Transparency, accessibility and accountability as regulative conditions for a postgraduate test of academic literacy

Avasha Rambiritch

Submitted in fulfilment of the requirement for the degree
D. Phil. Linguistics
In the Faculty of Humanities,
Department of English at the University of the Free State

Supervisor: Prof. A.J. Weideman
Co-Supervisor: Dr. S. Brokensha

January 2012

Table of Contents

List of tables

List of figures

Appendices

Acknowledgements

My sincere gratitude to the following:

- My supervisor Prof. Albert Weideman – for introducing me to the world of language testing. And for showing me how important it is to be passionate about the work we do. Thank you for your mentorship and your guidance. It has been an honour working with you.

- My co-supervisor Dr. Susan Brokensha – for the expert advice and kind words.

- Jurie Geldenhuys for taking the time to edit this document.

- My husband Anesh for the many hours you spent helping me complete this study. For your love and support and for encouraging me to pursue my dreams.

- My son Vibhav – the light of my life. And for the joy you bring into it.

- My parents Vasant and Thara Rambiritch for your encouragement and for believing that I could. Thank you for making this journey possible.

- My brother Shikar, for lending his expertise in the design of the TALPS web page.

- My colleagues at the Unit for Academic Literacy, University of Pretoria for their support in the face of so many challenges.

- God Almighty, for giving me the opportunity to pursue this study, the strength to ensure that I could and the courage to make sure that I did.

Abstract

This study is concerned with transparency, accessibility and accountability as regulative conditions for a postgraduate test of academic literacy. What it will propose to do is investigate how these can be incorporated into the design of one test, the Test of Academic Literacy for Postgraduate Students (TALPS), and theoretically accounted for in terms of a framework.

A main focus is to show that the questions raised here about the social dimension of language testing cannot be adequately answered by experts in the field like Messick (1989b; 1996), Bachman and Palmer (1996), and Kunnan (2000; 2004). Instead these questions can be answered in a "third idea, other than validity and usefulness" (Weideman 2009a: 239), as outlined by Weideman, an idea that does not foreground one concept but rather identifies a number of fundamental considerations for language testing. The argument here is that construct and other empirically based forms of validity are not enough to validate a language test and that what is needed, in addition, is a detailed look at issues of transparency, accessibility and accountability.

This study begins by contextualising the problem of poor academic literacy and outlining the need for academic literacy tests such as the Test of Academic Literacy Levels (TALL) and TALPS. This is followed by an in-depth study of previous work in the field of language testing. The literature on key concepts such as validity, reliability, accessibility, transparency and accountability is surveyed as well. An important part of this study is telling the story of TALPS from its initial conceptualisation to its final implementation. Included in this is a detailed study of the reliability and validity of the test, taking the form of a validation argument.

Subsequent chapters (5, 6 and 7) focus specifically on issues of transparency, accessibility and accountability as they relate to TALPS. This study would not be complete without the voices of the test takers. A detailed summary of the data collected from a questionnaire administered to students who wrote TALPS is offered as well. The questionnaire has been designed to elicit information, comments, questions and reactions from the testees about the test.

The final chapter in this study will attempt to provide a summary of the answers to the important questions that have been asked and answered in the course of this investigation. It will also consider the link between transparency, accessibility and accountability, and will focus briefly on other conditions in the framework that contribute to the design of fair and socially acceptable tests.

This study hopes to make a contribution to the field of language testing by concentrating on an area of testing that has been largely ignored – the social dimension. One of the aims of this study is to show the complementarity among the empirical, social and ethical dimensions of TALPS. It therefore provides a framework that incorporates a concern for the empirical analyses of a test as well as a concern for the social dimensions of language testing. Test developers are challenged to consider important questions related to every aspect of the test, leading to the design of fair, accessible tests that are designed by test developers who are willing to be accountable for their designs.

Key terms: academic literacy, testing, transparency, accessibility, accountability, constitutive, regulative, validity, construct, framework.

Samevatting

Hierdie studie handel oor deursigtigheid, toeganklikheid en verantwoordbaarheid as regulatiewe voorwaardes vir ’n nagraadse toets van akademiese geletterdheid. Dit poog om ondersoek te doen na die wyse waarop hierdie voorwaardes geïnkorporeer kan word in die ontwerp van een toets, die *Test of Academic Literacy for Postgraduate Students* (TALPS).

’n Belangrike fokus is om aan te toon dat die vrae wat omtrent die sosiale dimensie van taaltoetsing geopper word nie voldoende beantwoord word deur gesaghebbendes in die veld soos Messick (1989b; 1996), Bachman en Palmer (1996) en Kunnan (2000; 2004) nie. In plaas daarvan kan hierdie vrae beantwoord word deur middel van ’n “third idea, other than validity and usefulness” (Weideman 2009a: 239) – ’n idee wat volgens Weideman nie een konsep uitsonder nie, maar eerder ’n aantal fundamentele oorwegings vir taaltoetsing identifiseer. Die argument in hierdie verband is dat nie slegs konstruk en ander empiries-gebaseerde vorme van geldigheid voldoende is om ’n taaltoets geldig te verklaar nie, maar aanvullend daartoe ook ’n gedetailleerde oorweging van kwessies van deursigtigheid, toeganklikheid en verantwoordbaarheid.

Die studie begin deur die probleem van gebrekkige akademiese geletterdheid te kontekstualiseer en die noodsaaklikheid van akademiese geletterdheidstoetse, soos die *Test of Academic Literacy Levels* (TALL) en TALPS, uit te lig. Daarna volg ’n dieptestudie van werk wat reeds op die terrein van taaltoetsing gedoen is. Daar word ook ’n oorsig gegee van die literatuur oor kernkonsepte soos geldigheid, betroubaarheid, toeganklikheid, deursigtigheid en verantwoordbaarheid. ’n Belangrike deel van die studie is om die ontwikkeling van die TALPS vanaf die konseptualiseringstadium tot die uiteindelike implementering te skets. Hierby ingesluit is ’n omvattende studie van die

betroubaarheid en geldigheid van die toets. Dit is in die vorm van ’n bevestigende argument vervat.

Die daaropvolgende hoofstukke (5, 6 en 7) fokus spesifiek op die kwessies van deursigtigheid, toeganklikheid en verantwoordbaarheid met betrekking tot die TALPS. Die studie sou egter nie volledig wees sonder die insette van die toetsafleggers nie. Daarom is ’n vraelys ontwerp met die doel om hul inligting, kommentaar, vrae en reaksies omtrent die TALPS te verkry. ’n Gedetailleerde opsomming van hierdie gegewens word in die studie ingesluit.

Die slothoofstuk poog om ’n opsomming te gee van die antwoorde op die belangrike vrae wat in die loop van die ondersoek gevra en beantwoord is. Dit beskou ook die verband tussen deursigtigheid, toeganklikheid en verantwoordbaarheid en gee kortliks aandag aan ander voorwaardes in die raamwerk wat bydra tot die ontwerp van regverdige en sosiaal-aanvaarbare toetse.

Daar word gehoop om met hierdie studie ’n bydrae tot die veld van taaltoetsing te lewer deur te konsentreer op ’n aspek van toetsing wat nog grootliks geïgnoreer is – die sosiale dimensie. Een van die doelwitte is om aan te toon hoe die empiriese, sosiale en etiese dimensies van die TALPS mekaar komplementeer. Dit verskaf dus ’n raamwerk wat die noodsaak van sowel die empiriese ontledings van ’n toets as die sosiale dimensies van taaltoetsing inkorporeer. Die opstellers van taaltoetse word uitgedaag om belangrike kwessies in verband met elke aspek van die toets te oorweeg, sodat dit sal lei tot regverdige, toeganklike toetse waarvoor hulle verantwoordbaarheid sal aanvaar.

Sleutelbegrippe: akademiese geletterdheid, toetsing, deursigtigheid, toeganklikheid, verantwoordbaarheid, konstitutief, regulatief, geldigheid, konstruk, raamwerk.

Chapter 1

Language, learning and higher education

1.1 Introduction

In his foreword to the National Plan for Higher Education, Minister Kader Asmal wrote:

> The victory over the apartheid state in 1994 set policy makers in all spheres of public life the mammoth task of overhauling the social, political, economic and cultural institutions of South Africa to bring them in line with the imperatives of a new democratic order (Ministry of Education, 2001).

Of paramount importance in the new democracy was the transformation of the higher education system, the visions for which were articulated in the Education White Paper 3: A Programme for the Transformation of Higher Education (Department of Education, 1997), its main aim being "the establishment of a single, national co-ordinated system, which would meet the learning needs of our citizens and the reconstruction and development needs of our society and economy" (Department of Education, 1997). This "new democratic order" meant that since 1994, tertiary institutions have had to deal with the issue of accepting students whose language proficiency may be at levels that would place them at risk, leading to low pass rates and poor performance. This is a problem not specific only to students from previously disadvantaged backgrounds. Language proficiency is low even amongst students whose first language is English and Afrikaans, which are still the main languages of teaching and learning at tertiary level. Low levels of proficiency in English generally means that students are not equipped to deal with the kind of language they encounter at tertiary level. For many students academic language becomes a third or even fourth language.

Van Dyk (2005: 39) outlines three reasons for the low levels of academic literacy and poor pass rates. He states that the first of these is that the political

history of segregation and subsequent unequal distribution of resources in the South African educational system has negatively affected a large group of students referred to as historically disadvantaged students. The second reason for low levels of academic literacy may be that the South African educational system was a syllabus-driven (positivistic) approach. The third reason is that an increasing number of university students choose to study in English, which is not always their first language.

The reasons articulated by Van Dyk are confirmed in the discussions of other researchers in the field like Butler and Van Dyk (2004: 1), Webb (2002: 53) and Van Rensburg and Weideman (2002: 157). While the first and second reasons discussed by Van Dyk are part of a bigger picture and beyond the control of students, teachers and parents, one is led to question why students would choose to be educated in a language that hinders their academic success. One may agree with Webb when he states that “language is fundamental in academic training, and is either a facilitator to academic development or a barrier” (Webb, 2002: 53). Unfortunately, research seems to indicate that in many countries, language has become a barrier to academic success in higher education (Van Dyk, 2010:4). To deal with this, the ministry of education has requested the development of an “appropriate language policy framework” (see the National Plan for Higher Education, Ministry of Education, 2001), the result of which saw eleven languages, nine African languages, as well as English and Afrikaans, being recognised as official languages.

This language policy has led to some pertinent questions being asked, the most important being the practicality of a student being taught in the language of his or her choice, especially considering that the country now has eleven official languages. Other important questions deal with the cost that a policy like this would incur, as well as the fact that there are not many teachers/lecturers proficient in as many languages as this will require. Barry’s statement that the vision of the ANC to redress inequities of the past by offering a language policy

of this nature “is likely to remain a symbolic gesture in the foreseeable future” rings true (Barry, 2002: 105).

In keeping with the call for institutions of higher education to formulate language policies for teaching and learning, many universities have chosen the route of a dual medium of instruction. This means that in almost all cases the choice is between English and one other African language while in previously Afrikaans medium institutions like the University of Pretoria the medium became either English or Afrikaans. An important question raised by Webb pertains to the use of the two languages chosen by an institution as the languages of choice for teaching and learning. He asks whether courses are taught “in parallel fashion” (Webb, 2002: 57) in the two languages or are the two languages “used alternatively (or mixed) in dual medium fashion?” (2002: 58). In the case of the University of Pretoria, students choosing English or Afrikaans as the language of learning and teaching attend classes in that particular language. In many cases students can also choose to write the exam in one of these two languages (English or Afrikaans). However, despite a language policy being in place, resources, funding, staffing and large student numbers dictate that very often students have to attend a course in English, even if it is not the language they would choose as the language of learning and teaching.

An important question raised above but not yet answered is why students would choose to be educated in a language that hinders their academic success. Barry states that research has revealed that language and achievement are

> inextricably linked and the use of English as the language of learning and teaching by the majority of second language learners in South African schools should be seen as a major contributor to the poor pass rates and dropout rates of learners throughout the education system (Barry, 2002: 106).

In discussing this very issue, Van Rensburg and Weideman (2002: 157) ask “Why is instruction in the mother tongue so unpopular?” Their answer to this is the fact that “many parents are persuaded – and are probably correct – in

believing that English is the most important language of opportunity for their children" (2002: 157). They also observe that while parents may in this respect be correct "in selecting a strategy to have their children learn English, they demonstrably take the worst route, namely to choose English as the language of instruction from as early a grade as possible" (2002: 157). Barry notes in this regard that "English dominates the educational landscape in South Africa" (2002: 108) and that it is obvious that we are moving towards "a monolingual society" (2002: 108). The harsh reality is that students opt for English as the language of learning and teaching despite potentially low levels of proficiency in the language. They see proficiency in English as their ticket to the international world, arguing that it is the language that dominates the professional and business world: "Students realise that a high level of language proficiency is essential for successful participation within the global village and that technology has opened new contexts with wide ranges of purposes" (Barry, 2002: 108). This is a far cry from a time when parents were too afraid to allow their children to be educated in English for fear that it would lead to a loss of their indigenous language and culture. According to Barry, the trend in education today is that black parents are sending their children to previously white schools where the language of teaching is English, insisting that their children be taught in English from Grade 1 (2002: 108). However, there is little doubt that this insistence on the part of students and parents to opt for English as the language of learning and teaching has a detrimental effect on students' academic development and performance, leading to poor pass rates.

1.2 Language and learning in South African tertiary institutions

Very clearly therefore, language has remained a contentious issue in South Africa. The trauma of Bantu education still reverberates through the country. The effects will, no doubt, be felt for many years to come. The democratic attempt to right the wrongs of the past has made great strides in many areas, but has also created new challenges for which solutions need to be found:

> In sum, the legacy of the past was a fractured system and a set of HEIs [Higher Education Institutions] bearing the scars of their origins. As South Africa entered a process of social, economic and political reconstruction in 1994, it was clear that mere reform of certain aspects of higher education would not suffice to meet the challenges of a democratic country aiming to take its place in the world. Rather, a comprehensive transformation of higher education was required, marking a fundamental departure from the socio-political foundations of the previous regime (CHE, 2004: 230).

One of these challenges remains the issue of language and learning. The actual state of mother tongue teaching is far from clear but for many students who have been taught in their mother tongue, this is their first experience at being taught in English. Tertiary institutions, especially those considered previously advantaged, today need contingency measures to deal with this situation. Not accepting these students because of poor language proficiency would have simply been a repetition of the past.

The trend has been to set up specific programmes to assist these students. Different institutions have, however, taken different routes. Some have set up academic support programmes, department and units, while others have offered degrees and diplomas on an extended programme system, where the programme is extended by a year to ensure that the relevant academic support is provided. The academic support tends to concentrate on language proficiency, computer literacy and/or mathematics literacy. The Unit for Academic Literacy Departmental Self-evaluation (Unit for Academic Literacy, 2007: 4) points out that institutions in South Africa either set up discipline-specific development programmes, such as the University of Pretoria Foundation Year (UPFY), which was dedicated to increasing access for previously disadvantaged students in the natural sciences, or they would target critically important areas of ability that were known to cause concern (Unit for Academic Literacy, Departmental Self-evaluation, 2007). This same report points out that the former kind of intervention had the disadvantage of being as expensive as first-generation academic development approaches, as well as being equally unsustainable, since they remained dependent on external funding (Unit for Academic Literacy, Departmental Self-evaluation, 2007). Targeting specific important areas was

seen as a more viable solution. The University of Pretoria has therefore offered support in language and computer literacy. Today, the whole range of solutions, stretching from the general to the specific, is often combined, maximising their respective strengths (Unit for Academic Literacy, Departmental Self-evaluation, 2007).

1.3 TALL and the Unit for Academic Literacy

When the Unit for Academic Literacy (UAL) was established in 1998 as the Unit for Language Skills Development, there was already a concern about the high failure rate and "lower than acceptable levels of both computer and academic literacy" (Unit for Academic Literacy, Departmental Self-evaluation, 2007: 4). This awareness led to the adoption, at the University of Pretoria, of a model where computer and academic literacy courses became compulsory for obtaining a degree. These courses of 12 credits each shared the 24 credits conventionally allocated to first year courses. In the case of academic literacy, the UAL was tasked to measure, at the beginning of the academic year in January, the academic literacy level of each new student. If the level was too low, enrolling for an intensive year-long course to develop academic literacy became obligatory. Where the level was acceptable, faculty prescriptions for alternative language courses (where required) came into play. This model has worked well, and has been adopted by other institutions that have come to study and observe its advantages. It has also been confirmed by two external evaluations (in 2003 and 2007). The main advantages are that:

1. the compulsory course is part of the normal academic programme – a more desirable situation than if it were not since the latter is known to create bottlenecks and resistance; and
2. the need to develop academic literacy is addressed early, so that risk of failure associated with low levels of academic literacy is dealt with at the beginning of a course of study (Unit for Academic Literacy, Departmental Self-evaluation, 2007).

