems and were not disturbed or distracted while completing ALT; in other words, any threat to valid scores being obtained was kept to a minimum. The candidates were familiar with the method of accessing ALT as they had all done several Blackboard tests as part of their respective courses during their first six months at university. Candidates gained access to ALT by logging on using their student numbers and passwords and then selecting the test from their list of modules.

Each time slot was an hour and a half which the pre-testing had shown was more than adequate even for those who prefer to work more slowly. Most of the test-takers completed ALT in just over an hour and even those who took longer, finished within the time limit. Since the results of ALT were instantly available, candidates were given the option of seeing their score.

3.13 ANALYSES

The first analysis involved a qualitative investigation into the feedback from a questionnaire which, as I described above, was completed by a group of lecturers and students.

The quantitative data from these two ALT administrations were analysed using STATISTICA.

An internal reliability analysis was performed using Cronbach alpha to calculate the item-total correlation for each item in order to determine how each item related to the other items in the scale. The reliability analyses for each scale and overall assessment if those items were deleted, as well as the corrected item-total correlation for each of the remaining items, were then computed. A reliability study was then performed to examine the overall reliability of the test as a whole.

As part of the above analyses, an item analysis was done and from the results it was possible to determine which items performed poorly, and which items should be rejected in future versions of the test. In addition, the internal correlations of the different test sections were calculated to determine the degree of difference or similarity in the attributes they were testing.

Using the test-retest method of gauging consistency of measurement over a period of time, the results of the first test administration were compared with those of the second by using the

Spearman correlation. Since a correlation is only considered significant if its p value is smaller than 0.05, the p values of these correlations will also be given. The ICC (agreement) as well as ICC (consistency) was calculated to compare the results on the two tests.

In order to examine the criterion validity of ALT, the Spearman correlation system was also used to compare the results of both administrations of ALT with the June 2008 results of TALL. Finally, candidate performance in the different categories of scores on TALL were compared with the corresponding candidates' results on ALT so as to determine whether ALT would be helpful in screening the borderline candidates in code 3.

3.14 CONCLUSION

This chapter has described and discussed the operationalisation of the assessment instrument from its theoretical beginning to the procedures used for data collection and analysis. Both qualitative and quantitative data have yielded useful information regarding the reliability and validity of the test and these results will be presented and discussed in the next chapter.

CHAPTER 4

RESULTS: PRESENTATION AND DISCUSSION

4.1 INTRODUCTION

In Chapter 3 details of the design and operationalisation of ALT were given along with details of the sample groups that participated in the project and procedures for the administration of ALT and data collection. In this chapter, the qualitative results of the pilot testing as well as the quantitative outcomes of both ALT administrations are presented and discussed with reference to the research questions addressed in this study.

4.2 QUALITATIVE FINDINGS: FEEDBACK FROM THE QUESTIONNAIRE

In section 3.11.3 of the previous chapter it was pointed out that a questionnaire (see Appendix A) was distributed to a group of nine Health Science first year students and three lecturers from the Unit for Afrikaans and English all from the University of Stellenbosch, to gain qualitative feedback on ALT (see Appendix B). The comments and suggestions made by the lecturers and students contributed to an assessment of the construct, content and face validity of ALT. These remarks are reported according to the questions in the questionnaire in the paragraphs that follow.

Question 1: The tasks are designed to simulate actual tasks that a first year university student would be called upon to perform. To what extent do the 4 test tasks reflect the target language use situation, i.e. a university campus?

Summary of response: The lecturers were unanimous in their opinion that the tasks included in ALT reflected those that students would be called upon to perform in the TLU domain. The majority of the students were of the opinion that the first two tasks included in ALT reflected those that students would be required to perform at university. Student opinion was more divided for Tasks 3 and 4. Three of the nine students felt that Task 3 was 'totally relevant', two that it was 'moderately representative' and three that it was 'reflective in parts'. Task 4 was considered completely relevant by only one of the students with the remaining opinion divided between being 'moderately representative' and 'reflective in parts'.

Question 2: How interesting and relevant to university study did you find the test content?

Summary of response: The lecturers found the content of ALT mostly relevant to university study but remarked that because of the subject-specific nature of some of the content it would not always be interesting or relevant to all the test-takers. The student participants found the content of Tasks 2 and 3 both interesting and relevant to the TLU domain. However, Task 1 was considered only 'moderately interesting and relevant' and Task 4 'quite interesting but not necessarily relevant or visa versa' by the majority of the students.

Question 3: Is the level of difficulty appropriate for a first year university student?

Summary of response: All the lecturers agreed that the level of difficulty in Task 1 was appropriate for first year university students. In Task 2, one lecturer felt the level was appropriate, one that it was too difficult and one felt that some of the questions were too easy and some too difficult. The ability to infer meaning which constituted Task 3, was considered to be pitched at the correct level by two of the three lecturers, the third found the task too difficult. Similarly, in Task 4, two lecturers found the level of difficulty appropriate for first year level and one thought it too subject-specific and therefore too difficult. The students were all in agreement that the level of difficulty of the tasks was appropriate for a first year university student.

Question 4: In the multiple choice tasks, are the questions fair or are the distractors challenging but fair, too similar, confusing at times or very confusing?

Summary of response: The consensus among the lecturers was that they were 'challenging but fair'. Opinion amongst the students was divided almost exactly between those who thought they were 'challenging but fair' and those who found them 'confusing at times'.

Question 5: Is the order in which the tasks are placed in the test appropriate?

Summary of response: All the lecturers agreed with the order in which the tasks were placed in ALT. Six of the nine students thought that the order of the tasks in ALT was appropriate. One student felt that since the tasks had nothing to do with one another, the order was irrelevant. The other two agreed that the tasks should be arranged from easier-to-more-difficult but had their own opinions about which tasks were easier and which more difficult.

Question 6: Is the computerised format of the test easy to navigate, harder to use than a pencil-and-paper test or easier to use than a pencil-and-paper test?

Summary of response: There was general agreement among the lecturers that the computerised format of ALT was both easy to navigate and easier to use than a pencil-and-paper test. Most of the students found the computerised format of the test easy to navigate. One considered the format easier to use than a pencil-and-paper test and two thought it more difficult.

Question 7: Comment on the quality of the sound and visuals in the multimedia presented.

Summary of response: Two of the three lecturers and five of the students considered the quality of the sound and visuals 'good', the third lecturer and one of the students deemed it 'excellent' while three of the students thought it only 'adequate'.

Question 8: Comment on the length of the multimedia clips.

Summary of response: The lecturers all felt the length of the clips were appropriate for the specific tasks. All but one of the students thought the duration of the multimedia clips was too long. The remaining student thought the length appropriate for the relevant tasks.

Question 9: Are the instructions clear and unambiguous, quite clear, confusing at times or inadequate and confusing?

Summary of response: The lecturers all agreed that the instructions were clear and unambiguous. In Tasks 1 and 2, most of the students felt that the instructions were 'clear and unambiguous'. In Tasks 3 and 4, opinion ranged from 'clear and unambiguous' to 'confusing at times'.

Question 10: Did you find that being able to listen to the 'instructions' audio clip only once (Task 1) was adequate for the task, not enough to absorb all the information or completely insufficient?

Summary of response: All the lecturers thought that this was adequate for the task. Four of the nine students felt that only being able to listen to the clip once was not sufficient while two felt it was adequate. The other three did not give an opinion.

Question 11: Did you find that a single screening of the Psychology lecture (Task 2) was sufficient, necessary for 'real-life' simulation but had difficulty noting all the points or completely inadequate?

Summary of response: Two of the lecturers and six of the students thought it was 'sufficient' and one agreed with the remainder of the students that it was 'necessary for 'real-life' simulation but had difficulty noting all the points'.

Question 12: Did you find that being able to listen to the 'euthanasia discussion' audio clip only once (Task 3) was adequate for the task, not enough to absorb all the information or completely insufficient?

Summary of response: Two of the three lecturers and most of the students thought it was 'adequate for the task'.

Question 13: What listening skills do you think are being measured in the various tasks?

Summary of response: The table below is an illustration of their opinion.

TABLE 4.1: LECTURER AND STUDENT OPINION ON THE SPECIFIC LISTENING SKILLS BEING MEASURED IN THE TASKS

	Direct meaning comprehension skills	**Inferred meaning comprehension skills**	**Contributory factors such as stress and intonation**	**Listening and note-taking skills**
Task 1	Lecturer 1 & 2 Students 1, 2, 3, 4 & 5	Lecturer 2	Lecturer 2 Student 1	Lecturer 1, 2 & 3 Students 1, 6 & 7
Task 2	Lecturer 1, 2 & 3 Students 3 & 6	Lecturer 2 Students 1, 2, 5, 6 &	Lecturer 2 & 3	Lecturer 1, 2 & 3 Students 4, 6, 7 & 8
Task 3	Lecturer 1, 2 & 3 Students 4 & 8	Lecturer 1 & 2 Students 3, 6 & 7	Lecturer 1 & 2 Students 6 & 8	Lecturer 1, 2 & 3 Students 1, 2, 5, 6 9
Task 4	Lecturer 1, 2 & 3 Students 1, 6 & 7	Lecturer 1, 2 & 3 Students 4 & 8	Lecturer 2 & 3 Student 5	Lecturer 1, 2 & 3 Students 1, 2, 3, 6, 8 & 9

Question 14: Is the listening purpose e.g. main theme, specific detail, opinion, supporting arguments etc. made clear to candidates in the instructions?

Summary of response: Two of the three lecturers felt that the listening purpose in Tasks 1, 2 and 4 was made very clear to test-takers in the instructions. The other lecturer felt that the purpose was not clear enough in any of these tasks and all the lecturers felt that Task 3 should have been made clearer. Among the students, opinions varied as the majority felt that the purpose was 'very clear' in Task 1 and 2

but for Tasks 3 and 4, feedback was evenly distributed between 'very clear' and 'fairly clear'.

Question 15: Are there factors in the test which could advantage/disadvantage any particular candidate? For example: subject matter of the extracts, accents of the speakers, vocabulary use, etc.

Summary of response: All the lecturers commented on the fact that the subject-specific nature of some of the content particularly in Task 2 and 4 could be disadvantageous to some candidates in terms of background knowledge of the topic. However, the overall opinion was that the task content was fair. Seven of the nine students thought that there was a degree of bias. The basis for their argument was also the subject-specific nature of the topics of some of the tasks and the fact that the test is in English which could be a problem for non-English mother tongue speakers.

Question 16: If there were any questions with which you had specific problems, your feedback would be greatly appreciated.

Summary of response: The lecturers did not respond to this part of the questionnaire and only three of the students made comments. The first thought ALT was 'a great exercise ... and quite interesting'. The second had problems with the sound clip on Task 4 (this was as a result of not following instructions) but found the other three tasks 'very well set out'. The last student commented on the fact that he was interrupted by someone during the last task and so lost track of both the recording and the text.

The last section of the questionnaire was only for the lecturer's consideration and not for the students involved in the pilot project.

Question 1: How relevant did you find the test material in terms of measuring academic literacy as a whole?

Summary of response: Only two of the three lecturers responded to this question. One of the lecturers felt the test material was 'very relevant' the other felt that the subject-specific terminology over-emphasised the need for background knowledge.

Question 2: Do you think the test would be reliable as a way of 'streaming' incoming first year students?

Summary of response: Only two of the three lecturers responded to this question. One thought the test would be reliable while the other pronounced it 'mostly reliable'.

Question 3: Would you have confidence in the test results as an accurate measure of academic listening ability?

Summary of response: Only two of the three lecturers responded to this question. Both lecturers agreed that they would definitely have confidence in the results.

Question 4: What was your overall impression of the test? I would value your opinion on any issues which have not been covered in the questionnaire.

Summary of response: The following quote is from the only lecturer who chose to express an opinion: 'I think it is a good test – difficult, but good. I think you have selected situations that reflect accurately what students will be experiencing at university'.

4.3 QUALITATIVE FINDINGS: DISCUSSION OF RESULTS

In this section, the findings mentioned above will be analysed and discussed with reference to the research questions posed in Chapter 1.

4.3.1 Qualitative research questions

The following research questions are concerned with assessing the content, face and construct validity of ALT through the opinions of both colleagues and students. For face validity it is important for 'non-testers', such as students to 'comment on the value of the test' and for experts in the field to 'judge the test' in the case of content validity (Alderson et al., 1995:172). Matters of construct validity involve issues such as the layout, clarity of instruction and level of difficulty.

4.3.1.1 How representative and relevant are the tasks included in ALT?

Feedback on this question from both experts (lecturers) and non-experts (students) was very positive. All the lecturers felt that the tasks were completely relevant and representative and in the case of the students, the majority was satisfied with the relevancy of Tasks 1 and 2 with some reservations about Tasks 3 and 4. Since content validity is determined by professional opinion, the response from the lecturers seems to indicate that ALT can be deemed content valid with regard to the representativeness of its content. The students' response mostly indicates a positive reaction to the relevancy of the tasks included in ALT. In the feedback on Task 1, one of the students thought it 'very important to be able to follow such instructions'. Opinion on Task 2 included the following remark: 'the language used was of the exact quality that is used by lecturers while they are teaching'. A student's comment about Task 3 was that

'one should be able to follow a discussion to the extent that you can think for yourself, form an opinion and participate'. The following view was reported regarding Task 4: 'the task is relevant to listening to a lecture and trying to follow in the notes at the same time, requires summarising skills'. These quotes reflect some of the impressions of the test and could be indicative of good face validity.

4.3.1.2 Are experts in the field confident that the results of ALT would be an efficient indication of academic literacy levels?

This question was a matter of professional judgment and the lecturers involved in the project were divided in their opinions. One thought ALT was very reliable as an indicator of academic literacy levels, one didn't comment and one thought the tasks too subject-specific to yield a reliable demonstration of overall academic literacy.

4.3.1.3 Is the level of difficulty, layout, sound and visual quality of the clips and clarity of instruction included in ALT conducive to good construct validity?

Overall, both experts and students were satisfied with the level of difficulty of the tasks apart from Task 2 which some of the lecturers thought to be too dependent on background knowledge. There was general consensus on the computerised format of the test being easy to navigate and clearly laid out. The sound and visual quality of the audio and visual clips were considered good rather than excellent which, as I explained in the test description in the previous chapter, is sometimes a price that has to be paid for authentic, non-staged recordings. The majority of the participants in the pilot project thought the instructions were clear and unambiguous. This feedback along with the response to the other questions in the questionnaire mentioned earlier in the chapter seems to indicate that method effects were kept to a minimum and did not have a significant negative impact on the construct validity of ALT. On the advice of the statistician who assisted me in this project, it was decided that no changes be made to ALT as a result of the response to the questionnaire. His reasoning was that the beta group was too small to warrant making any significant changes to the items included in ALT.

4.4 QUANTITATIVE FINDINGS: ALT RESULTS

Various statistical analyses were carried out as part of a validation study of ALT itself as well as the two sets of scores it provided. An examination of how the various components relate to one another was carried out in addition to an investigation into the validity and reliability of the test as a whole in the assessment of academic listening.

4.4.1 Internal consistency

The first validation of the test involved computing the internal consistency for each of the four tasks by calculating the Cronbach alpha coefficients. The item-total correlation for each item was also measured in order to determine how each item related to the other items in the scale. A summary and discussion of the four tasks in each of the two test administrations is presented below. The two ALT administrations were carried out a month apart to determine reliability over a period of time. In the following correlation tables of the various tasks in the two test administrations, I will be focusing on the alpha coefficients and the item-total correlations. The other data is not specifically pertinent to my research questions but has been included to provide a complete overview.

TABLE 4.2: TEST ADMINISTRATION 1 TASK 1: INSTRUCTIONS

variable	Summary for scale: Mean=5.95652 Std.Dv.=2.40706 Cronbach alpha: .769989 Standardized alpha: .81589 Average inter-item corr.: .663146				
	Mean if deleted	Var. if deleted	StDv. if deleted	Itm-Totl Correl.	Alpha if deleted
Task 1: Question 1(test1)	5.756522	4.838646	2.199692	0.423296	0.752222
Task 1: Question 2_1(test1)	4.991304	4.869493	2.206693	1.000000	0.725375
Task 1: Question 2_2(test1)	5.113043	4.488077	2.118508	0.729738	0.716090
Task 1: Question 2_3(test1)	5.391304	4.027691	2.006911	0.738834	0.701115
Task 1: Question 2_4(test1)	5.373913	4.258888	2.063707	0.609960	0.723447
Task 1: Question 3(test1)	5.200000	4.613095	2.147812	0.513263	0.739896
Task 1: Question 4(test1)	5.695652	5.411918	2.326353	0.067942	0.798531
Task 1: Question 5(test1)	5.652174	4.818068	2.195010	0.353347	0.762706
Task 1: Question 6(test1)	5.330435	5.509175	2.347163	0.000117	0.812725
Task 1: Question 7(test1)	5.104348	4.876622	2.208308	0.472670	0.747275

TABLE 4.3: TEST ADMINISTRATION 2 TASK 1: INSTRUCTIONS

variable	Summary for scale: Mean=6.79208 Std.Dv.=2.01523 Cronbach alpha: .719138 Standardized alpha: .78050 Average inter-item corr.: .884940				
	Mean if deleted	Var. if deleted	StDv. if deleted	Itm-Totl Correl.	Alpha if deleted
Task 1: Question 1(test2)	6.554455	3.512014	1.874037	0.205456	0.728647
Task 1: Question 2_1(test2)	5.811881	3.519238	1.875963	0.922652	0.677755
Task 1: Question 2_2(test2)	5.871287	3.334249	1.825993	0.622296	0.671000
Task 1: Question 2_3(test2)	6.000000	2.730071	1.652293	0.839759	0.608338
Task 1: Question 2_4(test2)	5.990099	2.783418	1.668358	0.811238	0.615863
Task 1: Question 3(test2)	5.930693	3.155416	1.776349	0.607779	0.661836
Task 1: Question 4(test2)	6.544554	4.100354	2.024933	-0.152005	0.786917
Task 1: Question 5(test2)	6.445545	3.220174	1.794484	0.336273	0.708549
Task 1: Question 6(test2)	6.108911	3.476466	1.864528	0.189069	0.736015
Task 1: Question 7(test2)	5.871287	3.754549	1.937666	0.184839	0.721823

This task consisted of 10 items and as can be seen in the two tables of Task 1 shown above, question 4 shows a low item-total correlation in both administrations of ALT indicating that a revision of the item would be necessary for future versions of the test. Question 6 shows poor performance on the first testing but performs better in the second. The task requires candidates to listen for very specific information and deduce the meaning of words from their context. Questions 4 and 6 are multiple choice items where the distractors have only minimal differences between them which may have made the spread of answers too broad. The other items show an adequate item-total correlation. The task as a whole however, demonstrates an acceptable level of reliability with an alpha reading of 0.77 on the first test administration and 0.72 on the second.

TABLE 4.4: TEST ADMINISTRATION1 TASK 2: LECTURE EXTRACT

variable	Summary for scale: Mean=8.15652 Std.Dv.=2.69708 Cronbach alpha: .720324 Standardized alpha: .71654 Average inter-item corr.: .179535				
	Mean if deleted	Var. if deleted	StDv. if deleted	Itm-Totl Correl.	Alpha if deleted
Task 2: Question 1(test1)	7.678261	6.120522	2.473969	0.340247	0.704595
Task 2: Question 2(test1)	7.513043	5.592235	2.364791	0.613314	0.663286
Task 2: Question 3(test1)	7.330435	6.421059	2.533981	0.336444	0.704967
Task 2: Question 4(test1)	7.573913	6.630702	2.575015	0.132749	0.733965
Task 2: Question 5(test1)	7.452174	6.325267	2.515008	0.295160	0.710215
Task 2: Question 6(test1)	7.356522	6.377077	2.525287	0.333594	0.705060
Task 2: Question 7(test1)	7.460870	5.848555	2.418379	0.517060	0.679098
Task 2: Question 8(test1)	7.530435	5.854782	2.419666	0.479242	0.683750
Task 2: Question 9(test1)	7.469565	6.477358	2.545065	0.219706	0.720522
Task 2: Question 10(test1)	7.269565	6.883346	2.623613	0.136864	0.724562
Task 2: Question 11(test1)	7.643478	6.207484	2.491482	0.302613	0.710188
Task 2: Question 12(test1)	7.443478	5.821340	2.412745	0.542916	0.675787

TABLE 4.5 TEST ADMINISTRATION 2 TASK 2: LECTURE EXTRACT

variable	Summary for scale: Mean=8.83168 Std.Dv.=2.93388 Cronbach alpha: .815249 Standardized alpha: .82574 Average inter-item corr.: .294251				
	Mean if deleted	Var. if deleted	StDv. if deleted	Itm-Totl Correl.	Alpha if deleted
Task 2: Question 1(test2)	8.287128	7.025244	2.650518	0.473165	0.801640
Task 2: Question 2(test2)	8.099010	7.208067	2.684784	0.470662	0.801249
Task 2: Question 3(test2)	7.930693	7.884444	2.807925	0.327160	0.811998
Task 2: Question 4(test2)	8.237624	7.748667	2.783643	0.194803	0.828521
Task 2: Question 5(test2)	7.960396	7.501348	2.738859	0.495483	0.800653
Task 2: Question 6(test2)	8.019802	7.051578	2.655481	0.635054	0.787891
Task 2: Question 7(test2)	8.059406	7.504884	2.739504	0.366303	0.810133
Task 2: Question 8(test2)	8.059406	7.282576	2.698625	0.470070	0.801285
Task 2: Question 9(test2)	8.158416	7.252285	2.693007	0.415708	0.806728
Task 2: Question 10(test2)	7.960396	6.864782	2.620073	0.880697	0.772895
Task 2: Question 11(test2)	8.306931	6.582316	2.565603	0.659796	0.781793
Task 2: Question 12(test2)	8.069307	7.471338	2.733375	0.373863	0.809611

All twelve items in Task 2 are adequate with item-total correlations ranging from 0.13 to 0.88. The alpha readings of 0.72 on the first test administration, and 0.82 on the second, show good levels of reliability. Task 2 is a lecture comprehension task where candidates are expected to answer multiple choice questions based on what they have heard. Test-takers have to listen for specific details and be able to identify the main themes and supporting information. The input is explicit and the test-takers do not have to rely on inference to absorb the meaning. The type of listening required by the task is possibly the reason why the items performed well in this section of both test administrations.

TABLE 4.6: TEST ADMINISTRATION 1 TASK 3: DISCUSSION

variable	Summary for scale: Mean=5.43478 Std.Dv.=1.83124 Cronbach alpha: .312002 Standardized alpha: .32094 Average inter-item corr.: .039854				
	Mean if deleted	Var. if deleted	StDv. if deleted	Itm-Totl Correl.	Alpha if deleted
Task 3: Question 1(test1)	5.165217	2.926512	1.710705	0.132312	0.281859
Task 3: Question 2(test1)	4.747826	3.068261	1.751645	0.025225	0.326162
Task 3: Question 3_1(test1)	4.878261	3.431381	1.852399	-0.192281	0.418233
Task 3: Question 3_2(test1)	5.304348	3.009745	1.734862	0.172114	0.273975
Task 3: Question 3_3(test1)	4.608696	3.362415	1.833689	-0.130782	0.370508
Task 3: Question 3_4(test1)	4.800000	2.676706	1.636064	0.263883	0.219862
Task 3: Question 4(test1)	5.278261	3.192707	1.786815	-0.000340	0.327720
Task 3: Question 5(test1)	5.173913	2.512735	1.585161	0.444457	0.145347
Task 3: Question 6(test1)	4.947826	3.496516	1.869897	-0.225789	0.431885
Task 3: Question 7(test1)	4.791304	2.717527	1.648492	0.238954	0.232103
Task 3: Question 8(test1)	5.208696	2.665224	1.632551	0.354444	0.192602
Task 3: Question 9(test1)	4.878261	2.556684	1.598963	0.327811	0.184986

TABLE 4.7: TEST ADMINISTRATION 2 TASK 3: DISCUSSION

variable	Summary for scale: Mean=5.22772 Std.Dv.=2.06730 Cronbach alpha: .513971 Standardized alpha: .51514 Average inter-item corr.: .091473				
	Mean if deleted	Var. if deleted	StDv. if deleted	Itm-Totl Correl.	Alpha if deleted
Task 3: Question 1(test2)	4.920792	3.909147	1.977156	0.060057	0.530156
Task 3: Question 2(test2)	4.544554	3.798997	1.949102	0.119076	0.514712
Task 3: Question 3_1(test2)	4.742574	3.427187	1.851266	0.299623	0.461913
Task 3: Question 3_2(test2)	5.128713	4.349602	2.085570	-0.166482	0.556623
Task 3: Question 3_3(test2)	4.405941	3.313366	1.820265	0.553813	0.405691
Task 3: Question 3_4(test2)	4.623763	3.791277	1.947120	0.105501	0.520119
Task 3: Question 4(test2)	5.138614	3.912338	1.977963	0.211079	0.493632
Task 3: Question 5(test2)	4.950495	3.261230	1.805888	0.476144	0.412777
Task 3: Question 6(test2)	4.861386	3.266840	1.807440	0.420534	0.424652
Task 3: Question 7(test2)	4.455446	3.943219	1.985754	0.067438	0.524797
Task 3: Question 8(test2)	5.049505	3.484979	1.866810	0.419898	0.439881
Task 3: Question 9(test2)	4.683168	4.093665	2.023281	-0.054722	0.565324

The twelve items in Task 3 all demonstrate quite a low item-total correlation with Question 3 emerging as the poorest performer with two items in Test 1 and one item in Test 2 showing a negative value. Question 6 in Test 1 and question 9 in Test 2 also showed a negative value. The reliability coefficients for this task on the two test administrations are fairly low at 0.32 and 0.51 respectively. Since this task focused on inferencing skills it is no surprise that the reliability of the items have performed poorly. An explanation for this is that misunderstanding often occurs when there is conflict between what the speaker intended and the interpretation of the listener. Referring back to 2.8.4, research has shown that testing a listener's comprehension of explicitly stated or literal information is far easier than measuring a listener's inferential skills (Buck, 1991:86).

TABLE 4.8: TEST ADMINISTRATION 1 TASK 4: TUTORIAL EXTRACT

variable	Summary for scale: Mean=9.78261 Std.Dv.=4.86839 Cronbach alpha: .915511 Standardized alpha: .91913 Average inter-item corr.: .435255				
	Mean if deleted	Var. if deleted	StDv. if deleted	Itm-Totl Correl.	Alpha if deleted
task4_1(test1)	9.191304	20.55924	4.534230	0.604359	0.910511
task4_2(test1)	9.139131	20.37067	4.513388	0.669595	0.908377
task4_3(test1)	9.086957	20.26644	4.501826	0.728234	0.906603
task4_4(test1)	8.956522	20.98597	4.581044	0.681160	0.908780
task4_5(test1)	8.939131	20.87282	4.568678	0.750076	0.907300
task4_6(test1)	8.982609	20.50722	4.528490	0.780588	0.905836
task4_7(test1)	9.113044	20.48260	4.525771	0.655604	0.908841
task4_8(test1)	8.982609	20.72834	4.552839	0.715704	0.907602
task4_9(test1)	9.130435	20.34502	4.510545	0.680373	0.908036
task4_10(test1)	9.408695	20.39090	4.515628	0.656826	0.908785
task4_11(test1)	9.295652	21.04163	4.587116	0.480568	0.914616
task4_12(test1)	9.382608	20.81857	4.562737	0.545025	0.912430
task4_13(test1)	9.191304	20.07064	4.480027	0.722598	0.906594
task4_14(test1)	9.043478	20.57151	4.535583	0.685574	0.908063
task4_15(test1)	9.469565	22.68742	4.763131	0.134159	0.924349
task4_16(test1)	9.426087	21.56228	4.643520	0.382948	0.917388

TABLE 4.9: TEST ADMINISTRATION 2 TASK 4: TUTORIAL EXTRACT

variable	Summary for scale: Mean=11.4158 Std.Dv.=4.77990 Cronbach alpha: .937387 Standardized alpha: .94646 Average inter-item corr.: .597325				
	Mean if deleted	Var. if deleted	StDv. if deleted	Itm-Totl Correl.	Alpha if deleted
task4_1(test2)	10.62376	19.95857	4.467502	0.688901	0.933088
task4_2(test2)	10.61386	19.71086	4.439692	0.777605	0.931030
task4_3(test2)	10.58416	19.78159	4.447649	0.811155	0.930513
task4_4(test2)	10.54455	20.11786	4.485294	0.795987	0.931385
task4_5(test2)	10.54455	19.92288	4.463505	0.865095	0.930015
task4_6(test2)	10.50495	20.13661	4.487384	0.939975	0.929867
task4_7(test2)	10.63366	19.43569	4.408593	0.828467	0.929680
task4_8(test2)	10.54455	19.98567	4.470534	0.842763	0.930459
task4_9(test2)	10.63366	19.70041	4.438514	0.750639	0.931585
task4_10(test2)	10.84158	19.40037	4.404585	0.683321	0.933515
task4_11(test2)	10.85149	20.45371	4.522578	0.428467	0.940689
task4_12(test2)	10.90099	19.78866	4.448444	0.580860	0.936508
task4_13(test2)	10.74257	18.83425	4.339844	0.876210	0.927975
task4_14(test2)	10.54455	20.50582	4.528336	0.660504	0.934035
task4_15(test2)	11.12871	21.35753	4.621421	0.253248	0.944156
task4_16(test2)	11.00000	20.34830	4.510909	0.456535	0.939857

All sixteen items in this gap-fill task performed well and demonstrate a high internal rate of reliability and homogeneity which means they are all similar to one another relative to what is being assessed (Davies et al., 1999:75). The focus of the sample target language use (TLU) domain is fairly narrow in this task hence the similarity between component items. The alpha for the task in the first test administration was high at 0.93 with a slightly lower reading of 0.92 in the second test.