In order to assess the academic literacy levels of first year students, the Unit first made use of the English Literacy Skills Assessment for Tertiary Education

(ELSA Plus) developed by the University of Pretoria and Hough and Horne Literacy Consultants. However, in 2003 a switch from the ELSA Plus became necessary. Details of the reasons for the switch are outlined in a paper by Van Dyk and Weideman (2004a). In summary, their reasoning is that, "…the construct of the current test has become contested over the last decade, as a result of its dependence on an outdated concept of language, which equates language ability with knowledge of sound, vocabulary, form and meaning" (Van Dyk and Weideman, 2004a: 4). In addition to its construct being outdated, the test had to be hand-marked. It required large-scale and ever costlier administrative and logistical support. It was therefore decided that the Unit should develop its own test. It was in the year 2003 as well, that the panel for the first external evaluation recommended that the Unit change its name from the Unit for Language Skills Development to the Unit for Academic Literacy. According to the report:

> The major research, teaching and associated testing functions of the Unit are focused on academic language and literacy acquisition and its development, which are authentic academic activities and belong within an academic faculty, appropriately in the School of Languages in the Faculty of Humanities. In the long-term, consideration should be given to restructuring the Unit as an academic department. In the short term, a name change should be considered. We recommend the name, "The Unit for Academic Literacy" (Cliff, Crandall, De Kadt and Hubbard, 2003).

At the beginning of 2004, the newly developed Test of Academic Literacy Levels (TALL; in Afrikaans: TAG, Toets van Akademiese Geletterheidsvlakke) was used for the first time. Four universities (Pretoria, North-west, Stellenbosch and on one of its campuses, Free State) now use the test annually to determine the academic literacy levels of over 31 000 students. The success of these tests has been the subject of numerous papers, both presented at national and international conferences, and published in accredited journals (Weideman, 2003a; Van Dyk, & Weideman, 2004a; Van Dyk, & Weideman, 2004b; Van der Slik, & Wedeiman, 2005; Weideman, 2006a; Weideman, 2006b; Van der Slik, & Weideman, 2007; Weideman, & Van der Slik, 2008; Van der Slik, & Weideman, 2008; Weideman, 2009a; Van der Slik, & Weideman, 2009). The

TALL, TAG and TALPS (Test of Academic Literacy for Postgraduate Students) will be discussed in detail in chapter 2.

There are a number of modules offered by the unit as can be seen in Table1.1. These include the compulsory intervention modules (EOT 110 & 120). Students who are not at risk as determined by their test scores are required to take two other courses offered by the unit or as per faculty requirements. These could be two of the following: Academic Writing (EOT 162), Academic Reading (EOT 161), Legal Discourse (EOT 163) or Communication in Organisations (EOT 164). Find below a breakdown of the courses offered by the unit:

***Table 1.1**: Student numbers for courses offered by the UAL (2006-2010)*

	MODULE	DESCRIPTION	2006	2007	2008	2009	2010
	EOT 110	ACADEMIC LITERACY(1) 110	2615	2783	2259	2901	2880
	EOT 120	ACADEMIC LITERACY(2) 120	2474	2606	2143	2656	2646
	EOT 161	ACADEMIC READING SKILLS 161	1177	1297	1357	1282	1228
	EOT 162	ACADEMIC WRITING SKILLS 162	1387	1340	1334	1396	1439
	EOT 163	LEGAL DISCOURSE 163	705	695	809	664	565
	EOT 164	COMMUNIC. IN ORGANISATIONS 164	1711	1680	1741	1992	2069
	JNV 100	INNOVATION 100	270	251	251	245	165
FIRST YR MODULES (100-LEVEL)			**11546**	**10917**	**9894**	**11136**	**10992**
	JSQ 216	COMMUNICATION SKILLS 216	0	0	0	755	794
	JSQ 226	COMMUNICATION SKILLS 226	683	701	722	0	0
	UAL 210	WRITING ACADEMIC ESSAYS 210	0	0	0	21	0
SECOND YR MODULES (200-LEVEL)			**683**	**701**	**722**	**776**	**794**
	AFR 358	EDITING 358	30	25	18	21	21
	EOT 300	ADV. LANGUAGE PROFICIENCY 300	29	24	37	25	23
	TRL 352	LITERARY TRANSLATION 352	9	9	4	13	12
THIRD YR MODULES (300-LEVEL)			**68**	**58**	**59**	**59**	**56**
	AFR 767	EDITING 767	6	6	4	6	6
	EOT 702	LANG.INSTRUCTION &LEARNING 702	3	0	0	0	0
	TRL 751	LITERARY TRANSLATION 751	2	4	3	4	5
	TTS 751	ACADEMIC WRITING SKILLS 751	9	2	7	13	14
HONS MODULES (700-LEVEL)			**20**	**12**	**14**	**23**	**25**
TOTAL FOR DEPT			**12317**	**11688**	**10689**	**11994**	**11867**

(BIREP, 2011)

1.4 Problem statement

Tests have in general almost always been seen in a negative light (Shohamy, 1997; 2001; 2004; 2008; McNamara & Roever, 2006). Fulcher and Davidson, in an article which is an imaginary Socratic dialogue between J.S. Mills and Michel Foucalt about educational assessment, have Foucalt state that in society individual happiness is impossible because we are oppressed by the institutions of society, and one of the most evil of these is the test (Fulcher & Davidson, 2008: 407). The Foucalt character in this imaginary dialogue claims that

> testing is the method by which the powerful remain in power and decide what knowledge is to be valued. The test takers are mere objects that have no choice but to comply with the demands of the powerful. The purpose is to establish domination through endless testing, thereby placing value on what is cherished by the powerful, thus maintaining society's status quo (Fulcher & Davidson, 2008: 408).

Shohamy (2001), a leading theorist in the field of critical language testing, echoes these views. Her focus is on the voices and rights of the test taker and on pursuing the 'how' of testing rather than the 'why' (Shohamy, 2001: xii).

She differentiates between traditional testing and 'use-oriented' testing. Shohamy explains that traditional testing is concerned with topics such as methods for computing different types of reliability (i.e. how accurate test scores are), obtaining evidence of validity (i.e. the extent to which tests measure what they are expected to measure) and procedures for examining the quality of items and tasks (i.e. the extent to which test items and tasks measure the content being tested) (2001: 3). She points out that in traditional testing the focus is primarily on the test; the test taker is important only as a means for examining the quality of a test. Shohamy states that, in the traditional view,

> once the test is designed and developed, its items written and administered, its format piloted, items and statistics computed, reliability calculated and evidence of validity obtained, the role of the tester is complete. The task ends when psychometrically sound results are satisfactorily achieved (Shohamy, 2001: 4).

In her view traditional testing views tests as isolated events, detached from people, society, motives, intentions, uses, impacts, effects and consequences (2001: 4). 'Use-oriented' testing, on the other hand, sees testing as part of educational, social and political contexts. It is concerned with what happens to the test takers who take the tests, the knowledge that is created by tests, the teachers who prepare for the tests, the materials and methods used for tests, the decisions to introduce tests, the uses of the results of tests, the parents whose children are subject to the tests, the ethicality and fairness of the tests, and the long and short term consequences that tests have on education and society (2001: 4).

These are the very issues that are of concern to this study. Shohamy points out that in the field of testing, issues about the use of tests – i.e. intentions, effects and consequences – were neglected but that there has recently been a renewed interest in this topic. As a result of this, language testers have begun to address issues such as test ethicality, test bias, the effect and impact of tests on teaching and learning, and various issues related to the use of tests (see Spolsky, 2008). Shohamy concentrates on the voices and rights of the test taker as well as the power that tests have held over test takers. The uses of the results of a test can, for example, lead to detrimental effects. Shohamy explains this when she says: "It is often the performance on a single test, often on one occasion at a single point in time, that can lead to irreversible, far-reaching and high stakes decisions" (Shohamy, 2001: 16). Doing well on a test opens doors, performing poorly on a test shuts doors and shatters dreams. Yet, very often, the scores test takers receive are not questioned but quietly accepted "because of the blind trust they have in the authority of test results and their own limited power" (Shohamy, 2001: 16). One reason for this is that tests use the language of science, which grants "authority, status and power" (Shohamy, 2001: 21). Testing is therefore seen as a scientific discipline that cannot or should not be questioned, very often because very few members of the public understand its "language of science".

Critical language testing, in this perspective, is what is required to counter effectively the conventional uses and abuses of tests and test results. This "implies the need to develop critical strategies to examine the uses and consequences of tests, to monitor their power, minimise their detrimental force, reveal their misuses, and empower the test takers" (Shohamy, 2001: 131). Critical testing aims to encourage stakeholders in the field of language testing to ask important questions about the uses of tests and test results. The field of critical testing has broadened the field of language testing, moving it away from seeing test developers, test users and test takers as separate. Questions that the field is concerned with are:

> Who is the tester? What is their agenda? Who are the test takers? What is their context? What is the context of the topic being tested? Who is going to benefit from the test? Why is the test being given? What will the results be used for? What is being tested and why? What are the underlying values behind the test? What are the testing methods? What additional evidence is collected about the topic? What kinds of decisions are reached based on the test? Who, excluding the tester, is included in the design of the test and its implementation? (Shohamy, 2001: 134).

Clearly, defined like this, the field has become concerned with the ethical questions surrounding the field of testing, issues raised by McNamara and Roever (2006) as well.

McNamara and Roever attempt to find answers to important questions in the field by looking at both the psychometric approaches to fairness as well as the social dimensions of testing. The authors express their belief that psychometrics is not enough to validate a test, that a psychometrically good test is not necessarily a socially good test. What the authors suggest is needed is a consideration of the social dimension of testing. Like Shohamy, the authors too examine test use within given social contexts. Issues raised include the use of tests for immigration and citizenship as well as the use of tests to limit or control access into desired fields of study, countries of choice or chosen professions. It is their view on the way forward that is of particular importance

here. McNamara and Roever aim to create awareness about the importance of considering the social impact of tests. They stress the importance of an "adequate social theory to frame the issues that we wish to investigate" (McNamara & Roever, 2006: 253). These theories may be unfamiliar to language testers and, as they point out, will challenge "many of the fundamental epistemological and ontological assumptions of the field" (2006: 253).

McNamara and Roever argue for the broadening of the field of language testing, with input from fields as diverse as sociology, policy analysis, philosophy, cultural studies and social theory, "breaking down the disciplinary walls between language testing researchers and those working within other areas of applied linguistics, social science, and the humanities generally" (2006: 254). They stress the importance of what they call a "well-rounded training for language testers that goes beyond applied psychometrics" (2006: 255). They are quick to point out that they are not calling for the abandonment of training in psychometrics, but that they believe that testers should be well versed in psychometric theory, quantitative research methods, research on second language learning, and test construction and analysis. What they advocate is a training "that includes a critical view of testing and social consequences, whether those effects concern the educational sector (college admission, school graduation) or society at large (immigration, citizenship, security, access to employment)" (2006: 255).

It should be clear, then, that test developers, designers and users can today no longer ignore the social issues that surround the field of testing. Concerns raised by Shohamy (2001) and McNamara and Roever (2006) should be concerns of every test developer.

According to Fulcher and Davidson tests, when used correctly, "have the power to grant access to opportunities and goods that were previously unavailable to the ordinary people" (2008: 412). This is particularly true for South Africa, with

our history of apartheid and segregation. The majority of students who write TALL and who will write TALPS come from previously disadvantaged backgrounds, have received an inferior quality education or have been educated by teachers who have received an inferior quality education. Many of our students coming from rural areas may not have studied in English. TALL and TALPS are the kind of tests Fulcher and Davidson refer to above – these tests were designed with the objective of helping students achieve their goals and dreams. If such tests indicate that students have low proficiency levels, that will no doubt hamper their success at university level. When they write TALL at the University of Pretoria they already have access to a programme of study – it is successfully completing this programme that is often the problem. TALL and TALPS are used to identify a serious academic literacy problem and an intervention programme is designed to help develop the language that these students lack. Used correctly, tests can have positive effects. The need for accountability and transparency on the part of the test developer has not been lost on the test developers of TALL and TALPS, as will be demonstrated in the following chapters.

This thesis will therefore concentrate on an area of testing that has been largely ignored – the social dimension. According to McNamara and Roever "the social context of language assessment includes not just the designers and takers of a particular test, but also the purposes for which people take the test, and the ends to which the results of the test are put" (2006: xii). For Shohamy, asking questions about the social dimensions of language testing means asking questions about the "social and political issues of the uses of tests by focusing on the tester, the test-taker and other stakeholders" (2001: xv). A concern for the social dimensions of language testing means that one is forced to consider important questions related to every aspect of the test, from its design and implementation to the consequences of the uses of the test results as well as to the reason for giving the test, the effect of the test on the test-taker, concerns about the design of fair tests, the rights and responsibility of the test designer,

and the rights and responsibility of the test-taker. In short, the consideration of these social and political dimensions of language testing has broadened the field.

1.5 Key questions

This thesis will argue that construct and other empirically based forms of validity are not enough to validate a language test and that what is needed in addition is a detailed look at issues of transparency, accessibility and accountability. It will examine whether it is possible to build destigmatisation measures into the design of the test, rather than presenting a subsequent defence in the face of objections from those affected by the test results. It will attempt to determine whether it is possible to anticipate such objections, and deal with them by altering the design, presentation or administration of the test. It will attempt, furthermore, to determine acceptable levels of theoretical defensibility of the test design and social accountability in view of the uses to which the results of the test will be put.

In order to do this, an exposition will be given of TALPS from its initial conceptualisation to its final implementation in January 2008. A detailed study will be made of the concepts of accessibility, transparency and accountability as they relate to TALPS. A key question that this thesis will investigate is whether construct and other conventional forms of validity are enough to validate a language test or whether what is needed in addition is a detailed look at issues of transparency, accessibility and accountability, with reference to the proposed Test of Academic Literacy for Postgraduate Students (TALPS). Here are a number of related questions to consider in this respect:

1. What is transparency and accountability in language testing? Can test developers ignore the social dimension and ethical issues related to testing? If not, how can attention to such concerns be theoretically justified?

2. Is it possible to anticipate all issues related to transparency and accountability? Can these be anticipated to such an extent that it may be possible to design solutions to them into the test? If not, what is the minimum that can be or should be anticipated, and what can be done to keep the design and production process of a test, as well as its administration, open to academic and public scrutiny? Conversely: what design and administrative processes would inhibit fair academic and public scrutiny?
3. What is accessibility in language testing? How much information is available to the test takers about the test? How can test designers ensure further accessibility? The issue of the accessibility of the test is one that needs to be explored further. While there are several practical examples of how this was accomplished in the past, specifically with TALL (example on the UAL website, brochures, information provided at the Open Day), this needs to be explored further.
4. To what extent is it possible to build destigmatisation measures into the design of the test, rather than presenting a subsequent defence in the face of objections from those affected by the test results?
5. How much support can be derived and should be derived from empirical analyses to assist in taking decisions about the social, ethical and related dimensions of tests? Is there complementarity among the various empirical components and the social and ethical dimensions of a test? If so, how can a theoretical account of these be given?

1.6 Chapter outline

The rest of this study will comprise of the following chapters:

1.6.1 Chapter 2: Telling the story of a test

Chapter 2 will follow Shohamy's exhortation "to tell the story of a test" (2001). This chapter will begin with a discussion of TALL since this test was the sounding board on which TALPS is based. In keeping with the intention to tell the story of the test, this chapter will attempt to document the progress made with TALPS from its initial conceptualisation, design and development to its trial (pilot tests), the results of these trials and its final implementation in January 2008. Using the empirical evidence gathered in this process, conclusions and assertions will be made about the test.

1.6.2 Chapter 3: Theoretical framework for understanding foundational concepts in language testing

This chapter will explore the theoretical framework that has informed the research. It will discuss the need for a theoretical analysis or justification for applied linguistic designs and will provide a definition or explanation for the terms constitutive and regulative. This chapter will explore the theoretical framework that is implicit in the distinctions made by Weideman (2003a; 2003b; 2006a; 2007a; 2007b; 2009a), in order to give an explanation of these conceptions, and to ascertain whether a theoretically coherent account of some apparently disparate testing concepts is possible. The chapter will conclude with a detailed discussion of the concept of validity in language testing, with specific emphasis on the distinctions developed by Messick (1980; 1981; 1989a; 1989b; 1996), Bachman and Palmer (1996), and Kunnan (2004).

1.6.3 Chapter 4: The constitutive concepts underlying the design of TALPS

This chapter provides a detailed discussion of the validity and reliability measures of TALPS. It will take the form of a set of claims that will be used to validate the test.

1.6.4 Chapter 5: Transparency issues in testing academic literacy: The case of TALPS

Key questions addressed in this chapter are:

- What is transparency in language testing?
- Is it possible to anticipate all issues related to transparency?
- How may test developers of TALPS ensure transparency?

1.6.5 Chapter 6: The accessibility of TALPS

This chapter looks at issues related to the accessibility of TALPS. A first concern in this regard is the question of internal accessibility: How accessible is the test to test takers? A second concern dealt with in this chapter is the question of how the test should be used. Will it be used as a high stakes test that will deny students access into desired programmes, or will it be used as a low to medium stakes test that tests students' academic literacy levels and then places them in a programme designed to help improve their academic literacy if this is needed? This chapter will, moreover, include a summary of the data collected from a questionnaire administered to students who wrote TALPS. The questionnaire is designed to elicit information, comments, questions and reactions from testees about the test. Finally, this chapter will consider the various responsible choices open to the test developers in each of these (high and low stakes) eventualities.