TABLE 4.10: WHOLE TEST RELIABILITY

variable	Summary for scale: Mean=29.3304 Std.Dv.=7.23263 Cronbach alpha: .625064 Standardized alpha: .6515ε Average inter-item corr.: .324827				
	Mean if deleted	Var. if deleted	StDv. if deleted	Itm-Totl Correl.	Alpha if deleted
task 1 score	23.37391	38.66888	6.218431	0.388979	0.578123
task 2 score	21.17391	30.92627	5.561140	0.608997	0.421724
task 3 score	23.89565	42.63259	6.529364	0.285365	0.632467
task 4 score	19.54783	19.03032	4.362376	0.501615	0.555357

The reliability coefficient for the test as a whole is only 0.63, which falls short of Weir's (2005:29) suggested optimum of 0.8. A possible reason for this could be the poor performance of Task 3. Tasks 2 and 4 appear to be the more reliable subtests with item-total correlations of 0.61 and 0.50 respectively.

4.4.2 Spearman correlation coefficients for each pair of subtests on ALT

TABLE 4.11: ALT SUBTEST CORRELATIONS

Task	1	2	3	4
1				
2	0.37 (p=0.00)			
3	0. 04 (p=0.64)	0.35 (p=0.00)		
4	0.28 (p=0.00)	0.53 (p=0.00)	0.24 (p=0.01)	

Correlations of the different test sections were calculated to determine the degree of difference or similarity in the attributes they were testing. The purpose in having different test sections is to test different abilities and this is usually indicated by reasonably low correlation coefficients. The correlations are thus expected to be in the region of 0.3 to 0.5 (Alderson et al., 1995:184). Since a correlation is only considered significant if its p value is smaller than 0.05, the p value will be included in the following discussion. The correlation coefficients between respectively Tasks 1 and 2, Tasks 2 and 3 and Tasks 2 and 4 are within or close to, these parameters and all show a p value of 0.00 which is an indication of significance. However, the correlation coefficient between Tasks 1 and 4 (p=0.00) and Tasks 3 and 4 (p=0.01) which, although significant are slightly below the minimum of 0.3. Tasks 1 and 3 (p=0.64) show an insignificant correlation since the p value is larger than 0.05 which is likely to be an indication of quite different traits being measured in the two tasks.

4.4.3 Test–retest correlation

The Spearman correlation method was again used to gauge the consistency of measurement of ALT over a period of time. As was mentioned in the test procedure in the previous chapter, a retest of ALT was given a month after the first test administration and the correlation between the results of the two administrations was then statistically calculated. The results are given in Figures 4.1 and 4.2.

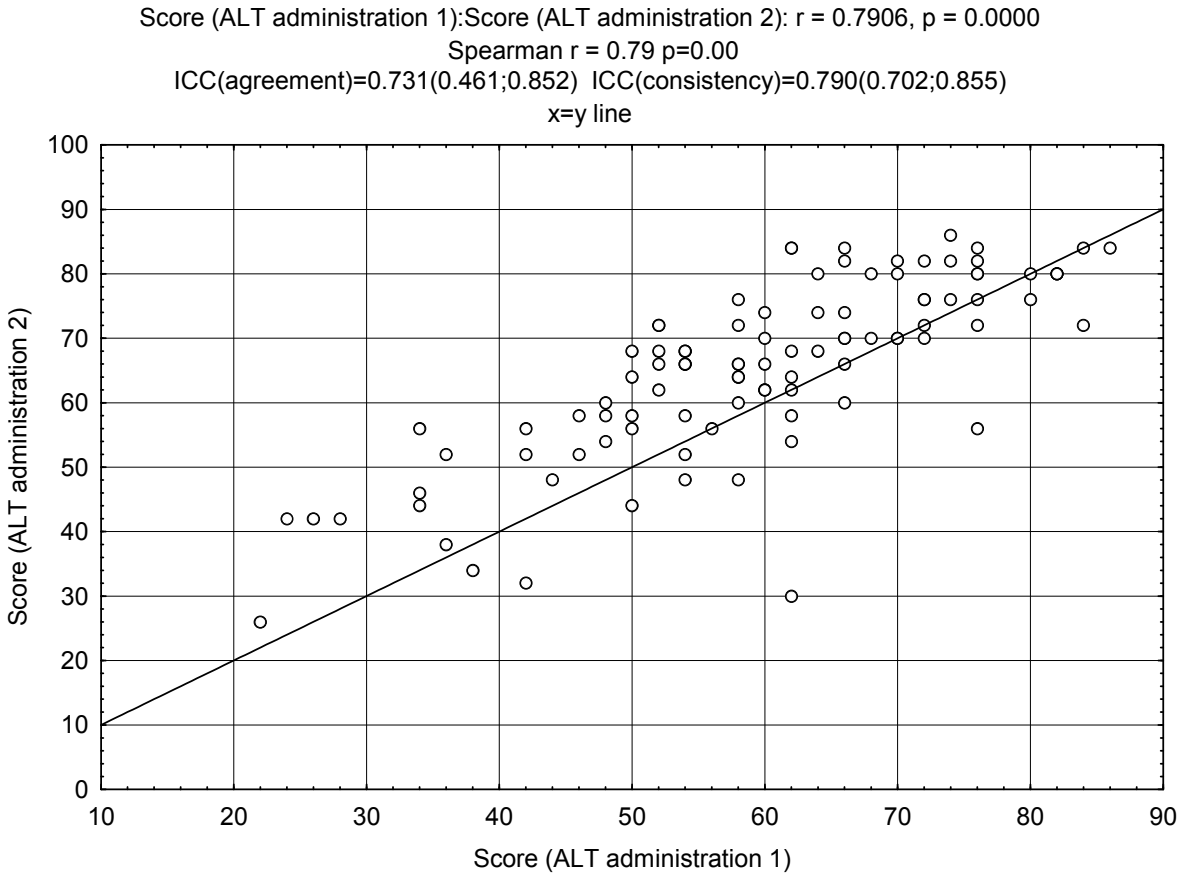

FIGURE 4.1: COMPARISON OF SCORES ON ALT ADMINISTRATION 1 WITH ALT ADMINISTRATION 2

The correlation value of 0.79 is considered as acceptable evidence that ALT shows good consistency of measurement and can thus be considered reliable. The p value of 0.00 is also evidence that the correlation is significant. The pattern that emerges from the graph as indicated by the clustering of circles above the line is that scores on the second ALT administration were higher than on the first. The intra-class correlation or ICC (agreement) in Figure 4.1 has a reading of 0.73 which indicates a good correspondence between the first and second administration. Because of an assumption that candidates would find the second sitting of ALT easier than the first, the 'agreement' reading of the ICC has a built-in penalty because of bias which the ICC (consistency) does not include. This explains the higher

reading of 0.79 for the 'consistency' measurement and confirms that the second sitting of ALT yielded better results than the first.

The Bland and Altman scatterplot (Figure 4.2) compares the difference between and the average of the two ALT administrations. The negative mean value shown in the figure is an indication that test-takers fared consistently better in the second test administration.

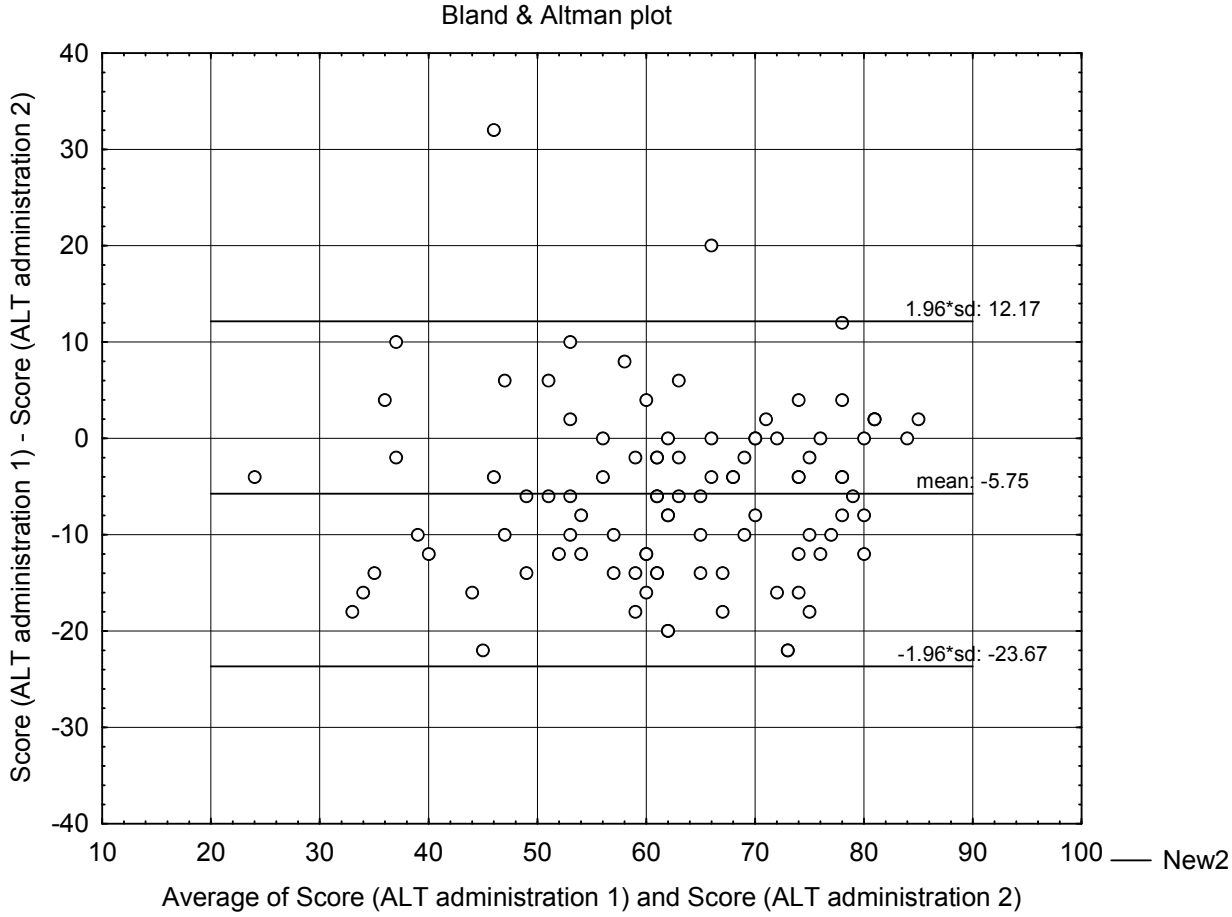

FIGURE 4.2: SCATTERPLOT OF THE DIFFERENCE BETWEEN SCORES ON ALT ADMINISTRATION 1 AND 2 COMPARED WITH THE AVERAGE OF THE SCORES ON ALT ADMINISTRATION 1 AND ALT ADMINISTRATION 2

4.4.4 Correlation with TALL (June 2008)

For reasons of concurrent validity, I decided to use the June 2008 TALL results as the criterion against which to measure the two administrations of ALT, since the tests were all taken within approximately the same time frame. According to Alderson et al. (1995:178), the correlation coefficient in a concurrent validity study should range from 0.5 to 0.7. The following scatterplots (Figure 4.3 and 4.4) illustrate that both the first and second ALT administration show a significant correlation coefficient with TALL. A correlation coefficient of 0.72 (p=0.00) on the first ALT administration and 0.67 (p=0.00) on the second signifies that both values fall well within the optimal boundaries.