1.6.6 Chapter 7: Accountability

This chapter will consider issues of accountability in language testing and will attempt to answer one of the key questions of the study: Are psychometric analysis and the empirical results yielded by such analysis enough? If test developers are to be publicly accountable, should their designs and motivations not be understandable to the public? This chapter looks at the use of terms like "dual accountability" and "public accountability" (Bygate, 2004: 19) as used by

leading theorists in applied linguistics and in the field of language testing. Specifically, the chapter will consider whether the notion of ‘standards’ is sufficient to allow a clear articulation of the idea of accountability, and whether such an idea does not in the first instance presuppose the idea of transparency. It will be argued that invoking ‘standards’ as a sufficient guarantee of accountability is still a criterion that fails to venture beyond conventional notions of ‘accountability’ and ‘fairness’.

1.6.7 Chapter 8: Regulative conditions for test design

This chapter will begin by providing a summary of the answers to the important questions that have been asked and answered in the course of this investigation. How much were test developers able to anticipate in the design of the test? Have issues of transparency, accountability and fairness been adequately considered and dealt with? This chapter will, also, in addition to considering the link between transparency, accessibility and accountability, focus briefly on other conditions in the framework, that contribute to the design of fair and socially acceptable tests.

1.7 Conclusion

This chapter has outlined the steps that had to be taken to deal with the poor academic literacy of students at tertiary institutions in South Africa. It has highlighted the need for tests like TALL and TALPS, tests that are designed and should be used in ways that benefit rather than disadvantage already disadvantaged students. It has also been pointed out that this study focuses on the social dimension of testing, with a consideration of concepts hitherto largely ignored in conventional approaches to the field of language testing.

The purpose of the next chapter is to document the story of TALPS, providing an overview of the steps that were followed, from its conceptualisation to its final implementation.

Chapter 2

Telling the story of a test

2.1 Introduction

In an imaginary dialogue between Mill and Foucault on educational assessment, Mill asks, “If the purpose of government is to ensure the happiness of the people, and happiness is knowledge (as Socrates claimed), is it not possible for tests to play some positive role”? He then answers, “So tests, used correctly, have the power to grant access to opportunities and goods that were previously unavailable to the ordinary people” (Fulcher & Davidson, 2008: 412).

This view is quite contrary to much of what is available in the literature on testing. Unfair tests, unfair testing methods and the use of tests to restrict and deny access have ensured a negative attitude to tests. Anyone reading the literature available on tests and testing is bound to come across numerous examples of tests or organisations that have used tests negatively (Shohamy, 2001; 2008). McNamara and Roever (2006: xii) make reference to one of the earliest examples of these: The Shibboleth Test, as recorded in the Book of Judges in the Hebrew Bible. Around three thousand years ago in the war between the Ephraimites and the Gileadites, both part of the Hebrew tribes, about forty-two thousand Ephraimites were killed for crossing into Gilead territory. The Ephraimites were given a simple language test. They were to pronounce a particular word (for “ear of grain”). This test was designed to distinguish the Ephraimites, whose language lacked a particular sound, from the Gilieadites, whose dialect included the use of this sound. Those who did not pronounce that particular sound were put to death (McNamara & Roever, 2006: xii). Shohamy (2001) uses the example of an Arabic test given to Hebrew speakers in Israel who spoke Arabic as a second language. According to Shohamy, because of the political conflict between Israel and the Arabs, the Arabic language held very low status and there is no motivation among Hebrew

speakers to speak the language. The national inspector of Arabic, who was responsible for the test, made it clear in a number of statements that "measuring the level of Arabic was a method of imposing a change in the status and role of the Arabic language" (2006: 60). This is just one more example of how tests are used, to "impose national ideologies and beliefs about languages and the suppression of diversity" (Shohamy, 2008: 369). Other examples in the literature concern the use of tests to deny access to immigrants seeking entrance to a foreign country or to deny access to an educational institution/field of study. Shohamy quotes an example of a university that used tests to ensure good enrolment figures for a particular language class. The tests developed by that department tested only grammar, "knowing *a priori* that this is a weak area among students" (Shohamy, 2001: 90). Students failed the test and had to enrol for the particular language course (2001: 90).

The fact of the matter is that tests have effects on test takers and opportunities are denied because of poor performance on a test. The issue of being denied access is one that is rooted in the history of our country. Chapter 1 has provided a detailed explanation of the issues surrounding language and learning at tertiary institutions in post-apartheid South Africa, looking specifically at the University of Pretoria and the intervention strategies applied to deal with these.

The purpose of this chapter is to tell the story of the design and development of a specific set of tests. While the focus in this thesis is on the Test of Academic Literacy for Postgraduate Students (TALPS), this story must begin with the Test of Academic Literacy Levels (TALL), since this test was the sounding board on which TALPS is based. TALL was designed to test the proficiency of the academic language of first year students. Butler's study (2007) has highlighted the need for a similar test for postgraduate students. Mill and Foucault's imaginary conversation is again relevant. In designing TALL and later TALPS, the test developers wanted to ensure that the test played "some positive role" and granted "access to opportunities and goods that were previously

unavailable to the ordinary people" (Fulcher & Davidson, 2008: 412). In order to determine the success of the test developers in this regard, this chapter begins with a brief discussion of TALL before documenting the progress made with TALPS, from its initial conceptualisation, design and development to its trial (pilot test), the results of these trials and its final implementation.

2.2 The design and development of TALL

The decision by the Unit for Academic Literacy (UAL) to switch from the ELSA Plus was motivated in the first instance by the fact that the construct was based on an outdated view of language. The first step for the developers of the new test, which would eventually be referred to as TALL (Test of Academic Literacy Levels), was to determine "what does a construct based on a theory of academic literacy look like?" (Weideman, 2003a: 59). We may define a construct as "an ability or set of abilities that will be reflected in test performance, and about which inferences can be made on the basis of test scores" (Davies, Brown, Elder, Hill, Lumley & McNamara, 1999: 7). In discussing the proposed new construct, Van Dyk and Weideman states that

> the test construct or blueprint defines the knowledge or abilities to be measured by that specific test…a construct is usually articulated in terms of a theory, in our case a theory of language, and more specifically a theory of academic literacy (Van Dyk & Weideman, 2004a: 7).

The proposed blueprint for the test of academic literacy for the University of Pretoria requires that students should be able to:

- understand a range of academic vocabulary in context;
- interpret and use metaphor and idiom, and perceive connotation, word play and ambiguity;
- 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;
- interpret different kinds of text type (genre), and show sensitivity for the meaning that they convey, and the audience that they are aimed at;
- interpret, use and produce information presented in graphic or visual format;

- 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;
- 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;
- 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;
- 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 (Weideman, 2003a: 61).

This construct differs greatly from that of the ELSA PLUS which, as stated earlier, was based on an outdated view of language which "equates language ability with knowledge of sound, vocabulary, form and meaning" (Van Dyk & Weideman, 2004a: 4). It is instead based on what they have called an "open view of language" (Van Dyk & Weideman, 2004a: 5) as indicated in the table below:

Table 2.1 Two perspectives on language

Restrictive	Open
Language is composed of elements: • sound • form, grammar • meaning	Language is a social instrument to: • mediate and • negotiate human interaction • in specific contexts
Main function: expression	*Main function*: communication
Language learning = mastery of structure	*Language learning* = becoming competent in communication
Focus: language	*Focus*: process of using language

(Van Dyk & Weideman, 2004a: 5)

This move towards a more open view of language is indicative of what Bachman and Palmer (1996: 23) refer to as authenticity, as will be highlighted in the third chapter below. Explained simply, what Bachman and Palmer (1996) call for is that there should be correspondence between the test tasks and the use

of language in real life situations. This call for designing authentic test tasks is very clearly a first step towards ensuring accessibility and fairness in language testing. It is imperative that test takers see a link between the test tasks and the use of language in real life. Working with a construct such as this ensures that the intervention designed for students who fail to achieve the required level on the test is based on the same construct or blueprint as the test.

The next important step in the design process is to design appropriate task types in line with the blueprint. An important decision made by the test designers was to use a multiple choice format for the test. The reasons for this, as outlined by Van Dyk and Weideman (2004b), were the size of the population and the need to have the results ready urgently. The multiple choice format allowed for the test to be marked electronically rather than manually, thus ensuring that the results were ready on time. Using this format has, according to Van Dyk and Weideman, allowed them to become more “inventive and creative than we would otherwise have been, if we had simply succumbed to the prejudice that one cannot test (this or that aspect of) language in this way” (2004b: 16). The following are examples of this inventiveness, based on a reading passage that was used to test the understanding of metaphor – a dimension of language use that conventionally might easily have been considered impossible to test in this format:

> We should understand the phrase "milk in their blood" in the first sentence to mean that both men
>
> (a) have rare blood diseases inherited from their parents.
>
> (b) are soft-spoken, mild-mannered young farmers.
>
> (c) don’t like to make profit at the expense of others.
>
> (d) are descended from a long line of dairy farmers.
>
> Paragraph 2 speaks of 'hatching a plan'. Normally, we would think of the thing that is hatched as
>
> (a) a door.
>
> (b) a loft.
>
> (c) a car.

(d) an egg.

Or consider this one, which is designed to test the knowledge of the candidate regarding what counts as evidence:

In the second paragraph, we read that "milk farms have been the backbone of this country" for centuries. Which sentence in the fourth paragraph provides evidence of the claim that it has been so 'for centuries'?

(a) The first sentence

(b) The second sentence

(c) The third sentence

(d) None of these (Van Dyk & Weideman, 2004b: 16).

Importantly, the authors of the article point out that the tasks the test takers are being asked to perform belong to a set of abilities or task types that are much broader in scope than that of a test that defines academic literacy in terms of skills, or reduces it to the mastery of sound, form and meaning (Van Dyk & Weideman, 2004b: 16).

The TALL tests, written between 2005 and 2008, generally had just more than 60 items distributed over six sections. The 2005 version of the test had a seventh section on academic writing. Students have 55 minutes to complete the test, which is out of 100, since about half of the items count 2 or 3 instead of 1 (Van der Slik & Weideman, 2008: 364). Below is a description of the sections, number of items and marks (more or less) allocated to each:

- Section 1: Scrambled text (5 items, 5 marks)
- Section 2: Knowledge of academic vocabulary (10 items, 20 marks)
- Section 3: Interpreting graphs and visual information (7 items, 7 marks)
- Section 4: Text type (5 items, 5 marks)
- Section 5: Understanding texts (20 items, 47 marks)
- Section 6: Text editing (16 items, 16 marks)

(Van der Slik & Weideman, 2008: 364).

TALL and TAG were initially not access tests – they were not used to determine whether a student gains access to a desired field of study or not. Instead they were conceived of as placement tests used to determine the level of a student’s

academic literacy. They are, however, being increasingly used for access, as at the Stellenbosch University where TALL and TAG form part of the Access Test Battery. The Battery consists of five tests: Language (TALL & TAG), Thinking Skills, Numeracy Skills, Physical Science (Chemistry & Physics) and Mathematics. Different faculty prescriptions require that all five tests or a combination of three tests be written. TALL and TAG are always a part of the combination. The aggregate received on the tests are used in combination with students' Grade 12 results to determine access. If students have poor academic literacy levels, this will definitely hamper academic success. Poor academic language proficiency has broader effects such as students not completing their studies in time, parents having to pay for extra years spent at university, loss of income for students for every extra year spent studying, and poor throughput rates for the higher education system (Weideman, 2003a: 56). TALL and TAG, as used at the University of Pretoria, are not high stakes tests that prevent or allow access but are instead low to medium stakes tests – requiring that students who are considered at risk based on their test score attend a specially designed intervention programme. The intervention programme is a year long (EOT 110 & 120) and is designed to help develop the language ability that students would need to be academically successful. The test designers of TALL were concerned, from as early as the design stage, with ensuring that the test they designed and used was not just a fair and reliable test, but that as test developers they were responsible and accountable for their designs.

Conventionally, it has always been accepted that a test must be valid and reliable. If a test is not valid then it does not test what it was designed to test and the inferences we make about the test taker based on their test scores are also in doubt. A test must also be reliable, stable or consistent. Reliability measures such as Cronbach's alpha or Greatest Lower Bound (GLB) are used to determine whether the test measures consistently in different situations. According to Van der Slik (2006) the fairness with which a test measures is crucially dependent on its reliability. In addition to this, test developers still

need to answer questions related to the consequences and impact of the test on test takers. Also, some empirical analyses may not be understood by those affected by the use of the test scores, and places an additional responsibility on test designers.

A first concern of test takers of TALL should be to determine whether it is indeed a reliable test. Evidence of this has been the subject of a number of papers (Van der Slik & Weideman, 2005; Van der Slik & Weideman, 2008; Weideman & Van der Slik, 2008; Van der Slik & Weideman, 2009). Empirical analyses point out that TALL is indeed a reliable test, as indicated in the table below. These statistical measurements are based on a number of different administrations of the test between 2005 and 2008 at the University of Pretoria, Stellenbosch University and North-West University (Potchefstroom and Vanderbijlpark campuses):

Table 2.2 Selected properties of the academic literacy test (2005-2008) (standard deviations in italics)

TALL	UP	US	NWU	Overall
N	15,202	13,886	675	29,793
Mean proportion correct (*difficulty)*	.65 *(0.05)*	.69 *(0.05)*	.49 *(0.13)*	.61 *(0.12)*
Mean Cronbach's alpha *(reliability)*	.92 *(0.01)*	.88 *(0.01)*	.91 *(0.03)*	.90 *(0.02)*
Mean Average *Rit* (*discrimination index)*	.45 *(0.01)*	.38 *(0.01)*	.45 *(0.02)*	.43 *(0.04)*

(Van der Slik & Weideman, 2009: 257)

In addition to the question of the reliability measures of the test, the test developers asked other pertinent questions, the answers to which would ensure further transparency. These questions, among others, dealt with whether the test was reliable or "robust enough" (Van der Slik & Weideman, 2009); whether there was variation in the results, and, if so, whether this was as a result of the technical inconsistency of the test or as a result of differences in the population

of students taking the test; how fair the test was in terms of miscalculations (students who should have passed but did not and students who passed but should not have); and whether the scores achieved by the populations as a whole differed across the various populations (Van der Slik & Weideman, 2009).

Although TALL is a reliable, stable test, Van der Slik and Weideman (2009: 257) point out that potential miscalculation occurs because tests are never entirely reliable measuring instruments. Miscalculations refer to those students whose score showed them to be at risk when they were not, and vice versa. Students who may have been wrongly identified as being at risk are given a second chance or borderline test, thus eliminating potential negative consequences they may experience with the test and their score.

Table 2.3 Potential misclassifications on the English version of the academic literacy test (Percentage of this test population). In italics the corresponding intervals (in terms of standard deviations) around the cut-off points

TALL	UP	US	NWU
Alpha based: Correlation between test and hypothetical parallel test			
2005	432 (13.0%) 63 – 74 *(.31)*	246 (14.2%) 63 – 74 *(.41)*	16 (11.8%) 64 – 71 *(.18)*
2006	439 (12.0%) 51 – 59 *(.25)*	432 (11.7%) 52 – 58 *(.25)*	20 (13.7%) 45 – 54 *(.26)*
2007	448 (11.5%) 47 – 55 *(.19)*	604 (14.5%) 54 – 61 *(.24)*	18 (12.8%) 43 – 52 *(.19)*
2008	179 (4.1%) 30 – 35 *(.15)*	152 (3.6%) 34 – 42 *(.24)*	26 (10.0%) 37 – 43 *(.15)*
Average % ***(Average sd)***	**(10.0%)** ***(.23)***	**(11.0%)** ***(.28)***	**(12.0%)** ***(.20)***

(Van der Slik & Weideman, 2009: 258)

Using the data above, Van der Slik and Weideman conclude that "overall, however, miscalculations occur more or less within the expected range of scoring points around the cut-off point, i.e. around .25 SD around the cut-off point" (2009: 258). In answering the question of whether the scores of the University of Pretoria, Stellenbosch University and North-West University

differ from each other in respect of the various administrations of TALL, it was found that while there are differences among the different student populations, these are not large. The reason for the differences can be attributed to a difference in the make up of the population rather than inconsistencies with the test or its administration. Van der Slik and Weideman (2009: 260) explain that the English proficiency at Stellenbosch University (SU) is (and has historically probably been) much higher than that of students at the University of Pretoria (UP) and the North-West University (NWU). The SU takes in students from more affluent areas, while it is possible that the NWU takes in more students from disadvantaged backgrounds. The table below highlights these findings:

Table 2.4 T-Statistics (and effect sizes) for TALL 2005-2008

	Max. score	UP vs. US *(d)*	UP vs. NW *(d)*	US vs. NW *(d)*	UP Mean (SD)	US Mean (SD)	NWU Mean (SD)
2005	100	–10.60 *(–.29)*	6.28 *(.62)*	8.94 *(1.13)*	71.75 (19.31)	76.89 (14.57)	59.70 (21.97)
2006	100	– 9.81 *(–.23)*	4.88 *(.41)*	7.49 (.68)	64.32 (20.02)	68.46 (16.54)	56.27 (19.18)
2007	100	– 9.21 *(–.21)*	5.57 *(.55)*	7.90 (.75)	61.11 (20.59)	64.98 (16.79)	50.44 (21.57)
2008	100	*Not available*	6.28 *(.41)*	*Not available*	62.59 (20.15)	*Not available*	54.34 (20.30)

(Van der Slik & Weideman, 2009: 260)

Van der Slik and Weideman explain these differences by stating that "since the administration of the test is, by all accounts, subject to the same set of standardised administrative procedures for test implementation, the differences that we have again noticed here can be explained, no doubt, by differences in the composition of the various student bodies" (Van der Slik & Weideman, 2009: 260).

Other interesting analyses available on TALL are that of the worst and best performing subtests based on their reliability measures. According to Van der Slik and Weideman (2009: 258), throughout the years two subtests have stood

out as the worst (**Interpreting graphic and visual information**) and the best performing ones (**Text editing**) as indicated in the table below:

Table 2.5 Selected properties of the relatively worst (GVI) and best performing (TE) subtests of TALL (2005-2008) (standard deviations in italics)

Interpreting graphic & visual literacy	**UP**	**US**	**NWU**	**Overall**
Average number of items				6.5
Mean proportion correct	.74 *(.06)*	.79 *(.05)*	.63 *(.07)*	.71 *(.09)*
Mean Cronbach's alpha	.63 *(.09)*	.57 *(.14)*	.63 *(.08)*	.61 *(.10)*
Mean average *Rit*	.58 *(.03)*	.56 *(.02)*	.59 *(.03)*	.57 *(.03)*
Text editing				
Average number of items				16.5
Mean proportion correct	.61 *(.13)*	.64 *(.05)*	.47 *(.04)*	.57 *(.11)*
Mean Cronbach's alpha	.90 *(.06)*	.89 *(.01)*	.92 *(.01)*	.90 *(.04)*
Mean average *Rit*	.65 *(.07)*	.63 *(.04)*	.69 *(.01)*	.66 *(.05)*

(Van der Slik & Weideman, 2009: 259)

Data like this allow test developers to ask and answer a number of questions about the test and the test taker. Examining such data is exactly what test developers need to do in order to make their tests more accessible and transparent. Van der Slik and Weideman (2005: 24) point out that the "ongoing refinement of a test, through various empirical means, can indeed continue to serve several responsible purposes". In terms of the data presented above, Van der Slik and Weideman (2009: 258) point out that the subtest **Text editing** has superior reliability (Cronbach's alpha .90) as compared to the subtest **Interpreting graphic and visual literacy** (Cronbach's alpha 0.61) which scores much less on the same measure. The question of whether this should be of concern to the test developers and designers is an important one. The authors of the article explain that both the subtests measure necessary components of the construct and cannot be excluded. The difference in the reliability scores,

however, is more than likely due to the difference in length of the subtests. They state that:

> In other words, since the Text editing subtest is more than twice the length of its lower performing counterpart, it is its length, the fact that its measurement is achieved over many more items, that gives it the edge. If one hypothetically had a subtest for Graphic and visual literacy that was also about 16 items long, its expected alpha would have risen from its current .61 (for TALL) and .56 (for TAG) to, respectively, 0.7 and 0.81 (Van der Slik & Weideman, 2009: 259).

The data briefly discussed above, points to the power of TALL in terms of its reliability. This evidence is freely available to anyone interested in knowing about TALL. TALL was not initially designed to be used as a high stakes test but as a medium to low stakes test, requiring that students whose academic literacy is below a certain level take a compulsory year long intervention programme. This compulsory programme does not add on a year to the student's programme – it is not an extended programme but is built into the degree.

Despite this, however, there is a stigma attached to taking the two courses (EOT 110 in the first semester and EOT 120 in the second semester) that make up this intervention programme. One of the questions sometimes asked by parents and by students is whether the student needs to take the test if she or he has done well at English/Afrikaans in Grade 12. What needs to be explained is that the test tests the academic literacy of students. It is used to determine whether that student is equipped with the knowledge, language ability and skills needed to deal with the kind of language she or he will encounter specifically at university level. Even a cursory examination of the curricula, final exit examination papers and textbooks will reveal that this is not the same as the English/Afrikaans they studied at school. Students still, however, see the courses in a very negative light. One reason could be that students feel that by having to attend the intervention programme they are seen as not proficient in the English/Afrikaans language. This could be especially true for students attending English courses simply because proficiency in English is seen as a ticket to the international

business world. Realistically, it is simply not fashionable not to be proficient in English. Students fail to understand that the emphasis is on developing their academic literacy – one can be proficient in English, for example be fluent in speaking it in conversation and transactional contexts, but not academically literate, leading to poor performance and possibly failure. While it may not be possible to erase this stigma completely, empirical analyses, like those discussed above, assist in convincing students of not just the relevance, but also of the reliability of these tests.

While the data as it is presented above may not mean anything to the layperson, proof that the test is reliable, that it is based on an appropriate construct, that in the event of miscalculations students get to write a second chance in writing a borderline test, that each item was carefully designed or chosen, that the items were piloted and weak items discarded, and that careful planning had gone into the administration of the test, all ensure that the test developers have considered and taken seriously the responsibility that comes with designing tests. While this may not mean that students are any happier to attend the intervention programme, knowing this does mean that they should be more willing to accept the score as a realistic indication of their academic literacy levels.

2.3 The need for TALPS

The UAL is focused not only on developing the academic literacy of undergraduate students. Its EOT 300 course is an academic writing course offered to postgraduate students. The majority of students enrolled at the University of Pretoria prefer English as the language of learning – this despite the fact that for most of these students English is a second, third and sometimes fourth language. The university attracts students not just from South Africa, but also from other parts of Africa and the world. Very often these students do not have English as a first language. TALL and the intervention programme are in place to deal with this at undergraduate level. For the last five years the EOT 300 course has been focused on helping to develop the academic writing ability

of postgraduate students. There has, over this period, been an increasing demand for the course, as supervisors recognised the poor academic literacy levels of their students. This has been the focus of a study conducted by Butler (2007). The study, completed as part of his PhD, is concerned with the design of a course for academic writing at tertiary level. He states that the "immediate context of this study derives from the concern that a number of academic departments from a variety of disciplines at the University of Pretoria have expressed about the academic writing ability (and general language proficiency) of their postgraduate students" (2007: 10). He explains that these students are unfamiliar with academic writing conventions, are often unable to express themselves clearly in English, and have not "yet fully acquired the academic discourse needed in order to cope independently with the literacy demands of postgraduate study" (2007: 10).

Information elicited from a questionnaire designed by Butler was used to gather information about the perceptions supervisors have of their students' academic literacy levels. Some of the findings relevant to this study are listed below:

- Supervisors appear to be aware of the general language status of their postgraduate students in the sense that additional language users of English outnumber mother tongue Afrikaans and English users respectively at the university. A large number of comments by respondents were also directed at the literacy problems of additional language users specifically.
- Supervisors generally believe that an adequate level of academic literacy is crucial in the successful completion of postgraduate studies.
- A large majority of respondents believe that their postgraduate students' academic literacy levels range from average to poor.
- Almost all respondents feel that students should already be academically literate when they are admitted to postgraduate studies.
- There is general agreement that measures and strategies to select academically literate students are not always successful. Less than 50% of their supervisors indicate that the academic literacy of their postgraduate students is formally assessed.
- Supervisors are generally prepared to accept support from the UAL in the development of their students' writing ability. The majority of supervisors also indicate that they share this responsibility with language and writing experts (Butler, 2007: 126).

In addition to the questionnaire that was administered, Butler also conducted personal interviews with supervisors. The intention was to collect as much information as possible about the academic literacy of postgraduate students. The information collected from the personal interviews confirmed that there are serious academic literacy problems experienced by these students, and that as a result of these problems students do not complete their studies in the required time. What became clear from the data derived from the questionnaires and the information from the interviews was the need for a "reliable literacy assessment instrument" (Butler 2007: 181) that would "provide one with accurate information on students' academic literacy levels" (2007: 181). According to Butler

> interviewees were, therefore, also questioned about the relevance of a postgraduate literacy test and, without exception, expressed their eagerness to have access to such a test for the early determination of the academic literacy of their postgraduate students. This would enable them to determine timeously the relevant developmental opportunities for their students that focus on addressing specific literacy difficulties (Butler, 2007: 182).

The need for an academic literacy test for postgraduate students had been identified. In June of 2006 the University of Pretoria was invited to participate in the South Africa-Finland Co-operation Programme in which the Higher Education Quality Committee (HEQC) was involved. The invitation was to compete with other higher education institutions for grants which had the following objectives:

- Improve the quality of teaching and learning;
- Promote quality related innovations in teaching and learning, research and community engagement;
- Improve institutional systems and methods to enhance quality. (Council on Higher Education, 2006: 1).

The project proposal submitted by the UAL was entitled the *Postgraduate academic literacy initiative.* The executive summary below outlines the aims of the project:

Executive summary

The Postgraduate academic literacy initiative will address the urgent institutional need to develop the academic literacy levels of the increasing numbers of post-graduate students of the University of Pretoria who do not have English as a first language, and may be at risk academically, both in terms of delivering quality writing, and in the completion of their studies.

The project will have two main deliverables:

- a postgraduate test of academic literacy levels and
- an intervention designed, *inter alia*, on the basis of the test results, for the development of the academic literacy levels especially of master's and doctoral students.

The results will be disseminated both nationally and internationally, and the initiative will indirectly benefit also some of our European partners (at the Radboud University of Nijmegen) who are engaged in a similar project.

(Weideman & Butler, 2006: 1)

In October of 2006 the university was informed that the UAL had been awarded a grant of R70 000 in the Category: Grants for Quality Related Innovations (Departmental Grants). With the institutional need for the test and the funding now available, work on the test began. The story of TALPS, i.e. its design and development is documented in detail below.

2.4 The TALPS Project

According to the project proposal entitled *Postgraduate academic literacy initiative* **(Appendix A)** the project had two main components: test development and intervention design. The focus below is specifically on the development of the test, which entails how the test developers set about

- researching an appropriate construct i.e. a definition of academic literacy that can be operationalised, before deciding on test task types and drawing up test item specifications that are aligned with this construct;

- developing test items within the various task types;
- piloting these items on selected groups of postgraduate students;
- after item selection, preparing a final draft of a final version of such a test

(Weideman & Butler 2006: 2).

Table 2.6 Schedule of steps to achieve the aims of the project

Target date	Action
30 November 2006	1. Set up project team. 2. Recruit master's students for researching aspects of the test design and development. 3. Finalize construct.
31 January 2007	4. Decide on task types. 5. Draw up item specifications.
31 March 2007	6. Complete development of test items.
30 April 2007	7. Complete first round of trials (pilot test).
31 May 2007	8. Complete second round of piloting. 9. Decide on course design principles.
30 June 2007	10. Finalise draft of first version of test. 11. Decide on task types for intervention.
31 October 2007	12. Write up results of project so far; make substantial advances with completion of master's dissertations involved.
30 November 2007	13. Complete first draft of materials for course.
End 2007	14. Produce project report.

(Weideman & Butler, 2006: 6)

Ethical clearance for the project was sought and granted by the Research and Ethics Committee of the University of Pretoria.

2.5 Deciding on a construct

A first step for the developers was to find an appropriate construct on which to base the test and the intervention. Bachman and Palmer (1996: 21) define a construct as the "specific definition of an ability that provides the basis for a given test or test task and for interpreting scores derived from this task." The developers chose to base TALPS on the same construct as TALL. TALL was in many ways a sounding board for TALPS – the success of TALL was one of the most important factors that motivated the development of TALPS. Both these tests are designed to test the same thing: the academic literacy of students, the only difference being that one is directed at first year students while the other is

intended for postgraduate students. The construct on which the test is based is discussed earlier in this chapter. What is important here is the fact that the construct decided on "constitutes a definition of academic literacy" (Weideman, 2003a: 61). The construct is a blueprint of what students should be able to do at tertiary level, and this is what the test tests. Weideman states that "these abilities and components echo strongly, we believe, what it is that students are required to do at tertiary level" (2003a: 61). Weideman explains that in the process of developing the construct for TALL this view of academic literacy had been discussed at conferences and seminars and with "trans-disciplinary panels of academics" (2003a: 61). The responses have confirmed that "the elements identified above indeed constitute a number of essential components of what academic literacy entails" (2003: 61). A further confirmation of the positive reception of the construct, according to Weideman, came in the form of offers from other institutions to become partners in the development of the test or to use the test (TALL) on their students (2003a: 61). TALL also proved to be a highly valid and reliable test. In light of this the designers of TALPS were more than justified in using a blueprint that had already proved successful.

2.6 Specifications

The next step for the developers of TALPS was to align the construct of the test with specifications. Davies *et al.* (1999: 207) define test specifications as a document which sets out what a test is designed to measure and how this will be tested. Davidson and Lynch (2002: 4) state that "the chief tool of language test development is a test specification, which is a generative blueprint from which test items or tasks can be produced". They state also that a well-written test specification can generate many equivalent test tasks. The discussion of specifications at this point is focused specifically on item type specification and how they align with the construct of academic literacy used for this test. The test developers turned to the specification of task types devised for TALL:

Table 2.7 Specifications and task types: TALL

Specification (component of construct)	Task type(s) measuring / potentially measuring this component
Vocabulary comprehension	Vocabulary knowledge test, Longer reading passages, Text editing
Understanding metaphor and idiom	Longer reading passages
Textuality (cohesion and grammar)	Scrambled text, Text editing, (perhaps) Register and text type, Longer reading passages, Academic writing tasks
Understanding text type (genre)	Register and text type, Interpreting and understanding visual and graphic information, Scrambled text, Text editing, Longer reading passages, Academic writing tasks
Understanding visual and graphic information	Interpreting and understanding visual and graphic information, (potentially) Longer reading passages
Distinguishing essential/non-essential	Longer reading passages, Interpreting and understanding visual and graphic information, Academic writing tasks
Numerical computation	Interpreting and understanding visual and graphic information, Longer reading passages
Extrapolation and application	Longer reading passages, Academic writing tasks, (potentially:) Interpreting and understanding visual & graphic information
Communicative function	Longer reading passages, (possibly also:) Text editing, Scrambled text
Making meaning beyond the sentence	Longer reading passages, Register and text type, Scrambled text, Interpreting and understanding visual and graphic information

(Van Dyk & Weideman, 2004b: 19)

Using this specification, the developers were able to decide on the subtests or task types that would be included in the test.

2.7 The eight subtests in TALPS

ELSA Plus, which was the original test used to determine the academic literacy of first year students, included a component testing the writing ability of students. This question needed to be hand marked. Hand marking up to six

thousand answers in a very short space of time to allow the results of the test to be available as quickly as possible could take up to five days to complete (Van Dyk & Weideman, 2004a: 4) and created an administrative nightmare. In the new version of TALL it was decided to leave out the writing section. The justification for this stemmed from the fact that the new test was a reliable test and was based on a construct that tested students' academic literacy without the need for students to write a text. Also, the students are tested at the very beginning of their studies, just as they enter university. The focus at this early stage is to determine their general academic literacy abilities rather than focusing on one particular aspect. A final reason is that, when such writing tasks are reliably marked, these marks correlate well with the rest of the test yielding very little additional information.

With regards to TALPS, it was decided to include a section on argumentative writing. At postgraduate level it is essential that students follow specific academic writing conventions and it is important to test whether students were equipped with this knowledge. Butler (2009: 294) states: "In the development of TALPS we have also considered the importance of testing students' productive writing ability specifically (in the production of an authentic academic text), as well as their editing ability". In addition to the question on writing there is a question that tests students' editing skills. Below is an outline by Butler (2009) of the eight sections that appear in TALPS. He has also given a brief explanation of what aspect of academic literacy each tests:

> **Section 1** of TALPS is a scrambled text in which sentences in a paragraph have been scrambled, and students have to rearrange the sentences so that the paragraph forms a cohesive whole. It therefore tests not only students' ability in recognising text relations, drawing on their interpretative abilities regarding the context, but also their ability to recognise lexical clues contained in the sentences. Put differently: it assesses students' command of various grammatical features of the text.
>
> In **Section 2**, students' knowledge of general academic vocabulary is assessed. The context created for this section is specifically that of the postgraduate academic environment, and the words tested are a selection

of items from the different levels of the Coxhead academic word list (Coxhead, 2000).

Section 3 deals with visual and graphic literacy. Students are asked to interpret graphic information augmented by a short text discussion. This section mainly involves simple numerical computations and making inferences based on such calculations.

The **fourth section** emphasizes the importance of students being able to recognise different written text types. Students are requested to match two groups of sentences with regard to similarity in text type.

Section 5 includes a longer text that students have to read and subsequently answer comprehension type questions on the content of the text. Questions focus on students' abilities to classify and compare information, make inferences, recognise metaphorical language, recognise text relations and distinguish between essential and non-essential information.