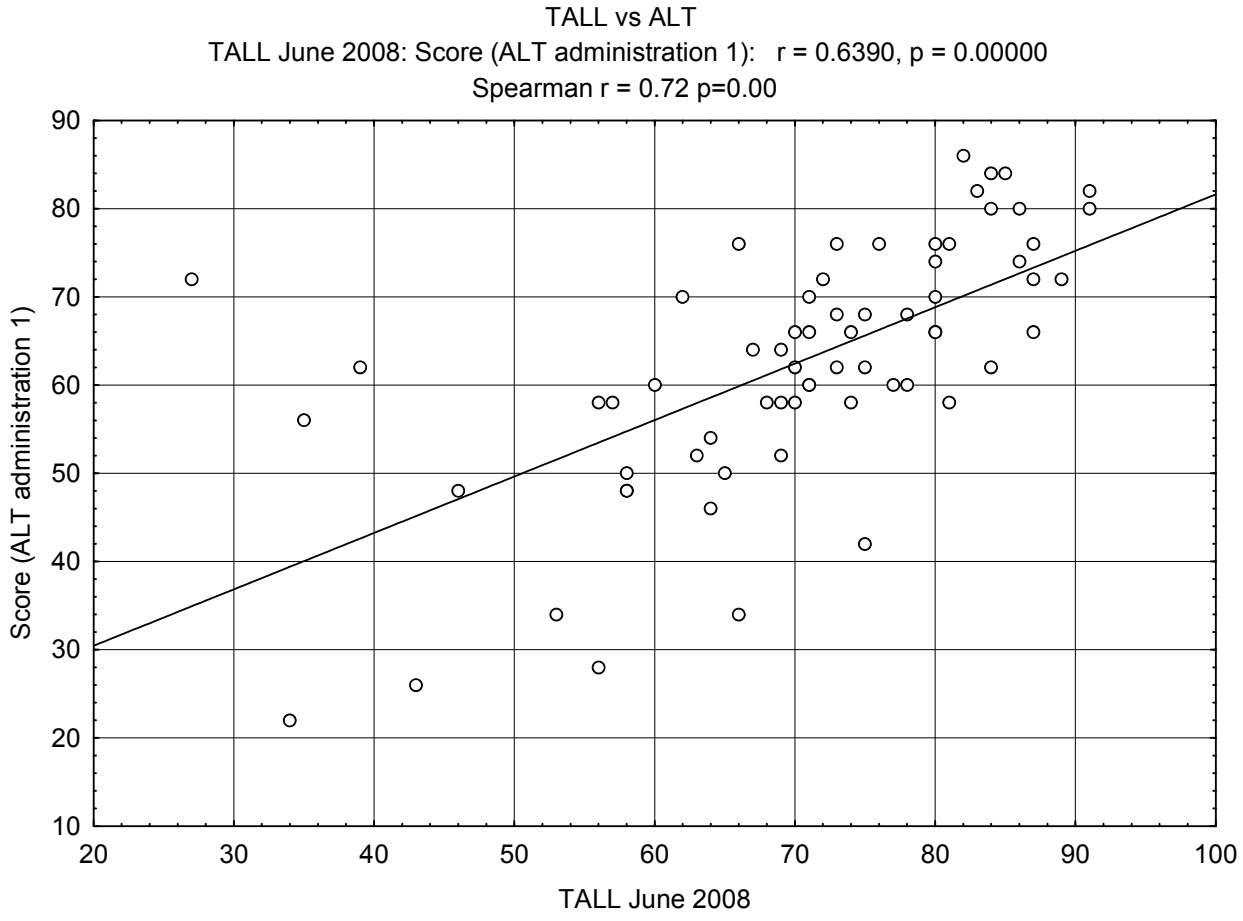

FIGURE 4.3: TALL JUNE 2008 CORRELATED WITH SCORE ON ALT ADMINISTRATION 1

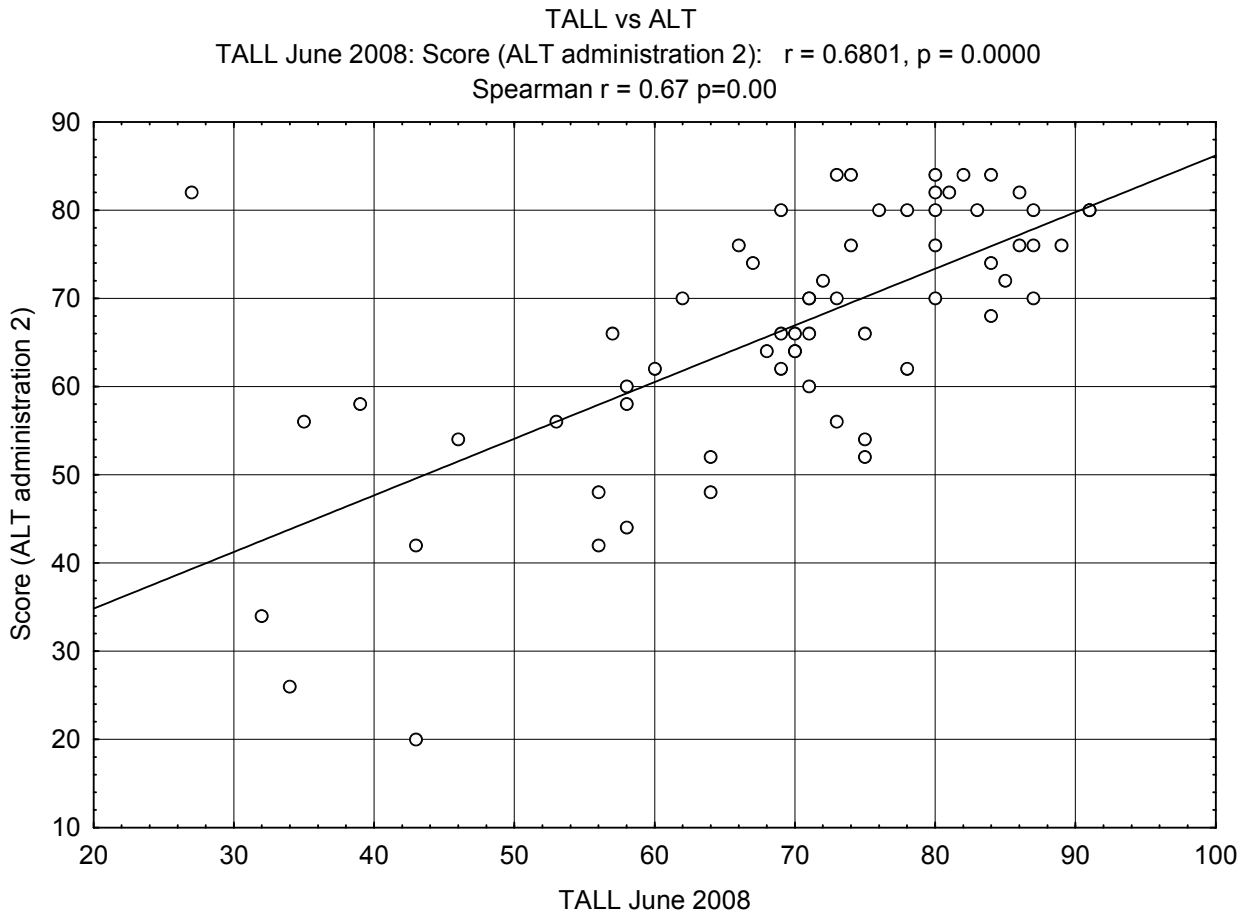

FIGURE 4.4: TALL JUNE 2008 CORRELATED WITH SCORE ON ALT ADMINISTRATION 2

4.4.5 Correlation between the scores of candidates in the three TALL scoring categories and their performance on ALT

As I mentioned in the previous chapter (3.3), students are sorted according to their TALL results into 'high risk' (code 1 and 2) and 'low to no risk' (code 4 and 5) categories. In the January 2008 TALL results which I used for this particular investigation, category 4 and 5 constitutes those students whose scores on TALL were higher than 58%. Category 1 and 2 is made up of candidates whose TALL score was lower than 54%. Candidates with TALL results of between 54 and 58% therefore form the borderline code 3 category and these students are later placed in either the top or the bottom group. Unfortunately, this is a very small group given the narrow score parameters and only seven percent of all the Bachelor of Science first year students fell into this category. It is therefore no surprise that the sample group contained only four code 3 students which was too few to reliably correlate with their scores on ALT. However, a comparison of these students' scores on TALL and on ALT yielded the following information.

TABLE 4.12: COMPARATIVE CODE 3 SCORES ON TALL AND ALT

	Score on TALL – January 2008	Score on ALT - administration 1	Score on ALT - administration 2
Student A	54%	24%	42%
Student B	54%	54%	68%
Student C	55%	34%	46%
Student D	56%	58%	66%

It seems feasible from this information that two of the four students could be placed in the top group and two in the bottom.

In order to get a picture of how the other two categories (1 & 2 and 4 & 5) on TALL correlated with ALT, the Spearman coefficients are given in the table below.

TABLE 4.13: CORRELATION BETWEEN TALL CATEGORIES (1 & 2 AND 4 & 5) AND THE TWO ADMINISTRATIONS OF ALT

	Code 1 and 2 (TALL)	Code 4 and 5 (TALL)
Administration 1 (ALT)	0.24 (p=0.28)	0.58 (p=0.00)
Administration 2 (ALT)	-0.02 (p=0.93)	0.41 (p=0.00)

These results indicate an insignificant correlation between candidates in the high risk category and their results on ALT. The low to no risk category of code 4 and 5 shows a much higher correlation. This could be as a result of a general lack of linguistic skills among the high risk group so the difference between reading and writing (as measured in TALL) and listening abilities (as measured in ALT) is more marked. In the low to no risk group it may be inferred that if candidates are accomplished at reading and writing they are more likely to succeed on listening tasks as well.

4.5 QUANTITATIVE FINDINGS: DISCUSSION OF RESULTS

Although some of the results of the various analyses and the patterns that emerge from them can only be speculated upon, there are some that serve as fairly conclusive evidence of reliability and validity.

4.5.1 Quantitative research questions

Even though the answers to these questions are based on statistical evidence, the reasons behind the various readings are still a matter for speculation and open to interpretation. In some cases the results were to be expected, in others the difficulty of interpreting linguistic ability in mathematical terms became apparent. The main patterns to emerge from the data will be discussed below in terms of the research questions mentioned in Chapter 1.

4.5.1.1 Does ALT provide internal consistency of measurement for each section?

This validation study involved the use of internal consistency coefficients to assess the degree of reliability for the four individual tasks as well as for each item included in the tasks. A low average inter-item correlation such as that found in Task 2 could be an indication of heterogeneity where different traits are being tested within the same task. Notwithstanding the evidence of heterogeneity in Task 2, the overall internal consistency for the task as shown by the Cronbach alpha coefficients of 0.72 and 0.82 in the first and second administrations respectively, display an acceptable level of reliability. Task 1, on the other hand, displayed an inter-item correlation of 0.66 which seems to indicate that the items in the task were more homogenous than in Task 2. Since Task 3 was based on implicit rather than explicit information and test-takers were required to make use of inferencing skills in order to respond to the questions, a low alpha and inter-item correlation was expected. The alpha reading of 0.31 on Task 3 in the first test administration was the lowest of the eight readings. This improved in the second sitting of ALT to 0.51 possibly because the candidates knew exactly what to listen for in the retest. Task 4's reliability coefficients of 0.94 and 0.92 on the two test administrations are substantially higher than the norm of 0.8 (Weir, 2005:29), which indicates strong reliability.

The test as a whole only displayed an alpha measurement of 0.63, which was to be expected since the test intentionally contained a variety of heterogeneous items. However, when parallel methods of checking for reliability were used in the test-retest, ALT showed that as a complete test it can be considered reliable.

4.5.1.2 Do the internal correlations of the different sections in ALT provide evidence of construct validity?

Alderson et al. (1995:183-4) recommends assessing the construct validity of a test by comparing the different components or subtests with each other. They also reason that since each subtest presumably measures different abilities, the correlations cannot be expected to be very high. In the correlation between Task 1 and Task 2 in ALT, the reading of 0.37 (p=0.00) falls within the range of 0.3 to 0.5 recommended by Alderson et al. (1995:184). This significant correlation stands to reason since the items included in both these tasks require the test-taker to listen for specific mostly explicit information but they are not so similar as to be testing exactly the same skills. However, when Task 1 was correlated with Task 3, the correlation was insignificant with a p value of 0.64 – since Task 3 focused on entirely different listening skills from those required in Task 1. Task 3 was based on implied information where candidates had to infer meaning from what they had heard in the audio clip, while Task 1 involved recognising and remembering specific instructions. The correlation between Task 2 and Task 3 (p=0.00) was within the optimum parameters of 0.3 and 0.5 which was surprising, since the traits that Task 2 was supposed to measure had more in common with Task 1 which had an insignificant correlation with Task 3. The correlation of 0.28 between Task 1 and Task 4, which proved to be significant with a p value of 0.00, is closer to the 0.3 mark. This was to be expected, since Task 4 also involved listening for facts and deducing the meaning of words from the context, in fact one might have expected the correlation to be even higher. The high correlation of 0.53 between Task 2 and Task 4 (p=0.00) was also somewhat unexpected as although there were traits shared by both tasks, essentially the two tasks were designed to test mostly different abilities. The low but still significant correlation of 0.24 (p=0.01) between Task 3 and Task 4 was again to be expected since the traits measured by the two tasks had little in common. These correlations all indicate that the individual tasks are each assessing different abilities to a greater or lesser degree which seems to demonstrate some measure of construct validity.

4.5.1.3 Is there a significant difference in scores from the first ALT administration and the retest?

As was mentioned earlier, ALT was administered twice to the same group of students with an interval of a month between the two sittings. In spite of the fact that test-takers fared better on the second sitting of ALT, the correlation between the results of the first and second test administrations is significant at 0.79 (p=0.00) which indicates a strong accord between the two sets of scores. This in turn, is also indicative of good test reliability.

4.5.1.4 Does ALT show concurrent validity by demonstrating a high correlation with TALL?

As I reported in the results, scores on TALL of June 2008 (because of being administered at approximately the same time as the two administrations of ALT) were used as the criterion for the correlation with the two ALT administration results. Alderson et al. (1995:178) state that 'a classic concurrent validation would involve comparing scores on the test in question with scores on some other test known to be valid and reliable'. Since TALL has proven reliability and validity (Van der Slik & Weideman, 2005; Van der Walt & Steyn, 2007), the correlation coefficient of 0.72 (p=0.00) measured on the first ALT administration and 0.67 on the second, provide convincing evidence of concurrent validity. In light of the fact that both TALL and ALT are language tests assessing academic communication skills, this result was not unexpected since there are aspects common to both test constructs, for example, understanding vocabulary from the context and making inferences from implied information. However, the correlation is not so high as to be an indication that the two tests are measuring exactly the same skills. The results therefore, seem to show that the reading and writing skills assessed in TALL and the listening skills measured in ALT, although related, are still different aspects of language ability.

4.5.1.5 What is the correlation between the scores of the borderline TALL test-takers and their performance on ALT?

As was reported earlier, one of the main aims of this research was to determine whether ALT would be able to assist in more accurate screening of the students placed in the borderline category based on their performance on TALL. Since this group consisted of only four students in the sample group, a reliable correlation between their TALL scores and their performance on ALT was not possible. However, just by comparing their scores on TALL and on ALT, some insight was provided by the difference in the candidates' performance on the two tests. All four candidates' scores on TALL were between 54 and 56%, yet two of the candidates' scores remained the same or improved on the two administrations of ALT while the other two test-taker's scores declined. Perhaps this is an indication that given larger numbers of code 3 candidates in future studies, ALT might yet be conclusively proved helpful as an additional screening tool.