Section 6 of the test assesses a number of academic literacy abilities. This question on grammar and text relations firstly provides students with a text they have to read where specific words have been omitted. Students then have to choose between 4 options regarding the place where these words have been left out in the sentences. The second part of the question requires that students, having been provided with the specific place where a word has been left out, choose between 4 options as to what is the correct word. The third part combines the formats of the first two in the sense that students are required to integrate the two tasks and do both simultaneously. They therefore have to find both the position where a word has been left out and the most suitable word that would fit that position. This section of the test assesses students' functional knowledge of sentence construction, word order, vocabulary, punctuation and at times communicative function (cf. Van Dyk & Weideman, 2004b), with the main focus on the former, i.e. on grammatical or structural features of the language.

In **Section 7**, students' grammatical knowledge of English is assessed in the sense that they have to edit a short paragraph in which a number of typical language errors occur.

The last section of the test (**Section 8**) provides students with the opportunity to produce a written academic text. Similar to TALL, the reading texts selected for use in TALPS are topical in the sense that they all relate to the same topic. Students are then required to make use of any information in the test on the topic and write an argumentative text of approximately 300 words in which they present a structured argument. The argument is within the context of Africa. They also need to ensure that they give due recognition to the sources used in the test that they choose to include in their argument (they have to include a short list of at least 2 sources at the end of their texts). They further have to ensure that

> the text adheres to generally accepted academic writing conventions (such as formality of register, logical structure, acknowledging sources, etc.) (Butler, 2009: 294).

2.8 Writing the Items

Writing the items for the test was shared between four members of staff of the UAL, University of Pretoria. TALPS was designed by a team of lecturers employed at the UAL, some of whom have considerable experience in the designing of tests of this nature (see project proposal for more info about this). Clearly, members of this team were equipped with the knowledge and experience to carry out this task. A schedule was drawn up and tasks were allocated to members of the team. The schedule included the initial target date as well as the date by when the task was completed. The team aimed to finalise the first draft by April 2007 and complete the first round of trials in May 2007 as indicated in the schedule below:

Table 2.8 Schedule of tasks and responsibilities

Initial target date	Task	Person responsible	Completed by	Comment
31 March	Vocabulary	Jurie	7 March	Done
	Text type	Avasha	26 April	Done
	Scrambled text	Avasha	26 April	Done
	Graphic and visual literacy	Gustav	7 March	Done
	Understanding texts	Albert	7 March	Done
	Grammar and text relations	Lynda	7 March	Done
	Editing task	Gustav	7 March	Done
	Writing: productive task	Gustav	—	Postponed until pilot is complete
17 April	Finalisation of first draft	Albert	2 May	Done

2.9 The process of development of TALPS

The process of the development of TALPS stretches over several drafts of varying length. These are listed below:

1. 1st draft = 173 items (150 minutes)
2. 2nd draft = 150 items (150 minutes)
3. 3rd draft = 100 items (120 minutes) (Piloted on UP EOT 110 students (May 2007)
4. 4th draft = 88 items (120 minutes) (Piloted on UP and UFS Postgraduate students (September 2007) + UFS Postgraduate students (February 2008)
5. Final draft = 76 items (120 minutes)

2.9.1 The first draft

The first draft of TALPS comprised of 173 items. The test was 150 minutes long and totalled 173 marks. It was made up of the following subtests:

Section 1 – Scrambled text (5 marks x 3)

This section included three sets of sentences that had to be re-arranged to create a paragraph. The final test would have only one set of sentences totaling 5 marks. The developers included three sets to test which of these worked better. The set working the best would be included in the final version of the test.

Section 2 – Interpreting graphs and visual information (16 marks)

There were 16 questions that made up this section. Each question carried 1 mark.

Section 3 – Dictionary definitions (5 marks)

This section comprised of 5 questions, each carrying 1 mark. The headwords used in this section were from Coxhead's academic wordlist.

Section 4 – Academic vocabulary (40 marks)

Once again, each question in this section carried 1 mark.

Section 5 – Text types (5 marks)

Here students were given two sets of sentences. The sentences are from different types of texts. Students had to match a sentence in the first set with a sentence from the second set. Each question carried 1 mark.

Section 6 – Understanding texts (60 marks)

In this section students had to answer questions based on a text. The total for this section was 60.

Section 7 – Grammar and text relations (22 marks)

In this section students were given three short paragraphs. In the first they had to indicate the **place** where the word was missing. This carried 3 marks. In the second paragraph they had to indicate the **word** that was missing. This carried 7 marks, and the last paragraph required them to indicate both the **place and the word** for 12 marks. In total this question carried 20 marks.

Section 8 – Text editing (10 marks)

The last section in this first draft of TALPS tests students' ability to edit a paragraph that contains 10 errors. Students had to locate the errors and then indicate what the correction was for that error. Each correction carried 1 mark.

At this stage the test did not include a question requiring students to write an argumentative text. The concern of the developers at this early stage was in writing the multiple-choice questions. These would then be analysed using TiaPlus Test and Item Analysis (Cito, 2006) to determine which items did or did not test well. Items that did not test well were discarded.

2.9.2 The second draft

The second draft of TALPS was 150 marks and 120 minutes long. Changes were made in the following sections:

Section 1 – This section now included only one set of sentences. The total marks remained the same.

Section 4 – 27 questions were retained, each carrying one mark. This section now carried 27 marks.

2.9.3 The third draft

The aims set out in the schedule above were met and the third draft version of the test became the first pilot for TALPS. This first pilot was completed in May 2007 with first year students at the University of Pretoria, since no other students were available at the time. These were students who were taking the Academic Literacy (EOT 110) module. These results were measured using TiaPlus Test and Item Analysis (Cito, 2006) which provides measures at item and test level. Before the first pilot, by which time the test had been reduced to 100 items, items were evaluated by the designers to determine the appropriateness/strength of the item. Most changes were made in the **Understanding texts** section. In the 100-item test this section had 45 items; in the 88-item test (which was the second pilot) it had 33 items and then 28 items. The final version of the test has 21 items in this section. Justification for this decision was drawn from the analyses done using the TiaPlus Test and Item Analysis Build 300 (Cito, 2006). According to this, seven items had very high p-values, meaning that a high percentage of the test population got this answer correct.

2.10 Piloting the test

2.10.1 The first pilot

The first pilot of TALPS had 100 items and four sections: **Dictionary definitions**, **Academic vocabulary**, **Understanding texts and Grammar and text relations**. The test was piloted on first year students who were registered for the compulsory academic literacy course (EOT 110/120). Students were given one and a half hours to complete the test. The test totalled 100 marks. It did not include the question requiring students to write an argumentative text.

Table 2.9 Descriptive statistics of the first pilot of TALPS (UP students)

	EOT 110 Students
N	1488
Number of selected items	100
Minimum test score	0
Mean/average *P*- value	58.28
Standard deviation	11.10
Cronbach's alpha	0.85
GLB	0.92
Standard error of measurement	4.30
Average *Rit*	0.25
Cut-off point	32.5
Percentage failed	1.28

Table 2.10 Descriptive statistics of subtests of the first pilot of TALPS

	Dictionary Definitions	Academic Vocabulary	Understanding Texts	Grammar and Text Relations
N	1488	1488	1488	1488
Number of selected items	5	28	45	22
Minimum test score	5	0	0	0
Mean, (average P-value)	3.34 (66.80)	18.72 (66.86)	23.16 (51.48)	23.16 (51.48)
Standard deviation	1.07	3.44	5.50	5.50
Cronbach's alpha	0.36	0.61	0.72	0.72
GLB	0.42	0.70	0.83	0.83
Standard error of measurement	0.85	2.16	2.89	2.89
Average Rit	0.53	0.30	0.28	0.28

2.10.2 The second pilot

The second pilot of TALPS was carried out on postgraduate students both at the University of Pretoria and the University of the Free State in September 2007. This test comprised eighty-eight items and totalled 120 marks. It included the following sections:

Section 1 – Scrambled text (5 marks)
Section 2 - Interpreting graphs and visual information (10 marks)
Section 3 – Dictionary definitions – (5 marks)
Section 4 – Academic vocabulary – (10 marks)
Section 5 – Text types (5 marks)
Section 6 – Understanding texts (40 marks)
Section 7 – Grammar and text relations (15 marks)
Section 8 – Text editing (10 marks)
Section 9 – Academic writing (20)

Below is an analysis of the results:

Table 2.11 Descriptive statistics of the second pilot of TALPS (UP and UFS Students)

	UP and UFS students
N	117
Number of selected items	88
Minimum test score	0
Mean/average *P*- value	61.33 (69.70)
Standard deviation	14.19
Cronbach's alpha	0.93
GLB	1.00
Standard error of measurement	3.82
Average *Rit*	0.37
Cut-off point	28.5
Percentage failed	0.85

Table 2.12 Descriptive statistics of subtests of the second pilot of TALPS

	Scrambled Text	Academic Vocabulary	Dictionary definitions	Under-standing Texts	Text Editing	Interpreting graphs and visual information	Grammar and Text Relations
N	117	117	117	117	117	117	117
Number of selected items	5	10	5	33	10	10	15
Minimum test score	0	0	0	0	0	0	0
Mean/average *P*- value	2.66 (53.16)	7.60 (75.98)	4.19 (83.76)	23.75 (71.98)	7.62 (76.15)	6.56 (65.64)	8.96 (59.72)
Standard deviation	1.92	1.83	0.93	4.89	2.85	2.89	4.52
Cronbach's alpha	0.85	0.54	0.35	0.78	0.88	0.82	0.89
GLB	0.94	0.74	0.41	0.94	0.92	0.89	0.96
Standard error of measurement	0.75	1.24	0.75	2.30	1.00	1.22	1.51
Average *Rit*	0.79	0.45	0.53	0.36	0.70	0.63	0.64

The TALPS second pilot was also carried out on a second batch of students from the University of the Free State.

Table 2.13 Descriptive statistics of the second pilot of TALPS (2nd batch of UFS students)

	UFS Students
N	112
Number of selected items	88
Minimum test score	0
Mean/average *P*- value	57.38 (65.21)
Standard deviation	15.13
Cronbach's alpha	0.93
GLB	1.00
Standard error of measurement	3.87
Average *Rit*	0.40
Cut-off point	28.5
Percentage failed	1.79

Table 2.14 Descriptive statistics of subtests of the second pilot of TALPS

	Scrambled Text	Interpreting Graphs	Dictionary Definitions	Academic Vocabulary	Text Types	Under-standing Texts	Grammar and Text relations	Text editing
N	112	112	112	112	112	112	112	112
Number of selected items	5	10	5	10	5	28	15	10
Minimum test score	0	0	0	0	0	0	0	0
Mean/average (*P*- value)	2.16 (43.21)	7.19 (71.88)	4.23 (84.64)	7.38 (73.84)	2.40 (48.04)	20.56 (73.44)	7.08 (47.20)	6.38 (63.75)
Standard deviation	1.84	2.26	1.08	2.03	1.67	4.22	4.40	3.19)
Cronbach's alpha	0.81	0.69	0.57	0.64	0.75	0.75	0.88	0.95
GLB	0.90	0.82	0.67	0.78	0.85	0.95	0.96	0.98
Standard error of measurement	0.81	1.26	0.71	1.21	0.83	2.11	1.55	0.91
Average *Rit*	0.76	0.52	0.61	0.50	0.71	0.36	0.62	0.85

2.10.3 The TALPS final draft version

The final draft version of TALPS was made up of seventy-six items and eight sections:

Section 1 – Scrambled text – (5 marks)

Section 2 – Interpreting graphs and visual information – (10 marks)

Section 3 – Academic vocabulary – (10 marks)

Section 4 – Text types – (5 marks)

Section 5 – Understanding texts – (25 marks)

Section 6 – Grammar and text relations – (15 marks)

Section 7 – Text editing – (10 marks)

Section 8 – Academic writing – (20 marks)

This version of the test totalled 100 marks. The section on **Dictionary definitions** was left out of this version of the test. As can be seen from the descriptive statistics of the drafts of TALPS, the **Dictionary definitions** question had p-values of 84.2 for both the 88 item pilots. Davies *et al.* (1999) explain that the higher the index, the easier the item. The closer the index is to 100% or 0%, the less differential information it can provide about candidates. They state that items that are excessively easy or very difficult are normally removed because they do not contribute to the test's discriminability (1999: 95). The pilot for this version of the test was carried out in September 2007 on two groups of students: postgraduate students from the North-West University and postgraduate students from the University of Pretoria. Below is an analysis of the results of the test:

Table 2.15 Descriptive statistics of the TALPS final draft version (NWU)

	TALPS
	NWU Postgraduate Students
N	175
Number of selected items	76
Minimum test score	0
Mean/average *P*- value	52.09 (65.11)
Standard deviation	12.86
Cronbach's alpha	0.91
GLB	1.00
Standard error of measurement	3.83
Average *Rit*	0.39
Cut-off point	32.5
Percentage failed	9.14

Table 2.16 Descriptive statistics of the TALPS final draft version (UP)

	TALPS
	UP Postgraduate Students
N	97
Number of selected items	76
Minimum test score	0
Mean/average *P*- value	51.48 (64.36)
Standard deviation	14.10
Cronbach's alpha	0.93
GLB	1.00
Standard error of measurement	3.80
Average *Rit*	0.42
Cut-off point	32.5
Percentage failed	12.37

Table 2.17 Descriptive statistics of the TALPS final draft version (NWU &UP)

	TALPS COMBINED
	NWEST &UP Postgraduate Students
N	272
Number of selected items	76
Minimum test score	0
Mean/average *P*- value	51.88 (64.84)
Standard deviation	13.32
Cronbach's alpha	0.92
GLB	0.99
Standard error of measurement	3.84
Average *Rit*	0.40
Cut-off point	32.5
Percentage failed	10.29

Clearly, as can be deduced from the tables above, TALPS is a highly reliable and valid test, ready to be used for the purpose for which it was designed – to test the academic literacy of postgraduate students.

2.11 Conclusion

A fair and responsible test is one that is valid and reliable but socially acceptable as well. While the focus of this study is a concern with the social dimension of testing, one cannot completely ignore any part of the testing process. Ignoring the empirical analyses of a test would be irresponsible – in the same way that the literature on language testing speaks volumes about irresponsible test developers/users who have ignored the social aspects. In order to effectively document the process of the development of TALPS, it is important to consider these facts. It should be the aim of test developers to design tests that are valid, reliable, accessible and transparent, by test developers who are willing to be accountable for their designs. This chapter has focused on the design and development of TALPS. The next part of this story, which is a discussion of the validity and the reliability measures of the test and takes the form of a validation argument, will be completed in chapter 4. Before that, chapter 3 will explore the theoretical framework that underlies this study and will include a detailed discussion of the validity of TALPS. The following chapters (5, 6 and 7) will focus respectively on the issues of transparency, accessibility and accountability. This is in keeping with the framework employed for this study that enables us to provide interested parties with a comprehensive analysis of TALPS.

Chapter 3

A theoretical framework for understanding foundational concepts in language testing

3.1 Introduction

In defining the characteristics of all tests, Davies (1990: 17) states that a test "is intended above all to clarify the difference in the matter under test, in what is being tested (proficiency, aptitude, achievement) among the candidates". This need to "clarify the difference" means, among other things, that comparative figures and data need to be studied. In the field of language testing, this has led to a heavy reliance of applied linguistics on the field of psychometrics. In fact, according to McNamara and Roever (2006: 1), "…psychometrics became the substrate discipline, prescribing the rules of measurement, and language was virtually poured into these pre-existing psychometric forms." Thus, psychometrics became the basis of language testing, the most important and for some the only way in which a test could be validated. However, language tests need more than psychometric data to be considered valid, not least because "language is rooted in social life and nowhere is this more apparent than in the ways in which knowledge of language is assessed" (McNamara & Roever, 2006: xiv). McNamara and Roever's observation that "a psychometrically good test is not necessarily a socially good test" (2006: 2) is an important one to this study, because a core concern here is the social responsibility that the test developers of TALPS have, not just to the test takers (postgraduate students) but to everyone affected by the test – supervisors, parents, test administrators and society at large.

This study is primarily concerned with transparency, accessibility and accountability as regulative conditions for a postgraduate test of academic literacy. It proposes to investigate how these can be incorporated in the design of one test (TALPS), and theoretically accounted for in terms of a framework. This chapter will show that questions posed by this study cannot be adequately

answered with reference to the conventionally acknowledged experts in the field like Messick (1989a), Bachman and Palmer (1996), and Kunnan (2004). Instead these questions can be answered in a "third idea, other than validity and usefulness" (Weideman, 2009a: 239), as outlined by Weideman, an idea that does not foreground one concept but rather identifies a number of important considerations for language testing. In this chapter, an attempt will be made to do the following:

- Discuss the need for a theoretical analysis or justification for applied linguistic designs;
- Provide a definition of and further explanation for the terms constitutive and regulative;
- Investigate how a number of important considerations for language testing can be incorporated within a theoretical framework for language testing;
- Provide a detailed discussion of the concept of validity in language testing, with specific emphasis on the distinctions developed by Messick (1980; 1981; 1989a; 1989b), Bachman and Palmer (1996), and Kunnan (2004).

3.2 The need for a theoretical analysis or justification for applied linguistic designs

It is useful to begin by acknowledging that the field of language testing falls within the scope of applied linguistics. Weideman defines applied linguistics "as a discipline that devises solutions to language problems" (2006a: 72). In this view applied linguistics presents the solution in the form of a design or plan, which is in turn informed by some kind of theoretical analysis or justification. It is this "theoretical analysis or justification" (2006a: 72) that is firstly of concern here. The need for a theoretical analysis or justification is outlined by Weideman in a paper entitled *A responsible agenda for applied linguistics: Confessions of a philosopher* (2007b). He observes here that while

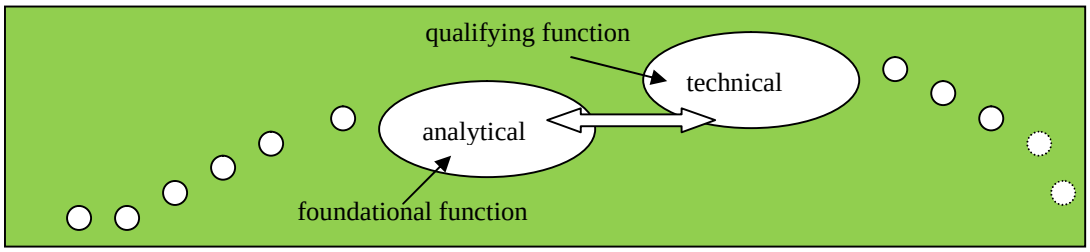
applied linguistic work can be so absorbing that very often one does not “take the time to stand back and take stock, or understand fully the disciplinary foundations we stand on” (2007b: 30), this ignorance quite easily sets one up for falling victim to theoretical fashions (2007b: 30). If we see applied linguistics as a discipline concerned with design, we must also “develop a theory of applied linguistics which shows what constitutive and regulative conditions exist for doing applied linguistics designs” (Weideman, 2007b: 29). Weideman explains that this theory or “agenda” (2007b: 29) needs a “certain yardstick” (2007b: 29). He has chosen the term ‘responsible’ as a measure of this. In addition to this, the discipline of applied linguistics needs a “broader theoretical framework” (2007b: 29), by which he means that one-sided emphases (e.g. on purely empirical determinations of validity) create undesirable design considerations.

The theoretical framework he is referring to therefore allows the articulation of “a responsible agenda for applied linguistics” (2007b: 30). The need for ‘responsibility’ or to “work with integrity” (2007b: 30) that is being referred to here is essential to this thesis in view of the fact that the issues raised here (transparency, accessibility and accountability) deal directly with the conceptualisation of responsibility, integrity and fairness in the field of language testing. The theoretical framework that will underlie this thesis will be drawn from that outlined by Weideman in the papers (2006a; 2007b; 2009a) which were referred to earlier.

The solutions to language problems that applied linguists design or devise are presented in the form of designs or plans – these could be the designing of language courses or, in the case of this thesis, the development of language tests (Weideman 2006a: 72). According to Weideman’s theoretical framework, the plan presented has two terminal functions: “a qualifying or leading function, and a foundational or basis function” (2006a: 72).

Presented schematically:

Figure 3.1 Leading and foundational functions of applied linguistic designs



(Weideman, 2006a: 72)

According to this schematic representation, the leading or qualifying function of a plan presented as an applied linguistic solution to a language problem is to be found in the technical aspect of design. The plan or design finds it foundational function in, or is based upon, the analytical or theoretical mode of experience. Explained simply, this theoretical framework suggests that:

a. The theory provides a rationale for the design but does not control it; and
b. The design therefore takes precedence, not the theory. While the theory is important it does not "prescribe" (Weideman, 2007b: 41) the design.

The context for this argument is provided by arguments against treating technically qualified objects, such as language tests, as mere applications of science. This does not mean that the technical design and development of an applied linguistic instrument, such as a language test, has nothing to do with scientific analysis, but that the role of scientific analysis is to provide subsequent theoretical justification for the design.

Weideman explains that "the context in which such a designed solution is implemented invariably has a social dimension, and that applied linguistic designs have ethical dimensions, since they affect the lives of a growing number of people" (2006a: 72). The following observations are relevant:

a. Applied linguistic work should be backed by some foundational framework to ensure that the notions of

responsibility and integrity can be articulated in a theoretically coherent and systematic way;

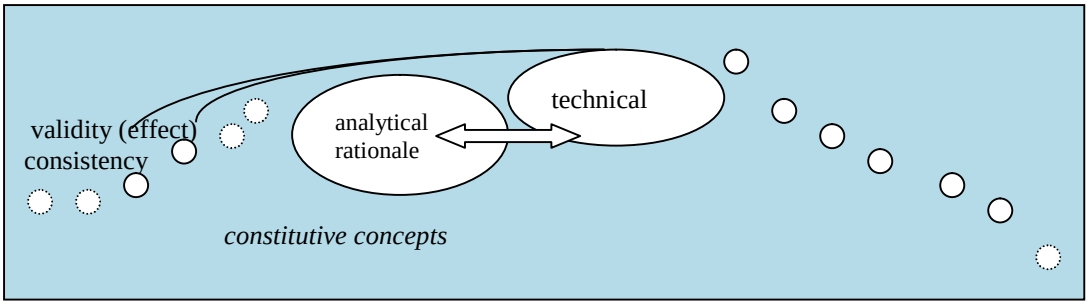
b. In the framework Weideman suggests that the plan or design has a leading or qualifying function and a foundational or analytical function;
c. The design cannot ignore the social and ethical dimensions present in the articulation of solutions to language problems.

3.3 Defining 'constitutive' and 'regulative'

The employment of the theoretical framework in this study serves to articulate coherently and systematically issues of responsibility and integrity – as well as to make allowances for other dimensions, such as the social and the ethical. This theoretical foundation is derived from a specific way of looking at the world. Not only is applied linguistics conceived of as a discipline of design, but the outcomes of such designs are characterised as technical objects by the philosophical framework being employed. Just as we do not exist in isolation, technical products such as tests do not exist on their own. They exist in the technical as well as in other modes of experience – they are not removed from but related to these other modes. Weideman states: "The conviction is a fairly simple one: nothing is absolute, and …, though one may distinguish between uniquely different modes of doing and being, all of these are connected to everything else" (2007a: 599). The technical dimension, in fact, coheres with a number of other dimensions of our existence. These relations yield two sets of concepts and ideas:

1. A set of **'constitutive'** concepts – defined here as "grounded upon". The technical design is grounded upon a set of concepts such as reliability/validity, and these are founding or constitutive concepts, as indicated in the figure below:

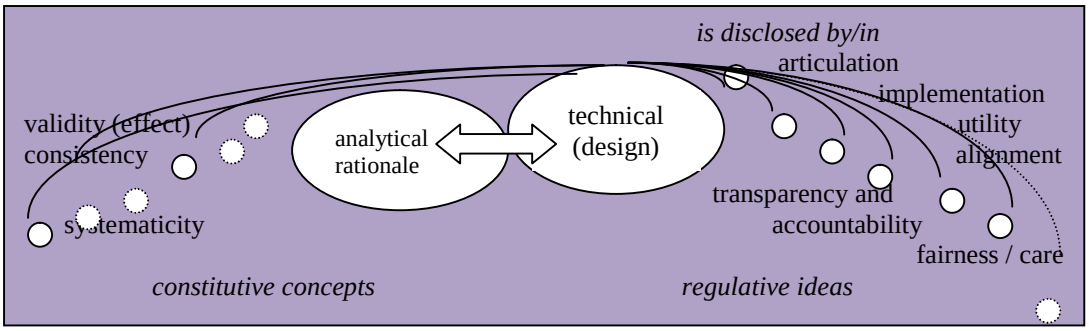
Figure 3.2 Constitutive concepts in applied linguistics



(Weideman, 2007b: 42)

2. The technical mode that qualifies the applied linguistic instrument that a language test constitutes does not exist on its own, but is guided by/led also by a set of **regulative** ideas – what can be defined as leading or guiding conditions**.** The regulative conditions ensure that issues related, for example, to the social, are anticipated in the design of the test as indicated in the figure below:

Figure 3.3 Constitutive concepts and regulative ideas in applied linguistic designs



(Weideman, 2007b: 44)

The theory does not "dictate or prescribe" (Weideman, 2007a: 599) the design but is used to provide a rationale for it. Equally important is the concern here for the social and ethical dimensions related to the "designed solution" (2007: 599) and that "the solution when implemented must also have ethical (and other) dimensions, i.e. must be transparent, accountable, theoretically and politically defensible and promote the interests of those affected by it" (2007a: 599). According to Weideman, validity and reliability are constitutive concepts

for the characterisation of the applied linguistics artefact, e.g. a language test or a course design. A test must do what it is designed to do and it must also be consistent. While these are the "necessary conditions for its design" (2007a: 602) they do not function in isolation but in harmony or accordance with other factors, qualities or modes such as the lingual, the social, economic, aesthetic, juridical and ethical dimensions of reality, and the way that these are reflected in concepts and ideas such as, respectively, the technical interpretability of the scores/outcomes of the test, the implementation of the test, its technical utility, alignment with needs of students and administrators, transparency, accountability and fairness.

3.4 Fundamental concepts in language testing

The framework employed in this study is based on a "representation of the relationship among a select number of fundamental concepts in language testing" (Weideman, 2009a: 241). The two main functions that have already been identified above are the technical mode and the analytical dimension. These do not function in isolation. The relation between the two is 'reciprocal' (Weideman, 2009a: 244). The technical mode interacts not only with the analytical mode, however, but is also connected with all other modes, as can be seen in Table 3.1 below. Weideman points out that the technical unity of multiple sources of evidence, the reliability of a test, its validity and its rational justification are foundational or constitutive applied linguistic concepts (2009a: 247). These may also be designated necessary requirements for tests (Weideman, 2009a: 247). Important is the fact that "each of these 'necessary' or foundational concepts yields a (technically stamped) criterion or condition for the responsible use or implementation of the technical instrument" (Weideman, 2009a: 247). This, according to Weideman, is why we say that tests should be reliable, valid and built on a theoretical base that is defensible in terms of a unity within a multiplicity of sources of evidence (Weideman, 2009a: 247) as opposed to focusing specifically only on one concept, such as validity.

This technical dimension of the applied linguistic design also links with the lingual, social, economic, aesthetic, juridical and ethical aspects. According to Weideman, the links between the technical, qualifying function of the test design and other aspects yield the ideas of technical articulation, test implementation or use, technical utility, the alignment the test has with learning and teaching language, its public defensibility or accountability, and its fairness or care for those taking the test (Weideman, 2009a: 247).

The theoretical foundation or framework being utilised in this study can be understood more easily if viewed in the form of a table:

Table 3.1 Constitutive and regulative moments in applied linguistic designs

Applied linguistic design	**Aspect / function / dimension / mode of experience**	**Kind of function**	**Retrocipatory / anticipatory moment**
is founded upon	numerical	constitutive	unity within a multiplicity of sets of evidence and conditions for (test) design
	kinematic		internal consistency (technical reliability)
	physical		internal effect / power (validity)
	organic		technical differentiation
	feeling		technical perception and intention
	analytical	foundational	design rationale (construct validity or theoretical defensibility)
is qualified by	technical	qualifying / leading function (of the design)	
is disclosed by	lingual	regulative	articulation of design in a blueprint / plan
	social		implementation / administration
	economic		technical utility, frugality
	aesthetic		harmonisation of conflicts, resolving misalignment
	juridical		transparency, public defensibility, fairness, legitimacy
	ethical		accountability, care, service

(Weideman, 2007a: 602)

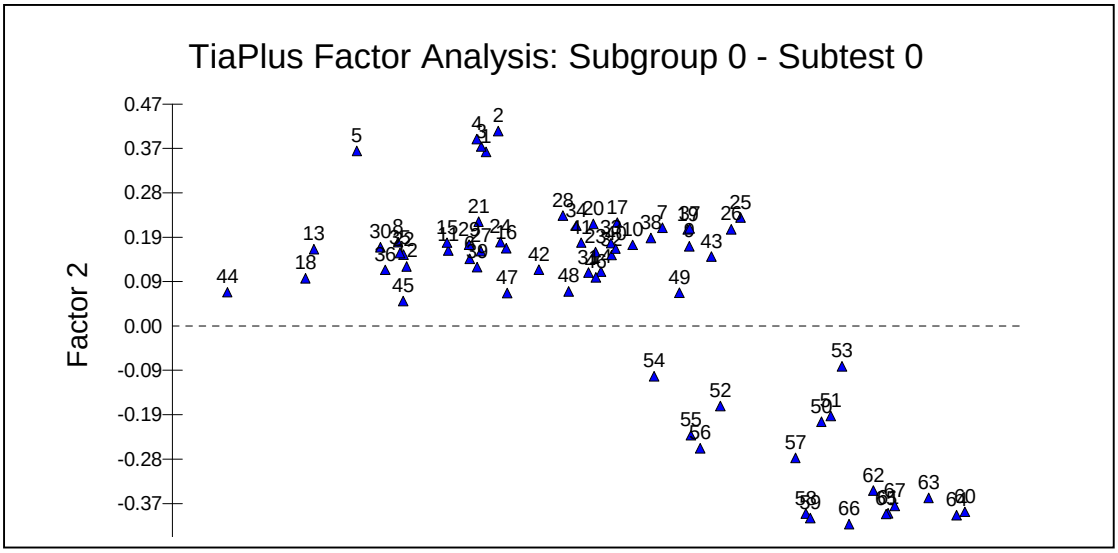
The table and the theoretical framework it articulates seem to suggest that if conditions such as consistency, validity, theoretical and social defensibility, transparency, accountability and fairness, are anticipated in the design of a test, then that test will fulfil the requirements of being a (psychometrically and socially) good test. It also suggests a move away from earlier beliefs about language testing – one view being purely asocial and psychometric, and, in the other, a move away from using assessment instruments to answer questions about the social aspects related to testing.

Weideman has pointed out that consideration of issues like transparency, accountability and fairness stem ultimately from "the love that we show for humanity" (2007a: 602). The use of the word 'love' may seem quite out of context in a discussion about language tests, something also pointed out to Weideman by an anonymous referee for this article. Yet it should be exactly this that should drive or motivate the work one does – for it is this that will force one to consider the effects and consequences of one's designs. This will ensure the design, development and use of fair and transparent language tests. Weideman's concern for the service that we perform for humanity is clear when, in arguing that applied linguists become accountable for their designs, he states that these designs or solutions to language problems should "relieve some of the suffering, pain, poverty, and injustice in our world" (2007b: 29). This concern for humanity, present also in the framework that he presents for test design and development, is the motivation behind choosing this framework over that of Messick (1989a), Bachman and Palmer (1996) and Kunnan (2004). In putting forward their model of test usefulness, Bachman and Palmer, for example, state that they regard their model as the essential basis for quality control throughout the entire test development process: "We would further argue that all test development and use should be informed by a model of test usefulness" (1996: 17). This model of test usefulness, though addressing issues largely ignored in the past and definitely a progressive move to a more unified view of language testing, still does not, however, provide answers to the key questions in this thesis, questions concerning the transparency, accessibility and accountability of test developers and the test.

The framework that will be employed for this study, unlike those developed by Messick (1989a), Bachman and Palmer (1996) and Kunnan (2004), does not subsume everything under one concept. What it does is highlight a number of important concepts in language testing. This view allows for a more open and flexible way of designing and using tests rather than the restriction of an "overarching or unified" (Weideman, 2009a: 239) concept. Such a view also

allows for what Weideman refers to as “trade-offs” (2006a; 2007b; 2009a). Weideman cites an example of such a trade-off. This had to with a decision that developers of TALL were faced with, a choice that developers had to make between the “technical consistency and the appropriateness or relevance of the tests” (Weideman, 2009a: 237). According to Weideman tests conventionally are considered adequate if they test a single ability. If the test measures more than one ability, this will show up in a factor analysis. The 2008 version of TALL, while still a reliable test, did show up some ‘outlying’ items as indicated in the figure below:

Figure 3.4 Measures of homogeneity and heterogeneity in TALL 2008



(Weideman, 2009a: 237)

Weideman explains that items further away from the zero line are “less closely associated with the measurement of a single factor” (2009a: 237). The designers could choose to leave out these items (1–5 and 50–67) to “enhance the technical consistency (reliability)”, or accept the ‘heterogeneity’ (Weideman, 2009a: 237) of the construct that is being tested. They chose to include the items, arguing that academic literacy is a “richly varied and potentially complex” (2009a: 237) ability and one would therefore have to “tolerate a more heterogeneous construct” (2009a: 237). The trade-off here was

between the technical consistency or reliability and the appropriateness of the construct. Tests cannot be developed and used in a vacuum. Because of the implications of the uses of test scores, every effort should be made to ensure that tests are valid, reliable, fair and accessible to every test-taker. This becomes easier when working within a framework that allows for considerations such as trade-offs.

Employing a framework that incorporates a concern for the empirical analyses of a test, as well as a concern for the social dimensions of language testing means that one is forced to consider important questions related to every aspect of the test: the validity and reliability of the test, the reason for giving the test, the effect of the test on the test-taker, concerns about the design of fair tests, the rights and responsibilities of the test designer and the rights and responsibilities of the test-taker. The value of this framework, according to Weideman, "lies in its separating out what is conceptually distinct, and, by so doing, enriching our theoretical understanding of the constitutive and regulative, necessary and sufficient conditions for language testing" (2009a: 249). An important question at this point would then be: What is there in the literature on language testing to help test developers design and administer tests that are valid, reliable as well as socially responsible and transparent?