4.6 CONCLUSION

The results presented and discussed in this chapter have provided useful qualitative and quantitative information regarding ALT. Perhaps the most important of these findings is the evidence of the reliability of ALT since no test can be valid if it is not reliable. Reliability measures were performed by inter-item and item-total correlations as well as the test-retest method of checking for consistency of measurement over a period of time. Issues of validity were the next most important to investigate and, based on the qualitative feedback in the completed questionnaires, the degree of face validity, content validity and some aspects of construct validity could be established. Statistical correlations between the different pairs of test components and between ALT and TALL provided further evidence of construct as well as concurrent validity. Additional research into the construct validity of each task in the test by means of factor analysis was beyond the scope of this particular study but would be an interesting exercise for the future. Other future research recommendations as well as an investigation into the relevance of this study will be described in the next chapter.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 INTRODUCTION

As I have explained in previous chapters, the University of Stellenbosch has implemented a Test of Academic Literacy Levels (TALL) which focuses primarily on reading and writing abilities. The results are used to place students into academic literacy intervention programmes designed to increase their chances of academic success. The rationale for this study has been to investigate the potential for an academic listening test (ALT) to assist in more accurate screening of students whose performance on the test is between 'high risk' and 'low to no risk', as classified by the coding system implemented by TALL administrators (Van Dyk, 2007). This chapter forms a summary of the study and discusses the outcomes of the research with relevance to the literature and theory on which ALT is based. The relevance and shortfalls of this project will also be described along with recommendations for future research into listening testing.

In Chapter 1 the context of the study was provided by the theoretical, empirical and practical reasons for my choice of research topic. The principal research questions, both qualitative and quantitative, that guided the study and led to the formulation of my hypothesis were also presented. The methodology used to address the research objectives was then described and the chapter concluded with an outline of the rest of the study.

The review of the literature relating to my study and described in Chapter 2 began with broadly based research into language testing in general, specific purpose testing, computer-based testing and academic literacy. This provided the necessary background knowledge to the field of listening testing and more specifically academic listening assessment.

Chapter 3 discussed the design, operationalisation and administration procedures of ALT. The processes used for the collection of qualitative and quantitative data were described as well as demographic details of the participants in both the pilot project and the main study. This chapter concluded with an explanation and rationale for the data analysis methods which were employed.

The results presented and discussed in Chapter 4 provided useful qualitative and quantitative data regarding ALT. One of the most important results to emerge from the validation study described in this chapter is the evidence of reliability, since no test can be valid if it is not reliable.

5.2 OUTCOMES

In a quest to confirm or refute my hypothesis that an academic listening test would be a useful added dimension to TALL it was necessary to collect and analyse extensive qualitative and quantitative data.

5.2.1 Qualitative data

The qualitative data resulted from response to a questionnaire (see Appendix A) by both 'experts' and students during a pilot project. The most important results to emerge from this data were evidence of construct, content and face validity.

It may be noted that construct validity can be assessed both qualitatively and quantitatively because it is concerned with how well the tasks included in a test, mirror real life requirements (Davies, 1990:23). Feedback from the lecturers involved in the project was that the tasks were completely relevant and representative. The majority of the students were also satisfied with the relevancy of Tasks 1 and 2 but opinion was more divided on Tasks 3 and 4. There are also certain factors which can threaten the construct validity of a test and prevent an accurate assessment of the construct. The degree of difficulty of items or tasks, the quality of the instructions and familiarity with computers are all factors which can affect the construct validity of a test. Inappropriate difficulty or inaccuracies serve to create construct-irrelevant variance and thus threaten validity. The questionnaire invited comment on these and other such factors and the response indicated that features of the design, layout and method of delivery of ALT that could adversely affect a candidate's performance, were not significant.

Content validity, as mentioned previously, stems from the professional judgement of experts and is concerned with the 'representativeness or sampling adequacy of the content' (Alderson et al., 1995:173). It clearly has strong links with construct validity and many of the questions in the questionnaire were aimed at assessing both these types of validity. Therefore, as described above, the relevancy and representativeness of the tasks would apply to content validity as well. The second question in the questionnaire pertained more specifically to content validity when the beta group was asked how interesting and relevant they found the content. Feedback from both lecturers and students was generally positive with some reservation on the part of the lecturers on the subject-specific nature of some of the tasks.

When asked which skills they thought were being measured in the various tasks, there was mostly agreement among the participants but some of the opinions did not concur with my intentions in the design of the tasks. These discrepancies will be discussed in the following paragraph.

With reference to 2.8.1, the traditional view that listening is divided into a two-stage process was used as a framework for ALT and was explained in the test description included in 3.10. Task 1 was primarily based on bottom-up or local skills such as listening for specific information but also included top-down processing skills in that candidates were required to deduce the meaning of words through the context. Most of the beta group participants agreed that this task involved direct meaning comprehension skills. However, in Task 2, where the emphasis was again on direct meaning from explicit information, the majority of the students and one of the lecturers thought that inferred meaning was required. In Task 3, which involved mainly top-down processing to infer the meaning from implied information, only two of the lecturers and three of the students were in agreement. Task 4 relied on listening for specific details and supplementing information with little or no inferencing required. However, all the lecturers and two of the students thought that inferencing was indeed necessary. This seems to confirm Buck (2001:106) and Brindley's (1998:173) opinion that it is extremely difficult to apportion different abilities accurately to set cognitive approaches. On balance, however, the content of the test seems to have been considered valid and representative of abilities needed in the TLU domain by most of the lecturers and students involved in the pilot project.

As described in section 2.3.2.2 face validity refers to a test's 'public acceptability' or the impression of a test by a non-expert. The students in the beta group were not supervised during the test administration and several later admitted that they had failed to read all the instructions and had thus experienced problems accessing some of the media files. Consequently, an accurate opinion on their test experience was not entirely possible. In the questionnaire the students were invited to give their opinion on the test as a whole but unfortunately only three of the nine students responded. Only one of the three had not experienced any technical hitches and thought the test was 'a great exercise ... and quite interesting'. The students' positive response to the relevancy of the tasks is, however, an indication of their confidence in ALT and could thus be interpreted as a sign of good face validity.

5.2.2 Quantitative data

As indicated in Chapter 2, researchers such as Bachman and Palmer (1996:21) and Messick (1996:243) have stated that enough empirical evidence has to be provided to justify the

interpretation of a test score. In the validation study carried out on ALT, statistical evidence of its reliability and validity was presented in the previous chapter. As the literature has shown, the two concepts have a somewhat complex relationship which is as yet not clearly understood (Alderson et al., 1995:186). It must be noted, however, that interpretations regarding test scores can never be considered completely valid (Bachman & Palmer, 1996:22).

In Chapter 2, communicative tests (like ALT) are discussed as being tests with fairly context-specific tasks that simulate real life and because there is likely to be more than one way of interpreting a text, often display low measures of reliability. This makes reliability of assessment very difficult, especially when candidates are asked to make inferences about a speaker's implied meaning (Buck, 2001:84-5). This reasoning is borne out in the first validation study that was performed on ALT. The item-total internal consistency coefficients were calculated for each task so as to assess the degree of reliability for the four individual tasks as well as for each item included in the tasks.

In Task 1, the items that showed the lowest levels of reliability were the questions that required the candidates to choose a response from very closely related distractors and relied on correct interpretation of the text. Task 3 showed the lowest levels of reliability presumably because this task involved making inferences about the opinions of the speakers. Task 2 and 4 emerged with the best reliability coefficients possibly because they relied very little on inferencing skills on the part of the test-takers: the correct responses could all be arrived at by listening to explicitly given information.

Another measure of reliability, the test-retest method was performed on ALT by administering the same test to the same candidates with an interval of a month between the two sittings. The results reported in the previous chapter show that the consistency of measurement between the two test administrations was fairly high at 0.79 ($p=0.00$) which indicates good consistency of measurement over a period of time. This reading is based on the intra-class correlation or ICC (consistency) measurement which unlike the ICC (agreement) does not include a penalty based on the proviso that candidates would find the retest of ALT easier than the first administration. Figures 4.1 and 4.2 confirm that this was indeed the case with scores from the second test administration reflecting consistently better performance than in the first sitting. This phenomenon was not unexpected since the test-takers knew what to expect as well as precisely which facts to listen for in the second session of ALT.

According to Bachman and Palmer (1996:22), when interpreting test scores the features of the test tasks need to be considered in order for the test to be considered construct valid. In order

to examine this, Alderson et al. (1995:184) recommend correlating the different components of a test to ascertain to what extent they all measure different attributes. On the basis that each test section supposedly measures something different, the correlation coefficient between these different subtests is not expected to be very high, in the region of 0.3 to 0.7. The results of the study performed on ALT show that only three of the six subtest correlations met the criterion with the other three being lower than 0.3. Given the nature of the task features, this indication that very different abilities are measured is not unexpected, particularly where the inferencing abilities of Task 3 are measured against the more direct comprehension skills of Tasks 1 and 4. The 0.28 (p=0.00) correlation measured between Tasks 1 and 4, although significant, is perhaps more surprising since both tasks involved listening for fairly specific detail. Alderson et al. (1995:187) point out that low inter-correlations between subtests are indicative of more heterogeneous items. As a consequence, a lower internal consistency reliability index is measured for the whole test. However, this could also be an indication of high construct validity.

In section 2.3.2.3, I described the importance of comparing an 'invalidated' test like ALT with a criterion such as another language test which has proven reliability and validity. Bachman (1990:290) maintains that the criterion can be used to validate the abilities that are assessed in the test. My investigation into the concurrent validity of ALT meant that the criterion and ALT had to have been administered in approximately the same time frame (Hughes, 2003:27). I therefore chose the TALL results of June 2008 since the two test administrations of ALT were carried out in July and August. As I reported in the previous chapter, the correlation measurements for both the test administrations fell well within the boundaries of 0.5 to 0.7 cited by Alderson et al. (1995:178) as being the usual parameters and showed significance with p values of 0.00. A relatively high correlation might have been expected since both tests assess receptive skills and both measure an aspect of language proficiency. A correlation higher than .90 could well have indicated that the abilities being measured were very similar. Thus, it would appear from this correlation that the skills required by TALL and ALT are different enough to lower the correlation but also sufficiently similar to prove that ALT shows concurrent validity.

As I reported in 4.4.5, a study of the correlation between the borderline candidates on TALL and their results on the listening test could not be reliably calculated given the very small number of students in the sample group. However, a comparison of their scores on the two tests gives an indication that ALT was in fact useful as an additional screening agent for sorting these students into either a higher or a lower group.

5.2.3 Final conclusions

The qualitative data to emerge from the validation study of ALT indicated that the method effects of the design, layout and mode of delivery that could adversely affect a candidate's performance, were kept to a minimum. Most of the lecturers and students involved in the pilot project pronounced it valid in terms of content and representative of abilities needed in the TLU domain. The students' positive response to the relevancy of the tasks is an indication of their confidence in ALT and could thus be construed as a token of good face validity. The quantitative data indicate that the tasks reliant on explicitly given information demonstrated the best reliability coefficients, possibly because they relied very little on inferencing skills on the part of the test-takers. A correlation of each of the tasks with the other tasks showed a significant correlation in five of the six correlations, which signifies fairly good construct validity. ALT, by means of a test-retest correlation, also showed consistency of measurement over a period of time as evidence of reliability. An investigation into the concurrent validity of ALT with the results of the June 2008 TALL as a criterion, showed that the skills required by TALL and ALT are sufficiently similar to prove concurrent validity. However, the correlation was not so high as to indicate that exactly the same abilities are being assessed. The scores of candidates, whose results in the January 2008 TALL were considered borderline, were compared with their performance on ALT. The comparison gave an indication that ALT could be useful as an additional and perhaps more accurate, means of sorting these students into either a higher or a lower group.

5.3 RECOMMENDATIONS FOR FURTHER RESEARCH

Since most validation studies are an ongoing process there is always scope for revision and improvement in any language test. This process is reliant on research and in the field of listening testing there appears to be much research that still needs to be done. According to Wagner (2002:26), it is a field that has attracted an increasing amount of attention over the last ten years but still has many unanswered questions. The following recommendations are some examples of research that could be investigated in the future.

- The two-factor model of listening assessment which I used as a framework for the operationalisation of ALT is an area that needs more research (Wagner (2002:26). A factor analysis study as part of this research would have been useful in validating the construct definition of ALT as shown in the model (see Figure 3.1). The literature has also shown that less competent students tend to use more bottom-up than top-down processing skills and it would be interesting to research whether this could account for

generally poor performance on inferencing tasks as was my experience with Task 3 in this study.

- As I mentioned in the previous chapter, one of the main aims of my research was to determine whether ALT would be able to assist in more accurate screening of the students placed in the borderline category based on their performance on TALL. Since I was unable to provide conclusive evidence of this owing to the small size of the sample, further research targeting this group in particular would be necessary in the future.

- Chapelle and Douglas (2006:91) maintain that computerised testing provides a heightened engagement between the test-taker and the task since everything happens on screen and the use of multimedia increases interactiveness. My observations during the course of this study definitely confirm this statement. Furthermore, computerised tests have the potential to deliver tests which are closer to real life with more authentic listening constructs than conventional tests. Thus, more research needs to be done on how test developers define the constructs they wish to measure and how successful the computer will be in assessing this. In this way a solution might be provided to the question of whether computers will significantly contribute to language testing in the future (Chapelle, 2001:21). A computerised version of TALL could form part of an investigation into whether this is indeed the case.

- The design of an Afrikaans academic listening test that replicates the construct of ALT could be a very worthwhile future project. A comparison of the scores of the two tests could potentially yield pertinent information that would contribute to research in the field of academic literacy.

5.4 FINAL REMARKS

In this study listening skills were assessed individually as well as in combination with other language skills, so ALT should not be seen as a 'pure' assessment of listening ability (Bejar et al., 2000:5). The integration of speaking, listening, reading and writing make up an individual's proficiency in a language but the mastery of these skills is seldom even (Lado, 1961:25). There has been some criticism of the skills-based approach to testing and Bachman and Palmer (1996:78) argue that rather than trying to differentiate between the four language skills, it would be more useful to pinpoint specific tasks involving language use. These tasks could then be described in terms of their various characteristics as well as the kinds of language ability they display. In this project both the identification of TLU tasks that will be

required of students in the future and the simulation of those tasks in an academic listening test formed the core of my study.

The significance of listening as a facet of academic literacy has been investigated in this research and it can be concluded from the results that listening competency although a component of an integrated set of language abilities (Rost, 2002:172), plays an important role in the assessment of academic competency. Moreover, the findings of this study confirm my hypothesis mentioned in 5.2, that a listening assessment instrument such as ALT could be a useful added dimension to TALL, the existing test of academic literacy levels currently used at the University of Stellenbosch.

BIBLIOGRAPHY

Alderson, J. C., Clapham, C. & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency*. London: Continuum.

Anderson, A. & Lynch, T. (1988). *Listening*. Oxford: Oxford University Press.

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Baker, E. L. (1998). *Understanding educational quality: Where validity meets technology. The fifth annual William Angoff Memorial Lecture.* Princeton, NJ: Educational Testing Service.

Barnard, H. (2008). *Foreign Direct Investment*. Gordon Institute of Business Science. [Online]. Available: http://www.gibs.co.za//home.asp?pid=2166 [02/05/2008]

Bejar, I., Douglas, D., Jamieson, J., Nissan, S. & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series Report No. 19). Princeton, NJ: Educational Testing Service.

Boughey, C. (2000). Multiple Metaphors in an understanding of academic literacy. *Teachers and Teaching: Theory and practice,* 6(3):279-290.

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics,* 18:171-191.

Brown, G. (1995). Dimensions of difficulty in listening comprehension. In D. Mendelsohn & J. Rubin (eds.). *A guide for the teaching of second language listening*. San Diego: Dominie Press. 59-73.

Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing,* 8:67-91.

Buck, G. (2001). *Assessing listening.* Cambridge: Cambridge University Press.

Call, M. (1985). Auditory short-term memory, listening comprehension, and the input hypothesis. *TESOL Quarterly*, 19:765-81.

Chapelle, C. (2001). *Computer applications in second language aquisition: Foundations for teaching, testing and research.* Cambridge: Cambridge University Press.

Chapelle, C. (2003). *English language learning and technology.* Philadelphia: John Benjamin's Publishing company.

Chapelle, C. & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge: Cambridge University Press.

Chaudron, C. & Richards, J. C. (1986). The effect of discourse markers on the comprehension of lectures. *Applied Linguistics,* 7(2):113-127.

Chaudron, C., Loschky, L. & Cook, J. (1994). Second language listening comprehension and note-taking. In J. Flowerdew (ed.). *Academic Listening: Research perspectives.* Cambridge: Cambridge University Press. 75-92.

Clapham, C. (1993). Is ESP testing justified? In D. Douglas & C. Chapelle (eds.). *A new decade of language testing research.* Washington DC: Teachers of English to speakers of other languages. 257-71.

Cliff, A. (2003). *Assessing the academic literacy skills of entry-level students using the Placement Test in English for Educational Purposes (PTEEP).* [Online]. Available: http://www.ched.uct.ac.za/seminars/archive2003/cliff2.pdf. [20/01/2008].

Cliff, A. & Ramaboa, K. (2007). The assessment of entry-level students' academic literacy: Does it matter? *Ensovoort,* 11(2):33–48.

Cooper, P. & Van Dyk, T. J. (2003). Measuring vocabulary: A look at different methods of vocabulary testing. *Perspectives in Education,* 21(1):67-80.

Davidson, F. & Lynch, B. (2002). *Testcraft.* Newhaven: Yale University Press.

Davies, A. (1990). *Principles of language testing.* Oxford: Basil Blackwell Limited.

Davies, A., Brown A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Studies in language testing 7: Dictionary of language testing.* Cambridge: Cambridge University Press.

Developing general listening skills. [s.n.] [s.a.] [Online]. Available: http://www.cambridgeesol.org/teach/cpe/listening/aboutthepaper/develop_gen_listening.htm. [23/01/2008].

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Dunkel, P., Henning, G. & Chaudron, C. (1993). The assessment of an L2 listening comprehension construct: A tentative model for test specification and development. *Modern Language Journal,* 77:180-191.

Dunkel, P. & Davis, J. N. (1994). The effects of rhetorical signally cues on recall. In J. Flowerdew (ed.). *Academic listening*: *research perspectives*. Cambridge: Cambridge University Press. 55-74.

Dunkel, P. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *MSU,* 2(2):77-93.

Elder, C., Erlam, R. & von Randow, J. (2002). *Enhancing chances of academic success among first year undergraduates from diverse language backgrounds*. [Online]. Available: http://www.fyhe.qut.edu.au/past_papers/papers02/ErlamPaper.doc. [23/01/2008].

Elder, C. & von Randow, J. (2007). *Enhancing chances of academic success amongst first year undergraduates from diverse language backgrounds*. Paper presented at EALTA conference, Sitges, Spain. (June)

Ferris, D. & Tagg, T. (1996). Academic listening/speaking tasks for ESL students. *TESOL Quarterly,* 30(2):297-320.

Flowerdew, J. (1994). Research of relevance to second language comprehension: an overview. In J. Flowerdew (ed.). *Academic listening*: *research perspectives.* Cambridge: Cambridge University Press. 7-29.

Fulcher, G. (1999). Assessment in English for Academic Purposes: Putting content validity in its place. *Applied Linguistics,* 20(2):221-236.

Gee, J. P. (1990). *Social linguistics and literacies: Ideology in discourses*. London: Falmer Press.

Glaser, G. R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist* 18:519-21.

Godwin-Jones, B. (2001). Emerging technologies: Language testing tools and technologies. *Language Learning & Technology,* 5(2):8-13.

Hansen, C. (1994). Topic identification in lecture discourse. In J. Flowerdew (ed.). *Academic listening*: *research perspectives*. Cambridge: Cambridge University Press. 131-145.

Hansen, C. & Jensen, C. (1994). Evaluating Lecture comprehension. In J. Flowerdew (ed.). *Academic listening*: *research perspectives.* Cambridge: Cambridge University Press. 241-268.

Hubbard, P. (2008). *An invitation to CALL: Foundations of computer-assisted language learning*. [Online]. Available: http://www.stanford.edu/~efs/callcourse/CALL8.htm [12/09/2008].

Hughes, A. (2003). *Testing for language teachers.* Cambridge: Cambridge University Press.

Jordan, R. R. (1997). *English for academic purposes*. *A guide and resource book for teachers.* Cambridge: Cambridge University Press.

Kellerman, S. (1992). "I see what you mean". The role of kinesic behaviour in listening and implications for foreign and second language learning. *Applied linguistics,* 13:239-258.

Krashen, S. D. (1985). *The Input Hypothesis.* London: Longman.

Kuehn, P. (1996). *Assessment of Academic literacy skills: Preparing minority and limited English proficiency (LEP) students for postsecondary education*. California State University (UCLA). Report for the Improvement of Postsecondary Education, Washington D.C.

Lado, R. (1961). *Language testing.* London: Longman.

LeLoup, J. W. & Ponterio, R. (2007). Listening: You've got to be carefully taught. *Language Learning & Technology,* 11(1):1, 4-15.

Lund, R. (1991). A comparison of second language listening and reading comprehension. *Modern Language Journal,* 75:196-204.

Lynch, T. (1998). Theoretical perspectives on listening. *Annual Review of Applied Linguistics,* 18:3-19.

McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.

McNamara, T. & Roever, C. (2006). *Language testing: The social dimension.* Oxford: Blackwell Publishing.

Messick, S. (1989). Validity. In R. Linn (ed.). *Educational measurement* (Third edition). New York: American Council on Education and Macmillan. 13-103.

Messick, S. (1996). Validity and washback in language testing. *Language Testing,* 13:242-256.

Mouton, J (1996). *Understanding social research*. Pretoria: van Schaik Publishers.

Mouton, J. (2001). *How to succeed in your master's and doctoral studies: A South African guide and resource book*. Paarl: van Schaik Publishers.

Popham, W. J. (1978). *Criterion referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Read, J. (2002). *The use of interactive input in EAP listening assessment*. [Online]. Available: http://www.sciencedirect.com/science. [28/01/2008].

Rost, M. (1990). *Listening in language learning.* London: Longman.

Rost, M. (2002). *Teaching and researching listening.* London: Pearson Education Limited.

Rubin, J. (1994). A review of second language listening comprehension research. *The Modern Language Journal,* 78(ii):199-221.

Shohamy, E. & Inbar, O. (1991). Validation of listening comprehension tests: The effect of text and question-type. *Language Testing,* 8:23-40.

Spolsky, B. (1995). *Measured words.* Oxford: Oxford University Press.

Stevenson, D. K. (1985). Pop validity and performance testing. In Y. P. Lee, A. C. Y. Y. Fok, R. Lord & G. Low (eds.). *New Directions in Language Testing.* Oxford: Pergamon Institute of English.

Van der Slik, F. & Weideman, A. J. (2005). The refinement of a test of academic literacy. *Per Linguam,* 21(1):23-35.

Van der Slik, F. (2005). Statistical analysis of the TALL/TAG 2004 results. Presentation to test development session, 1-3 June. University of Pretoria.

Van der Walt, J. L. & Steyn, H. S. (Jnr.). (2007). Pragmatic validation of a test of academic literacy at tertiary level. *Ensovoort,* 11(2):138-153.

Van Dyk, T. J. & Weideman, A. J. (2004a). Switching constructs: On the selection of an appropriate blueprint for academic literacy assessment. *Journal for Language Teaching,* 38(1):1-13.

Van Dyk, T. J. & Weideman, A. J. (2004b). Finding the right measure: from blueprint to specification to item type. *Journal for Language Teaching,* 38(1):15-24.

Van Dyk, T. J. (2007). Personal interview. 21 November, 2007. Stellenbosch.

Van Schalkwyk, S. (2008). Acquiring academic literacy: A case of first-year extended degree programme students at Stellenbosch University. Doctoral dissertation. Stellenbosch: University of Stellenbosch.

Wagner, E. (2002). Video listening tests: A pilot study. *Working Papers in TESOL & Applied Linguistics, Teachers College, Columbia University,* 2(1):1-39.

Wagner, E. (2007). Are they watching? Test-taker viewing behaviour during an L2 video listening test. *Language Learning & Technology,* 11(1):67-86.

Weideman A. (2003a). *Academic literacy: Prepare to learn*. Pretoria: van Schaik.

Weideman A. (2003b). Assessing and developing academic literacy. *Per linguam* 19(1 & 2):55-65.

Weideman A. (2006). Transparency and accountability in applied linguistics. *Southern African Linguistics and Applied Language Studies*, 24(1):71-86.

Weir, C. J. (1993). *Understanding and developing language tests*. Hertfordshire: Prentice Hall Europe.

Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave McMillan.

Widdowson, H. (1983). *Learning purpose and language use.* Oxford: Oxford University Press.

APPENDIX A

ACADEMIC LISTENING TEST QUESTIONNAIRE

For lecturers and students:

Please mark the appropriate boxes (you can choose more than one box) and provide additional comments in the 'specify' box if you wish to do so.

Task 1 is the 'Instructions' item, Task 2, the 'Psychology lecture', Task 3, the 'discussion on euthanasia' and Task 4, the seminar extract on 'Foreign Investment'.

1. The tasks are designed to simulate actual tasks that a first year university student would be called upon to perform. To what extent do the 4 test tasks reflect the target language use situation, i.e. a university campus?

	Totally relevant	Moderately represent-ative	Reflective in parts	Not really represent-ative	Specify
Task 1					
Task 2					
Task 3					
Task 4					

2. How interesting and relevant to university study did you find the test content?

	Very interesting and relevant	Moderately Interesting and relevant	Quite interesting but not necessarily relevant or visa versa	Neither interesting nor relevant	Specify
Task 1					
Task 2					
Task 3					
Task 4					

3. Is the level of difficulty appropriate for a first year university student?

	Correct level	Too difficult	Too easy	Some questions were too difficult/easy	Specify
Task 1					
Task 2					
Task 3					
Task 4					

4. In the multiple choice tasks, are the questions fair or are the distractors …?

	challenging but fair	too similar	confusing at times	very confusing	Specify
Task 1					
Task 2					
Task 3					
Task 4					

5. Is the order in which the tasks are placed in the test appropriate? If you feel they are not, please indicate in which order you would place them.

First task	Second task	Third task	Fourth task	Specify reasoning
Task __	Task__	Task__	Task__	

6. Is the computerized format of the test …?

easy to navigate	harder to use than a pencil-and-paper test	easier to use than a pencil-and-paper test	Specify reasoning

7. Comment on the quality of the sound and visuals in the multimedia presented.

Excellent	Good	Adequate	Poor	Specify reason

8. Comment on the length of the multimedia clips.

Appropriate length for the task	Some were too long (specify task)	Some could have been longer (specify task)	Some were too short (specify task)	Specify reason

9. Are the instructions ...?

	clear and unambiguous	quite clear	confusing at times	inadequate and confusing	Specify
Task 1					
Task 2					
Task 3					
Task 4					

10. Did you find that being able to listen to the 'instructions' audio clip only once (Task 1) was ...

adequate for the task	not enough to absorb all the information	completely insufficient	Specify reasoning

11. Did you find that a single screening of the Psychology lecture (Task 2) was ...?

sufficient	necessary for 'real-life' simulation but had difficulty noting all the points	completely inadequate	Specify reasoning

12. Did you find that being able to listen to the 'euthanasia discussion' audio clip only once (Task 3) was ...?

adequate for the task	not enough to absorb all the information	completely insufficient	Specify reasoning

13. What listening skills do you think are being measured in the various tasks?

	Direct meaning comprehen-sion skills	Inferred meaning comprehen-sion skills	Contributory factors such as stress & intonation	Listening and note-taking skills	Specify
Task 1					
Task 2					
Task 3					
Task 4					

14. Is the listening purpose e.g. main theme, specific detail, opinion, supporting arguments etc. made clear to candidates in the instructions?

	Very clear	Fairly clear	Not clear enough	Completely unclear	Specify
Task 1					
Task 2					
Task 3					
Task 4					

15. Are there factors in the test which could advantage/disadvantage any particular candidate? For example: subject matter of the extracts, accents of the speakers, vocabulary use, etc.

Definitely not	To some extent	Yes, some evidence of bias	Very biased	Specify reasoning

16. If there were any questions with which you had specific problems, your feedback would be greatly appreciated.

Feedback from lecturers only:

1. How relevant did you find the test material in terms of measuring academic literacy as a whole?

Very relevant	Mostly relevant	Only relevant to some aspects of academic literacy	Not relevant	Specify reasoning

2. Do you think the test would be reliable as a way of 'streaming' incoming first year students?

Very reliable	Mostly reliable	Not a very accurate indication of language ability	Not reliable at all	Specify reasoning

3. Would you have confidence in the test results as an accurate measure of academic listening ability?

Yes definitely	To some extent	Not in all the tasks	No, I wouldn't	Specify reasoning

4. What was your overall impression of the test? I would value your opinion on any issues which have not been covered in the questionnaire. Comments:

APPENDIX B

ACADEMIC LISTENING TEST (ALT)

Approx. 90 mins Marks: 100

Additional materials:

Paper and pen or pencil for note taking

Headphones are required for this test

Test Information

The test is divided into four sections. Each section will contain a recorded extract followed by questions in either a multiple choice or gap-fill format. Specific instructions will be given before each section regarding the type of response expected of a candidate as well as the steps that need to be taken to 'save' your task answers.