3.5 The concept of 'validity' in language testing

One would be forgiven for assuming that all questions find their answers in the concept of validity, for it is the concept of validity that seems to dominate the literature on language testing. In their book *Language testing and assessment* (2007), Fulcher and Davidson state that validity is today the "central concept in testing and assessment" (2007: 3). The conventional definition of validity as a quality of a test when it measures what it is designed to measure is no longer necessarily acceptable to all experts in the field. The term has undergone different interpretations, and the two different perspectives on the concept of validity can be divided into a traditional view and the current and more widely

accepted view. Traditionalists see validity as measuring what the test is designed to measure (cf. Borsboom, Mellenbergh & Van Heerden, 2004). They see validity as a property of a test. Van der Walt and Steyn (2007: 139) explain that the traditionalists consider validity "to be an inherent attribute or characteristic of a test, that a psychologically real construct or attribute exists in the minds of the test taker – this implies that if something does not exist, it cannot be measured" (2007: 139). This view of validity has conventionally identified the three different types of validity: criterion-related validity, content-related validity and construct validity.

The current view of validity does not separate the three types of validity identified above. In this view, construct validity is the central component, and the other two are aspects of it. Cronbach, who can be considered the 'father' of construct validity, together with Meel coined the term "construct validity" and defined it as follows:

> Construct validity is ordinarily studied when the tester has no definite criterion measure of the quality with which he is concerned and must use indirect measures. Here the trait or quality underlying the test is of central importance, rather than either the test behaviour or the scores on the criteria (Cronbach & Meel, 1955: 283).

The Cronbach and Meel (1955) article on construct validity heralded a major change in validation research (McNamara & Roever, 2006: 10), highlighting the possibility that validity may not be a "purely mathematical property like discrimination or reliability but rather a matter of judgement" (McNamara & Roever, 2006: 10). According to McNamara and Roever (2006: 11) validity is concerned with making inferences about the ability of the test takers being tested, based on their scores. Cronbach, in fact, emphasised the need for a validity argument that focused on collecting evidence for or against a certain interpretation of the test scores (McNamara & Roever, 2006: 10). However, it is only much later in his most influential writings on validation within measurement, according to McNamara and Roever (2006: 10), that Cronbach stressed the importance of the socio-political context, its influence on the field

of testing and the "role of beliefs and values in validity arguments which must link concepts, evidence, social and personal consequences and values" (McNamara & Roever, 2006: 11). In their discussion of this, McNamara and Roever point out that Cronbach also agreed with Messick that validity work has an obligation to consider test consequences and to help prevent negative ones. What we have here, according to McNamara and Roever (2006: 11), is a concern for social consequences as a kind of corrective to an earlier entirely cognitive and individualistic way of thinking about tests. As positive a move as this was seen to be, McNamara and Roever contend that there was "still difficulty in integrating his psychometrically inspired work on construct validity and his concern for social and political values" (2006: 12).

3.5.1 Messick on validity

What Cronbach began, in terms of beginning to reconceptualise validity by observing that it is limited to the technical meaning and interpretation of test scores (objective, technical results of the measurements), came to fruition in the highly influential work of Messick. Messick's article on validity, published in *Educational Measurement* (1989b) is still considered one of the most significant writings in the field. It is a work that has been and still is widely quoted by experts in the field of language testing. The article, simply entitled 'Validity', constitutes a major change in the way validity, and more specifically, construct validity began to be understood. In the opening lines to this article he states that

> validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment (Messick, 1989b: 13).

He explains that validity is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use, stating that "what is to be validated is not the test or observation device as such, but the inferences derived from test scores or other indicators" (Messick, 1989b: 13).

Messick stresses that validity is a matter of degree**,** not an all or nothing measure or property (1989b: 13). Equally important is his contention that validity is an ongoing process, an "evolving property and a continuing process" (Messick, 1989b: 13).

A key point that Messick raises in this article is the view of validity as a "unitary concept" (Messick, 1989b: 13). Since as early as the 1950's, test validity has conventionally been broken up into three types: content validity, predictive and concurrent criterion-related validity, and construct validity. Kurpius and Stafford (2006: 147) explain that content validity comes from two approaches: an empirical analysis of how well the test items reflect the content domain, and expert ratings of the relationship between the test items and the content domain. Items included in the test that do not cover the course content contribute to construct-irrelevance. If the purpose of a test is to predict some future behaviour or to establish current behaviour, evidence that the test items will do this accurately is sought. The relationship between the test scores and the variables external to the test (criterion) will provide this source of evidence for criterion-related validity (Kurpius & Stafford, 2006: 147). They explain that if the goal is for test scores to predict future behaviour, the concern is with predictive validity. Kurpius and Stafford (2006: 150) define construct validity as the degree to which all items on a test are interrelated and measure the theoretical trait/construct the test is designed to measure.

For Messick, however, content validity and predictive/concurrent criterion-related validity do not qualify to "bear the name 'validity' and to wear the mantle of all that name implies" (Messick, 1980: 1015). He justifies this by stating that because content validity provides "judgemental evidence in support of the domain relevance and representativeness of the content of the test instrument, rather than evidence in support of inferences to be made from test score" (Messick, 1989b: 17), it is not validity at all. He dismisses criterion-related validity as a type of validity because it

> relies on selected parts of the test's external structure. The interest is not in the pattern of relationships of the test scores with other measures generally, but rather is more narrowly pointed towards selected relationships with measures that are criterial for a particular applied purpose in a specific applied setting (Messick, 1989b: 17).

But while he dismisses content validity and criterion-related validity as types of validity, he does not dismiss the value of the evidence they provide. Instead he suggests "the use of labels more descriptive of the character and intent of each aspect" (Messick, 1980: 1014). The labels he suggests appear below:

Table 3.2 Alternative descriptors for aspects of test validity

Validity designation	Descriptive designation
Content validity	Content relevance-domain specifications
	Content coverage-domain representativeness
Criterion validity	Criterion relatedness
Predictive validity	Predictive utility
Concurrent validity	Diagnostic utility
	Substitutability

(Messick, 1980: 1015)

For Messick the compartmentalising of validity into different types "leads to confusion and, in the face of confusion, oversimplification" (Messick, 1980: 1014). A consequence of this, according to Messick, is the assumption on the part of test users that any one type of validity would do, so that once evidence of one type of validity is forthcoming, one is relieved of responsibility for further inquiry (1980: 1014). The point Messick makes is that there are not different types of validity, but different kinds of evidence, and that the points associated with each of these terms are important ones, but that their distinctiveness is blurred by calling them all 'validity'. A worse consequence, as indicated earlier, would be the belief on the part of test users that any one type of validity would be sufficient to validate a test. As stated earlier, Messick

does not dismiss the value of the evidence that content-validity and criterion-related validity provide, stating that this evidence does contribute to score meaning.

However, he sees construct validity as based "on an integration of any evidence that bears on the interpretation or meaning of the test scores" (Messick 1989b: 17). Messick states that

> construct validity is indeed the unifying concept of validity that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships. The bridge or unifying theme that permits this integration is the meaningfulness or interpretability of the test scores, which is the goal of the construct validation process (Messick, 1980: 1015).

Messick's view on validity is, according to Van der Walt and Steyn (2007), "a more naturalistic, interpretative one" (2007: 139). Messick, like Cronbach before him,

> saw assessment as a process of reasoning and evidence gathering carried out in order for inferences to be made about individuals and saw the task of establishing the meaningfulness and defensibility of those inferences as being the primary task of assessment development and research (McNamara & Roever, 2006: 12).

Messick introduces the social dimension into this picture by arguing two issues:

> That our conceptions of what it is that we are measuring and the things we prioritise in measurement, will reflect values, which we can assume will be social and cultural in origin, and that tests have real effects in the educational and social contexts in which they are used and that these need to be matters of concern for those responsible for the test (McNamara & Roever, 2006: 13).

Messick used a, by now well-known, matrix to summarise his theory on validity:

Table 3.3 Facets of validity as a progressive matrix

	TEST INTERPRETATION	**TEST USE**
EVIDENTIAL BIAS	Construct validity (CV)	CV + Relevance/Utility (R/U)
CONSEQUENTIAL BIAS	CV + Value Implications (VI)	CV + R/U + VI + Social Consequences

(Messick, 1989a: 10)

Messick explains this matrix by stating that the

> validity of test interpretation and test use, as well as the evidence and consequences bearing theron, are treated here in the unified faceted framework …, because of a conviction that the commonality afforded by construct validity will in the long run prove more powerful and integrative than any operative distinctions among the facets (Messick, 1989b: 20).

Construct validity is to Messick the "integrating force that unifies validity issues into a unitary concept" (1989a: 10), it "binds the validity of test use to the validity of test interpretation" (1989a: 10) and "binds social consequences of testing to the evidential basis of test interpretation and use" (1989a: 10). Two important questions arise at this point: a) What implications does this matrix have for test developers, and, b) Does Messick's view of construct validity answer important questions in this study, i.e. whether construct and other conventional forms of validity are enough to validate a test and whether issues of accessibility, transparency and accountability can be anticipated in the design and use of a test?

Messick has long been accepted as the expert on construct validity. His (1989b) article has been called the "most cited authorative reference" (Shepard, 1993: 423) on the topic of validity. Yet a close reading of the literature on validity, construct validity and language testing reveals that Messick's views have not been unquestioningly accepted by all experts in the field. Concerns and questions are addressed carefully, probably for fear of upsetting the applecart on which has carefully been placed the concept of construct validity. The reasoning behind this could be, according to Weideman, "the massive influence

of the views of Messick and the institutional base that he represented" (Weideman, 2009a: 239).

Whatever the reason, Messick's views do raise a number of concerns for this study. A first concern lies in the challenging and complex nature of Messick's work. Today the work of developing tests no longer lies in the hands of psychometrists and measurement specialists alone. Taylor (2009: 22) points out that there are growing numbers of people involved in selecting or developing tests and they often find themselves doing this without much background or training in assessment. The present day emphasis on the importance of testing means that many professionals in a variety of fields have to play the role of test developer – such as the language teacher who wants to design a test to test the writing levels of her class, but has no formal training in designing tests. Her first step then would be to consult the literature available on the designing of language tests – leaving her with the daunting task of unravelling Messick's concept of validity. Despite the availability of all of Messick's writing, McNamara and Roever still write about the need to make language test development and validation work "more manageable" (2006: 33) while Shepard refers to the "complexity of Messick's analysis" (1993: 427). While Messick has made an influential contribution to the field of testing, his work is not easily accessible to the lay person who needs to understand the field of testing, nor does he present us with a framework or guidelines to assist in the designing of tests that are accessible and transparent. Why then the huge influence of Messick in the field of testing? The main reason probably is that language testing was for a long time focused on psychometrics, and that Messick's predecessors worked firmly within that tradition. Could it be that Messick's consideration of the social consequences in testing came at exactly the time when the field needed such a change? Is it possible that Messick's theory was so readily accepted and became so influential because it seemed as if he was offering a new way of looking at and evaluating/assessing tests, one that included a consideration of the consequences of the uses of test scores, one

that “extends the boundaries of validity beyond test score meaning to include relevance and utility, value implications, and social consequences” (Shepard, 1993: 424)? Herein lies a second concern. Despite Messick’s incorporation of the social dimension in his theory of construct validity, there is still a heavy reliance on the collection of empirical data and statistical measures. Messick stresses this in his work. He states that

> the essence of this unified view is that the appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that the integrating power derives from empirically grounded construct interpretation (Messick, 1989a: 8).

The “watchword for educational and psychological measurement is to maximise the empirically grounded interpretability of the scores and minimise construct irrelevancy” (Messick, 1989b: 89), he states.

A core question in this study is what social responsibility test developers have not just to test takers, but to everyone affected by the test – supervisors, parents, test administrators and society at large. It is not the intention of this study to ignore or discredit the contribution that empirical considerations or constitutive elements have made to the field of language testing, but instead to question whether these are sufficient to answer all questions about the process of measuring language ability. Despite the influence of Messick’s work, others in the field have raised questions and concerns.

McNamara and Roever explain Messick’s validity matrix as follows:

Table 3.4 Understanding Messick's validity matrix

	WHAT TEST SCORES ARE ASSUMED TO MEAN	**WHEN TESTS ARE ACTUALLY USED**
USING EVIDENCE IN SUPPORT OF CLAIMS: TEST FAIRNESS	What reasoning and empirical evidence support the claims we wish to make about candidates based on their test performance?	Are these interpretations meaningful, useful and fair in particular contexts?
THE OVERT SOCIAL CONTEXT OF TESTING	What social and cultural values and assumptions underlie test constructs and hence the sense we make of scores?	What happens in our education systems and the larger social context when we use tests?

(McNamara & Roever, 2006: 14)

McNamara and Roever translate the matrix to make clear "the way in which Messick's theory takes theoretical account of the aspects of the social dimension of assessment" (2006: 13). More importantly, they point out a glaring flaw in the matrix. In explaining Messick's validity matrix, McNamara and Roever state that "aspects of the social context of testing are more overtly present in the model, in the bottom two cells of the matrix" (2006: 13). This for them then raises the question of the

> relationship of the fairness oriented dimensions of the top line of the matrix to the more overtly social dimensions of the bottom line, a question it could be argued that Messick never resolved and remains a fundamental issue facing our field (McNamara & Roever, 2006: 13).

McNamara and Roever here raise an important point. Messick has stressed the unifying and integrating nature of his view of construct validity. On closer inspection of his matrix, one is forced to admit that there is no close integration or unifying of the different concepts. While Messick's matrix asks us to consider questions about the social dimension of language testing, these questions have been relegated to the bottom row of the matrix. The empirical and social still exist, but may in such a view continue to operate as separate entities in the field of testing. Kane, in discussing this states that the "model based on construct validity is elegant and conceptually rich and suggestive, but

it is not easy to implement effectively, because it does not provide a place to start, guidance on how to proceed, or criteria for gauging progress and deciding when to stop" (Kane, 2011: 8).

Shepard (1993) also raises a number of concerns with Messick's validity matrix. Unsurprisingly, a first concern has to do with the faceted nature of his matrix. Shepard states that the faceted presentation "allows the impression that values are distinct from a scientific evaluation of test score meaning" (1993: 427), an argument similar to the one raised by McNamara and Roever above. She states that the separate rows in Messick's table make it appear that one would first resolve "scientific questions of test score meaning and then proceed to consider value issues" (1993: 427). She also observes that the sequential segmentation of validity gives researchers tacit permission to leave out the very issues that Messick has highlighted, because the "categories of use and consequence appear to be tacked on to 'scientific validity', which remains sequestered in the first cell" (1993: 427). Shepard's final point on this is related to what she refers to as the "complexity of Messick's model" (Shepard 1993: 429), stating that both the model and the chapter on construct validity stress that construct validity is a "never-ending process" (429) and that while this may be true, the "sense that the task is insurmountable allows practitioners to think that a little bit of evidence of whatever type will suffice" (429).

This is undoubtedly ironic, as this was exactly the reason why Messick saw a need for a unified view of validity – so that test developers will not simply use one type of validity to validate a test. Messick argues against the compartmentalising of validity into different types, claiming that it "leads to confusion and, in the face of confusion, oversimplification" (Messick, 1980: 1014). He also goes on to claim that the distinctiveness between them is blurred by calling them all 'validity'. These are, then, his motivating reasons for unifying the concept of validity. This raises concerns. Messick's main aim was unifying the concept of validity. When we use the words 'unify' or 'unifying'

we refer to things that are the same as or uniform or not different or not varying. Is it possible that Messick's use of this term causes the very confusion he refers to above? If the concept of validity is a unified one, then it would make sense to see everything under that concept as being or meaning the same. If content, criterion and face validity are unified or the same as or not varying, it would potentially make sense to use any one type to validate the test – they are, after all, uniform or in Messick's words 'unified'. He states that "to speak of validity as a unified concept is not to imply that validity cannot be differentiated into facets" and that "the distinctions introduced may seem fuzzy because the facets of validity are not only intertwined but overlapping" (1989a: 9). Validity then is not just unified but both unified and faceted. In his attempts to unify the concept of validity, Messick has created a most complex network of arguments. It is hard to avoid the conclusion that Messick's attempts to view validity as a unified concept has to some extent created the very problems he wanted to avoid. This is compounded by the complexity of his writing which makes it inaccessible to many readers. There remains also a suggestion of too much of a reliance on the importance of empirical data. What the field requires is shared emphasis on empirical data as well as social effects.

Messick's claim that "validation is a continuing process" (1989b: 13) suggests that it is never-ending. Bachman and Palmer reiterate this view when they state that construct validation is an "on-going process" and that "we should not give the impression that a given interpretation is 'valid' or 'has been validated'" (1996: 23). Is it not possible that there could be an end to the process of validation? Should there not be a valid test at the end of the process? As Weideman asks, "Is it inconceivable that the process of producing evidence will confirm that, to the best of the test designer's knowledge, the test has the desired effect, i.e. it yields certain objective scores or measurements?" (Weideman, 2009a: 242). Is it not acceptable to ask whether an instrument that has undergone a process of validation may be shown to be a valid test? Does the validation not demonstrate that it does what it was designed to do? The

answer here should be an affirmative or a negative, however qualified the ‘yes’ or ‘no’ might be.