When you have completed the test and saved all your answers, click on 'finish'.

Thank you very much for taking part in this study.

Media quality check

Will your computer play the media files?

Play the movie clip below to test whether the computer's media playback is correctly configured and that the volume is set to the correct level. Can you see the movie? (the quality is a bit poor)

Does it play correctly? Can you hear the sound?

If the answer to any of the above is **NO**, get one of the computer lab's technical assistants to help you configure the computer, or log out and try a different computer.

If the answer to all of the above is **YES**, select **True** and click **Save and View Next**.

General Information and Disclaimer

This test is part of a research project for a master's thesis. The purpose of the test is to assess your listening abilities in a university context. The results of the test will be used purely for research purposes and will not be made public. Data will be referred to collectively and not individually, so no student's identity or student number will be disclosed in the thesis. No candidate will be disadvantaged by taking the test.

By ticking the box below you are declaring that you have read and understood the information given above and consent to participate in the study.

I hereby give my permission for data, resulting from this test, to be used for research purposes, and in particular to be compared with my TALL results. (Yes/No option given)

Task 1: Instructions

The recording you are about to hear is a lecturer giving his class important instructions. You may listen to the clip only **once**. As 'virtual' students in the class, you may need to take notes so as to remember the specific details. The notes will not form part of the assessment. A multiple choice exercise follows the recording. Select one of the 4 alternatives (a, b, c or d), in response to all the questions except number 2 where the procedure is different. The directions for question 2 are included in the exercise.

When you are ready, select 'True', then click on **Save and View Next** to listen to the audio clip. It will only play once and you will not be able to return to it, so listen carefully.

'INSTRUCTION' AUDIO CLIP

Multiple choice questions:

1. The lecturer mentions handouts at the beginning of the recording. Which of the following most closely echo his words?

   a) He expects everyone to have a handout.
   b) He believes everyone has a handout.
   c) He suspects that everyone has a handout.
   d) He has distributed the handouts. b

2. The lecturer recommends certain steps which must be taken in order to access course material.

   Put the following steps in the correct order by selecting the step you think comes first next to a), the second next to b) and so on.

   a) Click on Ling 2
   b) Go to the Departmental webpage

c) Scroll down to Language Change
d) Click on undergraduates b;d;a;c

3. Where would a student listening to these instructions expect to find the 'copies' that the lecturer refers to?

a) In Lisa's office.
b) In his office.
c) In the thesis office.
d) In Delise's office. d

4. The speaker emphasizes one particular part of his instructions to the class. Which of the following **best** describes his purpose for doing this?

a) He wants to make certain that everyone understands the instruction.
b) He wants to ensure that the students take responsibility for their actions.
c) He wants the students to know that the information is very important.
d) He wants to make sure everyone remembers this particular instruction. c

5. What is meant by the word 'mandatory', heard in the text? Choose the word that you think, in the context, is closest in meaning.

a) obligatory
b) necessary
c) required
d) recommended a

6. According to the lecturer some of the recommended reading material…

a) provides background knowledge for the more advanced students.
b) is too advanced for some of the students.
c) is only useful if you have no background knowledge.
d) is useful for additional background knowledge. d

7. What is your interpretation of the phrase 'well worth your while' used by the lecturer in the passage? Does it mean …?

a) very easy to understand
b) very useful indeed
c) very well known
d) very reliable information b

Task 2: Psychology mini-lecture

In this task you will be watching and listening to a video clip of a first year Psychology lecture. You will only be allowed a **single** viewing and are, therefore, strongly advised to take notes to help you remember the most important points made by the lecturer. Your notes will not form part of the assessment and are solely to help you answer the multiple choice questions that follow the video screening. In each of the multiple choice questions, choose only one of the 4 alternatives (a, b, c or d) by selecting the appropriate button.

Select 'True' when you are ready and click **Save and View Next** to watch the video. Please take note that you will not be allowed to pause or restart the video once it has begun. Once you have viewed the lecture extract you can continue to the questions which follow.

VIDEO CLIP

Multiple choice questions:

1. What is the main theme of the lecture? Is it to determine whether …
   a) there are separate brain functions for object recognition and facial recognition?
   b) there is a part of the brain which is specially adapted to recognizing objects?
   c) there is a special module of the brain which is dedicated to recognizing faces?
   d) object recognition and facial recognition occur in the same part of the brain? c

2. Which of the following statements matches the lecturer's stated opinion?
   a) The brain has a special part dedicated to facial recognition.
   b) The brain has a special part dedicated to neither facial nor object recognition.
   c) The brain does not have a special part dedicated to facial recognition.
   d) The brain has a special module dedicated to both facial and object recognition. a

3. Which of the following words is the closest in meaning to 'dissociation'?
   a) dissolute
   b) inseparable
   c) separate
   d) unique c

4. According to the lecturer, the term 'single dissociation' applies to …
   a) the study of an individual's ability to succeed in task A and task B.
   b) an individual with limited ability on task A and task B.
   c) the study of two individual's abilities on task A and task B.
   d) an individual with limited ability on task A but not task B. d

5. Individuals who are unable to recognize familiar faces but otherwise have normal visual abilities are said to suffer from:
   a) prosopanomia
   b) prosopamnesia
   c) prosopagnosia
   d) prosopopoeia c

6. What makes object recognition a relatively easy task?

   a) The components are similar to those of other objects.
   b) The components are easily identifiable.
   c) The components are associated with other objects.
   d) The components are easily confused. b

7. What area of the brain is damaged when a person is unable to recognize familiar faces?

   a) the occipito-temporal region
   b) the occipito-apraxia region
   c) the cortical-temporal region
   d) the occipito-parietal region a

8. What does it mean if there is no dissociation on two tasks?

   a) A person is able to recognize familiar faces but not general objects.
   b) A person is able to recognize objects more clearly than familiar faces.
   c) A person is able to recognize an everyday object but not a familiar face.
   d) A person is able to recognize general objects as well as familiar faces. d

9. Why is it sometimes difficult to distinguish between faces? Choose the response that is **closest** to the reason given by the lecturer.

   a) People have only minute differences in their facial components.
   b) Everyone has the same components which make up their facial features.
   c) People's features may look the same but everyone is unique.
   d) People's features are confusing because they all look exactly alike. b

10. Which of the following will José not be able to recognize?

    a) his dog
    b) his car
    c) his shoes
    d) his mother d

11. Which of the following is the minimum requirement for double dissociation?

    a) Two patients, each performing the same pair of tasks.
    b) One patient performing a single task.
    c) Two patients, each performing two different pairs of tasks.
    d) One patient performing two separate tasks. a

12. In the opinion of the lecturer, which of the following provides the strongest argument for facial recognition occurring in a particular part of the brain?

    a) multiple dissociation
    b) double dissociation
    c) single dissociation
    d) spatial dissociation b

Task 3: Euthanasia discussion

At university, students are encouraged to critically analyse and evaluate what they have been taught in their respective courses. Discussion with one's fellow students is a good way of doing this. In the following extract, you will hear two Law students, Tess and Danie, discussing the legal aspects of 'euthanasia'. Listen carefully to their opinions and supporting arguments (notes would be helpful) before answering the multiple choice questions that follow.
Take note that the audio clip will only play **once**. When you are ready, select 'True' and then click **Save and View Next** to start listening to the clip. When the clip is finished you can continue to the questions.

'DISCUSSION' AUDIO CLIP

Multiple choice questions:

1. Which of the following statements **best** describes Tess' justification of passive euthanasia?
   a) It's not considered murder because the patient is dying and there is no chance of him/her making a recovery.
   b) It's not considered murder if doctors have decided a patient cannot be kept alive for ever on life support machines.
   c) It's not considered murder because it's common practice for doctors to switch off life support machines on a daily basis.
   d) It's not considered murder because the patient's family have made the decision to switch off the life support machines. a

2. What is the main theme of the discussion? Is it whether …
   a) euthanasia is justified?
   b) active euthanasia is murder?
   c) euthanasia should be legalized?
   d) suicide should be considered a crime? c

3. Decide whether the following opinions are expressed by only one of the speakers or whether they agree. Type T for Tessa, D for Danie or B for both, in the boxes provided.
   The doctor who overdosed his patient with morphine was justified in his actions. **D**
   An individual's rights are infringed by not allowing him to seek assistance in his suicide. **B**
   The 'right to life' does not mean the same as the 'right to death'. **T**
   A person should not be able to kill someone who is suffering. **T**

4. What is Danie's supporting argument for 'assisted suicide'?
   a) A person should have the freedom of choice to decide their own future.
   b) It is an individual's right to die with dignity.
   c) A person is entitled to ask for help if they are unable to end their own life.
   d) Suicide is not a criminal offence in South Africa. d

5. Tess argues that legalizing euthanasia would …:

a) open the floodgates for international criticism.
b) allow people to give up too easily.
c) encourage people to make rash decisions.
d) make a mockery of the right to life. b

6. When Danie suggests putting someone 'out of their misery', what does he mean?

   Choose the option you think most closely matches his sentiments.

   a) The person is so miserable that it will be a blessing not to have him/her around.
   b) The person is in so much pain that his life is not worth living.
   c) A person living an undignified life should at least be allowed a dignified death.
   d) The person is going to die soon anyway, so why prolong the agony. d

7. Why does Tess point to the fact that 'we are all human'?

   a) To support her argument that doctors make mistakes.
   b) To show that circumstances change and one should never give up hope.
   c) To justify the fact that we all make mistakes at some time.
   d) To show that the will to live is an integral part of the human spirit. a

8. Danie makes a statement that strong medical evidence is necessary before decisions can be made on the legalization of euthanasia. Tess shows her agreement by …:

   a) adding an extra piece of information to support his statement.
   b) reinforcing Danie's point by rephrasing what he has just said.
   c) quoting from the South African Law Commission.
   d) stating that the doctor and the patient's family should be in agreement. b

9. Reference is made to the Hartman case where a doctor is found guilty of assisted suicide but not sent to prison. This leads Danie to the assumption that …

   a) South Africa is about to follow the Netherlands' example of legalizing euthanasia.
   b) the court concluded that the doctor's actions were justified.
   c) legal opinion is shifting in favour of the legalization of euthanasia.
   d) the court didn't consider the doctor a threat to society. c

Task 4: 'Foreign Direct Investment' tutorial extract

In the following audio clip you will be listening to a senior lecturer in Business Science speaking about Foreign Direct Investment. You may listen to the audio clip **twice**. Before listening to the recording, select 'True' and click on **Save and View Next** and read through the gap-fill exercise (you will not be able to fill in the words) so as to get an idea of the central theme and structure of the text. You will notice it is a **summary** and not a transcription of what you hear.

When you are ready, select 'True' and click **Save and View Next** to listen to the audio (the text will re-appear on your screen). While listening to the extract for the first time, read the text

again and once again you will not be able to start typing in the words. As before, when you are ready, select 'True' and click on **Save and View Next** to complete the exercise.

The audio clip will play for a second time and you can type in the missing words. The words that come before and after the gap, will indicate the type of word that has been omitted, so make sure your responses fit into the grammatical context of the sentence. Words must be spelt correctly so that the computer is able to recognize them. A list of words is provided alongside the text which can be used to check the spelling of your answers.

Words:

benchmark	deciding	governance	managerial
benefit	determining	government	multinationals
budget	domestic	important	overachiever
capability	dominant	industry	profit
capacity	economy	internal	revenue
challenges	efficiency	infrastructure	shareholders
competence	entrée	investors	talent
control	experience	labour	threshold
corporations	expertise	local	underperformer
deals	foothold	management	ventures

'HELENA BARNARD' AUDIO CLIP (played 2x)

Gap-fill exercise:

Foreign direct investment is often considered to be merely a source of capital for a country but there are more complex issues which need to be considered. One of the primary considerations is the way in which **1** _____ are structured. It is possible for a change of **2** _____ to take place without money coming into the country. Therefore, what is of real significance is **3** _____ control. Foreign investors, with their accompanying new technologies, practices and training strategies can result in improved performance, independent of capital funds.

South Africa is regarded as a 'chronic **4**_____' when it comes to FDI. An explanation for this could be that there is a very strong **5** _____ sector in South Africa and foreign firms find it difficult to get a **6** _____. Generally, however, the inflow of technologies and managerial **7** _____is considered to be beneficial to South African industry.

An aspect of multinational investment is that there is a **8** _____ effect. If a firm is below this, then it is below the basic level of **9** _____. This can result in the firm

being forced to close due to the arrival of foreign firms. This clearly has a very negative effect on a country's local **10 _____**. However, firms that are above the threshold, **11 _____** from the challenges presented by the arrival of foreign firms. It provides an incentive to improve their goods and services as well as an opportunity to learn from the **12 _____**. Therefore, when considering the advantages and disadvantages of foreign direct investment**,** the **13 _____** factor is the situation at home.

In conclusion, it is clear that governments and multinationals are motivated by very different goals. Multinationals, aim to make money and provide their **14 _____** with good returns. Governments, on the other hand, have a duty to their citizens. However, both **15 _____** and **16_____** play important roles in ensuring that South African firms remain above the threshold level where foreign direct investment has strong beneficial effects.

MEMO

1. deals	2. control	3. managerial	4. underperformer
5. domestic	6. foothold	7. expertise	8. threshold
9. competence	10. capacity / infrastructure	11. benefit	12. multinationals
13. deciding/determining		14. shareholders/investors	
15. infrastructure/government/governance		16. governance/government/infrastructure	

APPENDIX C

TRANSCRIPTION OF TASK 1: 'INSTRUCTIONS'

You all have a handout I presume, okay, good. I will also put the (um) overheads for today on the Internet, as I put (the) a sub-portion of last week's overheads on the Internet. You can find it if you go to the Departmental webpage. Then you click on undergraduates and then you'll click on, there's a big picture there and you scroll down, underneath the picture and then it says Ling 1 Ling 2 Ling 3, right, so you click Ling 2. So and then Ling 1 Ling 2 Ling 3 and er click Ling 2 then, there's a whole lot of courses and I think towards the bottom of that list there is Language Change and then you should see a link there saying week one week two. Yes, I'd prefer not to do that because then I can't take them out either, okay um but I have made copies and put them in Delise's office in a file. So you go there, sign out the file, make your copies and bring it back okay (um, ja), because if, obviously if you don't bring the file back then the book's out the library then everyone's gonna suffer, okay? Not all the readings are mandatory um some of the readings are definitely well worth your while others are maybe a little bit more advanced or to give you a bit more background knowledge (um) and that I'll indicate using the tuts I'll give you a little thing that says which readings are good better best, okay.

APPENDIX D

TRANSCRIPTION OF TASK 2: 'PSYCHOLOGY LECTURE EXTRACT'

So today, we'll be exploring, one of the things we'll be exploring, is 'does the brain have a special module for facial recognition'? Okay? Is there a special part in the brain that is totally committed and only committed to facial recognition? Or, is facial recognition, is that function captured in the same part of the brain as general object recognition? Okay? What do you guys think, what does your gut feeling tell you? Do you think that there should be a special part of the brain dedicated to facial recognition? Or do you think that it's all just part of the general object recognition function? Special or not special? Special, okay that's what my gut feeling tells me.

So there's an important word that you need to know when we talk about whether the brain has a special module for facial recognition, this word being dissociation, okay? Previous classes have struggled with this concept so let me clarify it for you: dissociated means unrelated, not related to one another. Okay? So when there are differing levels of performance between tasks, when you're better on one task than on another task the functioning or the performance that you need to undertake for that task, okay, the function required for that task, can be said to be dissociated from the function required for another task. It appears that the two tasks functioning have nothing to do with one another. That they're separate, so they are dissociated. Okay? When things are associated it means that they have something to do with one another. Here we're talking in terms of dissociation, to what extent do they have nothing to do with one another? So, again dissociation means that there are differing levels of performance between tasks. Single dissociation, we talk of one individual performing on both tasks, so, one individual has impaired performance on task one but not on task two. So for this one individual it appears that the functions required for each task are not related therefore, it is dissociated: because it's one person we talk about we talk about single dissociation. So if a person with prosopagnosia can recognise common objects but not faces, now we know they can't recognise faces because they've got prosopagnosia, but if they can recognise objects, common objects being everyday objects, then we can say that recognising common objects and recognising faces, those two functions are dissociated, they have nothing to do with one another. Okay?

So, an individual is asked to recognise a shape and they can do that successfully however, they can't recognise a face. Okay? So they know they looking at a face they just can't

recognise it. Now whether it's a face of a family member, or whether it's the face of someone so famous that you'd expect your average individual from a given population, like student population to recognise it, they can't recognise the faces might be indicative of prosopagnosia.

Now what happens if these two tasks actually have different difficulty levels? As I suggest here this would undermine the theory that faces are special. What happens if performing one task is just inherently more difficult than performing a different task? Now when you have different levels of performance can one say that they are necessarily dissociated? Maybe they not dissociated maybe it's got nothing to do with whether the functions are related to one another or not. Maybe it's all explained by the fact that the two functions have different levels of difficulty: that one is just more difficult to perform than the other. So this is what Damasio and colleagues argue, they argue that distinguishing between common objects, everyday objects, so distinguishing between a shoe and a dog as an example, is easy because those objects are made up of different components, it's very difficult to confuse them. Shoes have toe caps, dogs have toes and pasterns. Okay? Shoes have vamps and dogs have ribs, you can't possibly confuse the two. You don't look at your dog and want to put it on your foot to wear it as a shoe. Okay? You don't look at your trainers and go 'ah good boy, good boy'. Okay? It's very difficult to confuse them. So, what Damasio and colleagues suggest is that distinguishing between everyday objects requires a super-ordinate level of recognition. All you need to know is what are the components that make up objects from this category, okay so dog has a tail, its got fur, it barks, you know that's important because cats also have tails and fur, um and shoes, shoes have soles, they have eyelets and if you didn't know it they have vamps. Okay, it's a new thing that I learnt, I can't look at my shoes in the same way anymore. Okay?

Faces, on the other hand, they argue, they are a lot more difficult to distinguish from one another. Okay? Why? Because they are visually confusing, they have the same components. Everybody has two eyes, a nose, two ears and a mouth and for everybody these components are in exactly the same place relative to one another. Okay? However, for each person's face there is only one example of that face in the entire world. Even if you have a twin, okay an identical twin, your faces might look the same but depending on the level of sophistication of your er measure where you're comparing the two faces you will find miniscule differences. So for each face there is only one example of that face in the entire world and you have to distinguish that from six billion other faces, or maybe you don't know six billion other people but at the very least you have to be able to distinguish that from thousands of other faces that you've seen before in your lifetime. So therefore, faces are more unique and more difficult to distinguish from one another. So if a person with prosopagnosia demonstrated no

dissociation, take note, demonstrates no dissociation of two tasks, on two tasks that have equal difficulty, what does that mean? This individual can't see faces or they can see faces, they can't recognise faces, they can't distinguish particular faces. Okay? Now they perform another task which is equally difficult for them and there is no dissociation. What does that mean? Pardon? Anybody? There's no dissociation, if there's no dissociation it means there is association, which means equally difficult, which means they also couldn't recognise this other object, this common object 'cause you know that they can't recognise faces. So, if they perform equally bad on a separate task, like recognising a common object, it means they also couldn't recognise this object and if that's the case, then this suggests that we don't have a special module in the brain for facial recognition. Do you all understand this? Are you all with me? Is there anybody who's not? Do you want me to go over it again? Okay, you're all with me, great.

So, for example, if this individual struggles to identify this shape and also, because they've got prosopagnosia, could not identify this face, then one can say well maybe both functions are actually related to one another. So, maybe facial recognition and object recognition happens in the same part of the brain because we know that for prosopagnosia a particular part of the brain, the enfero, no I'm lying to you, the occipital-temporal region must be damaged, for prosopagnosia. So we know this area is damaged that's why they can't recognise faces but now they also can't recognise objects, so it suggests that maybe recognising objects happens in that exact same part of the brain.

But now we come to double dissociation, it's an alternative way of asking the question, can face, is facial recognition located in a particular module of the brain? Now with double dissociation, we looking at, at least at a minimum two individuals and we comparing their performance on two separate tasks. So here we have farmer Joe and farmer Joe he can distinguish between faces. Okay? So when he looks at his family he knows just by looking at them he can tell you who's his wife and who's his brother and who's his cousin, who are his kids et cetera et cetera. Okay? He does not need other cues to give him that information. However, when you ask him to identify cows, you show him a cow, he's very confused, it has no real meaning for him, he can't identify the cows. If you tell him he's looking at a cow and he knows, he's learnt well cows are milked then it, then it's fine but just by looking at it, if you just show him a picture, it's confusing. The moment the animal moves, all of a sudden he goes 'okay these are cows'. Alright? So, on two tasks, the one is general object recognition being recognising the cow, the other one is facial recognition. He cannot recognise cows but he can recognise faces. Okay? Then we have Hosè and Pedro, now Hosè here is very confused because he looks at Pedro but he doesn't know he's looking at Pedro, he's got prosopagnosia.

Now, until Pedro starts talking, Hosè's going to be very confused as to who he's looking at. Okay? However, Hosè can recognise cows, if you had to ask him 'show me the cow' he'd point to a cow but if you had to ask him 'point to Pedro' and there were a couple of different people in front of him, he wouldn't be able to point to Pedro. Okay? So, on these two tasks he can recognise the common object but he cannot recognise the face. So, person A, farmer Joe, cannot recognise cows but can recognise faces, person B, Hosè, can recognise cows but cannot recognise faces. This seems to suggest that the two functions are dissociated from one another and it's a, it's a, it's a stronger argument that facial recognition happens in a separate part of the brain to object recognition. This double dissociation is a stronger argument than the single dissociation argument.

APPENDIX E

TRANSCRIPTION OF TASK 3: 'EUTHANASIA DISCUSSION'

Tess: Well passive is where we, you put the machine off, when a person's dying, so, that's not considered murder in the country, I mean that's done every day. Doctors switch machines off when they, if people are on life support.

Danie: and without the machines they won't be able to survive anyway …

Tess: … live anyway, ja, so (um) I think what, what we're talking about today is more active euthanasia so that's basically like in the case, S versus Hartman, where a doctor actually put morphine, an overdose of morphine, in the guy's, (um) in his drip so that basically, it was an active form of killing him.

Danie: 'cause the patient was suffering and …

Tess: He was, but, it is still considered murder.

Danie: But, then you think again, (um) a person (um) like suicide in South Africa isn't seen as a crime but what about a person that (um) isn't capable of (of) committing suicide. Like you infringe against his rights by not allowing him to get assistance from someone else to commit suicide.

Tess: I s'pose you are (you are) going against his right to decide his own sort of future but in our constitution we have the right to life which was said in UK versus Pretty, um the (united) United Kingdom case, that the right to life doesn't mean the right to death. So, if South Africa had to legalize you giving consent to your own death, I think it would open the floodgates to, to people just allowing themselves to ja to die and I don't think that's … I think that would, there has to be a line drawn.

Danie: But, you have to (um) look at their, their basic human rights, right to dignity, freedom of choice.

Tess: I think the one that tops all is, right to life, don't you think.

Danie: Ja but (um) a person that is suffering and in (um) it is, it's obvious that he he's going to die anyway why do you keep him alive? Why don't you put him out of his misery? (Tess: well, we're only human) Let him die in dignity and …

Tess: But don't you think we're only human and (and, and) how often have doctors made mistakes about terminal ill patients and how much longer they have to live. (um) If you just, if you had to allow euthanasia and say a doctor says it's, a person is terminally ill and he's made a mistake and this person decides to you know give the consent to die, but actually he would have had a healthy life in front of him.

Danie: That's why you need (um) medical evidence like (it) it must be, it can't be just a, a light decision they …

Tess: … must support it quite heavily (reformulation) Well, in the report by the South African Law Commission, on euthanasia and the artificial preservation of life, (um) they believe that the, the legal standing as it is today should (should) stay, like as far as active euthanasia is concerned (um), that it should still be punishable. But they did (um), they did qualify it, they said that if a medical practitioner (um) does the, does the euthanasia, and that the family must support it, then it should be legalized. So what do you think?

Danie: As you said, in the Hartman case, the (the) guy was found guilty of murder (Tess: ja) by er helping his dad (um) like ...

Tess: ja he (he) put gave him an overdose of morphine

Danie: Ja (um), he was found guilty of murder but there wasn't any punishment so it looks like the (the) court is is leaning towards legalizing euthanasia (Tess: ja) and it's and countries like Netherlands they have already done that. (Tess: ja) I know I just think South Africa would follow up on that.

Tess: … follow on that (um) ja, well I believe euthanasia could be okay but I think it would have to be very strictly, there'd have to be very strict regulations definitely. Like if you just legalized (um) you know euthanasia in terms of putting someone out of their misery you know I think you'd have to do define all everything that has to do with it, otherwise you could just as easily open the floodgates. And as was said in R versus McCoy (um), if a person kills another, maybe (um), in sympathy towards a person suffering, doesn't that sort of allow that person to cross the threshold and it's easier to do it the second time round.

Danie: Ja, I agree that you can't take like law into your own hands, there has to be sufficient medical like (Tess: back-up) support and back-up and but I still believe if a person is, is suffering and his (his) chances of (of, of) living a happy and like healthy life aren't very great then, then I think it would be best ...

Tess: Put him out of his misery

Danie: Yes

Tess: Ja, (um) you know in Law, there's always the public policy element to take into consideration and um I don't know if euthanasia, is … do you think it's in, in the best interests of society to, to have it legal?

Danie: (um) in certain circumstances (um) I just don't see how the society would (um) be in favour of a person suffering.

APPENDIX F

TRANSCRIPTION OF TASK 4: 'FOREIGN DIRECT INVESTMENT TUTORIAL'

We tend to think of foreign direct investment (as) as something that brings capital into a country but in fact (um), it's far more complicated than that. First of all, (the) the way deals could be structured, you could have a change (of) of control, without actually having money flowing into the country. (um) What really matters, is the managerial control, because the moment you have, let's say, a Swedish or a Danish or (a) American firm (um) taking over a South African firm, they are going to bring along their latest technologies, their latest practices. They're going to make sure, want to make sure, that the South African operations are as competitive as they could be and (to) to achieve that, they need (to) to invest a lot in training and (um) just make expectations, high expectations, (make) make that clear to the local firms. So what we tend to see is that (um), performance improves, independent of the capital funds, (er, the the) the relevant capital flow.

(um) There's a whole debate because South Africa's a chronic underperformer in terms of FDI and (um) there's a guy at Harvard, whose name escapes me for the moment, who said that perhaps the (the) problem is a very strong domestic sector and because (of) local firms are so strong, it's very hard for a firm to (to to) get a foothold. (um) Obviously, (the) the closer (your your) your local firms are to (to) the level of the foreign firm, the less learning there'll be. But, by and large, there (there) quite a large number of South African industries (where) where (the) the inflow of technologies and managerial expertise are (are) really helpful.

(um) The point about (um) multinational investment is, there's a threshold effect and if you're above the threshold, life looks very different to if you're below the threshold. If you lack (um) firms of a basic level of competence, what's going to happen is that the entry of foreign firms will simply wipe them out and you don't want to decimate your local capacity. However, if they above that threshold level, what you typically find is that the local firms are challenged by the entry of a foreign firm and they either (um) have to jack up their offering to make sure that they don't lose customers or they could sort of steal with their eyes, see what the multinational is bringing along and (and) introduce that in their own firms. So, what we see is that (um that) above that threshold level you actually have very strong beneficial (er) effects. (so) So really (um), both the anti-globalization activists and the (the) governments are right (um) but it's not the fault of the multinationals. (the the) The determining factor is at home and what's actually happening at home.

Multinationals and governments have very different goals. Multinationals want to make money and they want to give a good return to their shareholders. Governments have a responsibility to their citizens. So, in some cases, (you you) you have imperialist behaviour (er) but you don't always have it and certainly multinationals are (um) sort of a-political in a sense (that) that as long as they can make money (they) they perfectly happy to get in bed with (with) just about anybody.

I think what we really need to do is to focus on South African firms, South African infrastructure, (er) governance (um). Multinationals will be good for South Africa if our own firms and infrastructure is above that threshold level. So, rather than try (and) and shape the behaviour of multinationals, let's get our own house clean.

Source: http://www.gibs.co.za//home.asp?pid=2166