It is only when one unquestioningly accepts Messick’s definition of validity and construct validity that one is not allowed to ask such questions. Fortunately, experts in the field are asking such questions. McNamara and Roever make two thought provoking statements: that although “validity theory investigated the technical qualities of tests in the interests of fairness, it did not address the wider social function of tests” (2006: 248), and that “despite Messick’s efforts to build a unitary approach to validity that acknowledged the social meaning of tests, validity theory has remained an inadequate conceptual source for understanding the social function of tests” (2006: 249). Davies and Elder claim that “Messick’s conceptual clarity can be analysed less charitably” (2005: 799), stating that because validity can only be achieved through validation, what Messick does is offload all the problems of validity onto validation, “leaving validity as an abstract and essentially empty concept” (2005: 799). Shepard’s concluding concern with Messick’s matrix deals with the question of subsuming everything under the umbrella of construct validity. She states that “most theorists agree that validation includes the whole of Messick’s framework, not only the first box. But can all of the implied questions be subsumed under construct validation without degrading its scientific meaning?” (1993: 428).

It would be useful at this point to look at one further reinterpretation of Messick’s matrix. Weideman, by “turning some of the terms around so that we make a small adjustment to the matrix” (2009a: 239) suggests the following reading of Messick’s distinctions:

Table 3.5 The relationship of a selection of fundamental considerations in language testing

	***adequacy* of…**	***appropriateness* of…**
inferences made from test scores	depends on multiple sources of empirical evidence	relates to impact considerations / consequences of tests
the design decisions derived from the interpretation of empirical evidence	is reflected in the usefulness / utility or (domain) relevance of the test	will enhance and anticipate the social justification and political defensibility of using the test

(Weideman, 2009a: 239)

This matrix can be read as a number of claims or requirements for language testing, as follows (left to right, top to bottom):

(1) The technical adequacy of inferences made from test scores depends on multiple sources of empirical evidence.
(2) The appropriateness of inferences made from test scores relates to the detrimental or beneficial impact or consequences that the use of a test will have.
(3) The adequacy of the design decisions derived from the interpretation of empirical evidence about the test is reflected in the usefulness, utility, or relevance to actual language use in the domain being tested.
(4) The appropriateness of the design decisions derived from the interpretation of empirical evidence about the test will either undermine or enhance the social justification for using the test, and its public or political defensibility (Weideman, 2009a: 240).

The matrix above is thought provoking for a number of reasons. The most obvious of these is that the matrix is not a validity matrix, and construct validity does not appear in every cell. Instead the matrix is concerned with the "relationship of a selection of fundamental considerations in language testing" (Weideman, 2009a: 240).

The literature on language testing nonetheless is dominated by the concept of construct validity. Even Bachman and Palmer (1996) and Kunnan (2004),

though proposing concepts of test usefulness and test fairness, incorporate the concept of construct validity in their models. As indicated above, there are a number of problems with Messick's matrix and its heavy emphasis on construct validity. According to Weideman, the matrix above is clearly not "solely about validity" (2009a: 240). He states that concepts in the matrix, "while obliquely related to the technical power of a test … rather articulates the coherence or systematic fit of a number of concepts related to language testing" (2009a: 240). These concepts, as indicated in the matrix, would be the empirical evidence such as for reliability and validity, as well as ideas on the impact, usefulness/utility as well as the social justification and political defensibility of the test. The emphasis here is on a "select number of fundamental concepts in language testing" (2009a: 240) rather than subsuming all concepts under validity. The unitary concept of validity does "blur the distinctions" (to use Messick's words), but for reasons different to Messick's. Subsuming everything under construct validity undermines the importance of these other concepts, whereas seeing them as a number of fundamental concepts in language testing, we acknowledge their contribution to the responsible design of language tests. In arguing against the need for a unitary concept of validity, usefulness or fairness, Weideman states that concepts like "technical appropriateness, technical meaningfulness (interpretation) of measurements (test scores), utility, relevance, public defensibility and the like must be conceptually distinguishable to make sense" (2009a: 241). He observes that if one does not distinguish what is "conceptually distinct, the distinction so avoided subsequently obtrudes itself upon the conceptual analysis" (2009a: 241). Experts in the field, though, have stressed the need for a unified view of language testing, but have attempted to achieve this through the unified concepts of construct validity, usefulness or fairness.

Despite the questions, concerns and issues raised, Messick's work on validity and construct validity still dominates the literature on language testing. While

other experts in the field like Bachman and Palmer (1996) and Kunnan (2004) have differing views, these have been proposed very carefully without directly challenging the work of Messick. Are experts in the field too afraid to challenge Messick because of his influence in the field? While there are differing views, these are seen simply as alternatives to Messick, as Fulcher and Davidson observe about Bachman and Palmer's (1996) notion of test usefulness: "The notion of test 'usefulness' provides an alternative way of looking at validity, but has not been extensively used in the language testing literature" (2007: 15).

It is important to note, however, the significance of Messick's (1981; 1989a; 1989b) contribution to the field of language testing. While language testers now argue over the appropriateness of his validity theory, there is a lot about his work that cannot be refuted. One such example is his statement that

> what is to be validated is not the test or observation device as such but the inferences derived from test scores or other indicators – inferences about score meaning or interpretation and about the implications for action that the interpretation entails (Messick, 1989a: 5).

In fact Messick has stressed time and again the importance of validity as an "inductive summary of both the existing evidence for and the actual as well as potential consequences of score interpretation and use" (Messick, 1989a: 5). Borsboom *et al.* (2004: 1063), in setting forth their theory of validity and validation, argue the exact opposite of Messick's view. While Messick believes in the importance of the need to interpret scores, Borsboom *et al.* (2004: 1063) state that "no view could be farther apart from the one advanced here". They claim that "validity is not a judgement at all…it is the property being judged" (2004: 1063). Borsboom *et al.* were correct – no view could be further apart from this. As Weideman points out: "The scores of tests are indeed (technically qualified and theoretically grounded) objects. On their own…they are meaningless" (Weideman, 2009a: 242). So while one may agree with Borsboom *et al.* that validity is about the simple, factual question of whether a

test measures an attribute (2004: 1061), it is impossible to ignore the glaring flaw in their work – that numbers must be interpreted, they mean nothing on their own, they are not absolute.

3.6 Test usefulness

While Messick used a matrix to set out his theory on validity, Bachman and Palmer propose a model of test usefulness along with three principles that they believe provide a basis for answering the question "How useful is this particular test for its intended purpose(s)?" (Bachman & Palmer, 1996: 17). The model appears below:

> **Usefulness = Reliability + Construct validity + Authenticity + Interactiveness + Impact + Practicality**
>
> (Bachman & Palmer, 1996: 18)

The three principles they propose are:

1. It is the overall usefulness of the test that is to be maximised, rather than the individual qualities that affect usefulness.
2. The individual test qualities cannot be evaluated independently, but must be evaluated in terms of their combined effect on the overall usefulness of the test.
3. Test usefulness and the appropriate balance among the different qualities cannot be prescribed in general, but must be determined for each specific testing situation (Bachman & Palmer, 1996: 18).

Bachman and Palmer believe that the most important consideration in designing and developing a language test is the use for which it is intended. Therefore the most important quality of a test is its usefulness. They believe that "test usefulness provides a kind of metric by which we can evaluate not only the tests that we develop and use, but also all aspects of test development and use" (1996: 17). For Messick (1989a; 1989b) validity is the most important consideration, for Bachman and Palmer (1996) the most important quality is test usefulness, and for Kunnan (2004) the concept of test fairness is fundamental. These are clearly differing views.

Bachman and Palmer, like Messick, strive for ‘unity’. In the case of Messick the call was for the unifying of the concept of validity and in that of Bachman and Palmer (1996) the unifying of the six test qualities affecting test development and use that were mentioned above. The model above is a representation of their “view that test usefulness can be described as a function of several different qualities, all of which contribute in unique but interrelated ways to the overall usefulness of a given test” (Bachman & Palmer, 1996: 18). While for them there is tension among the different test qualities, this need not lead to the total abandonment of any. They argue that rather than emphasising the tension among the different qualities, test developers need to recognise their complementarity. According to Bachman and Palmer, the three principles outlined above

> reflect our belief that in order to be useful, any given language test must be developed with a specific purpose, a particular group of test takers and a specific language use domain (i.e. situation or context in which the test taker will be using the language outside of the test itself) in mind (Bachman & Palmer, 1996: 18).

Equally important to them is that “it is essential to take a systemic view, considering tests as part of a larger societal or educational context” (Bachman & Palmer, 1996: 19).

The move by Messick (1989a; 1989b) and Bachman and Palmer (1996) to include a concern for the social consequences in the field of language testing is indeed a positive move. It is their attempts to unify the field in terms of a single concept (validity/utility) that raises questions. Messick proposes a unified notion of validity while Bachman and Palmer propose a unified notion of test usefulness. What do these unified notions propose to do then? Do they unify the field of language testing? Is it simply a unified notion of validity? Or are they trying, through some ‘unified’ or ‘overarching’ concept such as usefulness or validity or fairness to unify the field itself? If the latter is the intention, then

one must question whether this ‘unification’ adequately addresses concerns about the incorporation of the social dimension in language testing in the design and administration of language tests.

Of the six qualities that contribute to Bachman and Palmer’s (1996) theory of test usefulness, reliability and validity are the essential measurement qualities. They provide the justification for using numbers and data as a basis for making judgements, inferences or decisions about the test taker. Reliability is defined as “consistency of measurement” and “an essential quality of test scores, for unless test scores are relatively consistent, they cannot provide us with any information at all about the ability we want to measure” (Bachman & Palmer, 1996: 20). What is important to note in their view is that the two measurement qualities, reliability and construct validity, are essential to the usefulness of any language test. However, they stress, as well, that while “reliability is a necessary condition for construct validity, and hence for usefulness, reliability is not a sufficient condition for either construct validity or usefulness” (Bachman & Palmer, 1996: 23). Stated simply, a test can be reliable but not valid or even useful. An important point raised at this juncture by Bachman and Palmer is that while language testers have been told that the qualities of reliability and validity may conflict, what is important is that test developers recognise their ‘complementarity’, that “test developers need to find an appropriate balance among these qualities, and that this will vary from one testing situation to another” (Bachman & Palmer, 1996: 18). What we can infer from this is that all six test qualities in their model are considered to be equally important, that the qualities will vary from one test situation to the next, but should not lead to abandoning any quality in favour of another.

The call here once again is for a unified, ‘systemic’ (Bachman & Palmer, 1996: 18) view of language testing. The call is also for a balanced, complementary perspective, or for a unified view to emerge from such a balanced perspective. Is this unified view only possible through the notion of usefulness or validity?

Is this unity possible only by subsuming everything under either one of these concepts? Weideman (2007a) sees reliability and validity as the two essential measurement qualities, referring to them as constitutive concepts (2007b; 2009a). Weideman specifies that reliability and validity are the founding concepts and are necessary conditions for test design. He refers to them as the "base or foundation" (2007a: 602) of the design.

Bachman and Palmer (1996), like Messick (1989a; 1989b), see construct validity as the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores. A crucial question for them is: "To what extent can we justify these interpretations?" (Bachman & Palmer, 1996: 21). In defining construct validity, Bachman and Palmer define a construct as the specific definition of an ability that provides the basis for a given test or test task and for interpreting scores derived from this task (1996: 21). Construct validity then, for them, is used to refer to 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 (1996: 21). Equally important, in their understanding of construct validity, is the fact that "construct validity is an on-going process of demonstrating that a particular interpretation of test scores is justified, and involves, essentially, building a logical case in support of a particular interpretation and providing evidence justifying that interpretation" (Bachman & Palmer, 1996: 22).

The third test quality that Bachman and Palmer highlight is authenticity. They define authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a TLU (target language use) task" (Bachman & Palmer, 1996: 23). Explained simply, what Bachman and Palmer call for is that there should be some alignment between the test tasks and the use of the language in real life situations. Another reason for them to consider authenticity as important is that it has a potential effect on test takers' perceptions of the test and, hence, on their performance:

> One way in which test takers and test users tend to react to a language test is in terms of the perceived relevance, to a TLU domain, of the test's topical content and the types of tasks required. It is this relevance, as perceived by the test taker, that we believe helps promote a positive affective response to the test task and can thus help test takers perform at their best (Bachman & Palmer, 1996: 24).

While this notion may look like the earlier concept of 'face' validity, i.e. the judgement of the appropriateness of the test by lay persons, the call for the designing of authentic test tasks is very clearly a move towards ensuring further accessibility and fairness in language testing. It is imperative that test takers see a link between the test tasks and the use of language in everyday situations, or in the particular context that their language ability is being tested for. The concept of authenticity as described by Bachman and Palmer can be closely related to the concept of (technical) alignment in the framework employed in this study. Such alignment, according to Weideman (2006a; 2007b; 2009a), is a regulative condition for test design and deals, among other things, with the link between the testing and the teaching that follows. Just as there should be authenticity or alignment between the test tasks and the use of the language in real life situations, there should also be alignment between what is tested and what is taught in an intervention programme that may follow the measurement. Students need to believe that the intervention or teaching that follows will help them achieve the desired results, in the same way that what gets tested is perceived to be in harmony with the target situation.

The teaching that follows a test will result in one of two things: positive washback or negative washback. With negative washback, the emphasis in the teaching is to get students to pass the test rather than to help them develop a particular skill or language ability. In the case of TALL and TALPS there is what we can call "positive alignment" between the test and the teaching. Students who do not achieve a particular score on the test must undertake the intervention programme which is based "on the same definition of academic

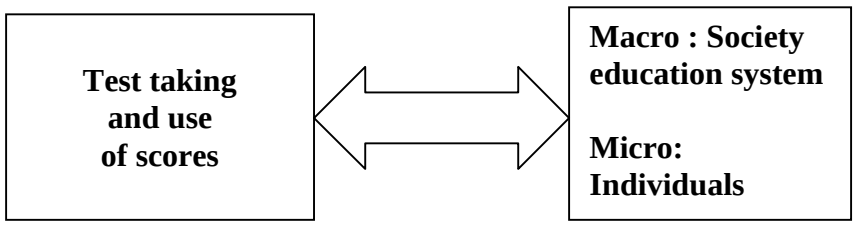
literacy as the test" (Weideman, 2006a: 82). The aim of the test is to test the academic literacy of the students, while the aim of the academic literacy course is to help develop the academic literacy of students.

Interactiveness, the fourth test quality Bachman and Palmer discuss, is defined by them as the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task (Bachman & Palmer, 1996: 24). They list the individual characteristics that are the most relevant for language testing as the test taker's language ability (language knowledge and strategic competence, or metacognitive strategies), topical knowledge, and affective schemata. They observe:

> The interactiveness of a given language test task can thus be characterised in terms of the ways in which the test taker's areas of language knowledge, metacognitive strategies, topical knowledge, and affective schemata are engaged by the test task (Bachman & Palmer, 1996: 25).

There is, of course, another level of 'interactivity' that is relevant. This is the interaction not only between test taker and test, but that between test designer and developer, test taker, test administrator and the users of test scores. All of these come into view when we look at the various social dimensions of the test that come together in what Weideman calls the implementation, or social use of such a technical instrument (2006a: 83). Bachman and Palmer call this "test taking and use of test scores" (1996: 30). They use the following diagram to explain the fifth test quality, impact:

Figure 3.5 Test impact



(Bachman & Palmer, 1996: 30)

The diagram indicates that the taking of a test and the use that is made of that test score has an impact on the individual test taker as well as on society, and on the education system as a whole. They ask: “What are the consequences for society, the educational system, and the individuals involved, of basing our decisions on test scores, rather than on some other criterion such as seniority or personal connections?” (Bachman & Palmer, 1996: 30). Test takers are affected by three aspects of the testing procedure:

1. the experience of taking and, in some cases, of preparing for the test,
2. the feedback they receive about their performance on the test, and
3. the decisions that may be made about them on the basis of their test scores (Bachman & Palmer, 1996: 31).

Concerns such as these indicate the move towards a more transparent process in language testing. It points to a concern with the social consequences of language testing – an area largely ignored in the past. Unlike in the past – where the concern was solely on the test developer and not on the test taker - specialists in the field, like Bachman and Palmer (1996), Shohamy (2001) and McNamara (2005) today ask questions about everyone affected by the test. This includes test takers, teachers, society and education systems.

In discussing the concern for the impact of the test on society and education systems, Bachman and Palmer ask relevant questions about the impact or consequences of the uses of test scores on individuals and society. They state that tests are not developed in a “value-free psychometric test-tube” (1996: 30). Two things are obvious from this: The first is that for Bachman and Palmer, tests cannot be developed and administered without a concern for values and consequences, and the second is that psychometric data are not enough to answer the questions related to values, impact and consequences. Bachman and Palmer ask that users of tests be aware of how tests impact on test takers and to ensure the “fairness of the decisions” (1996: 32) to be made. Fair decisions are the results of fair tests. Fair tests are only possible if test developers and users are responsible and accountable and if tests are not designed and administered

in a vacuum. For Weideman (2006a; 2007b), concerns about the impact or consequences of the uses of tests requires test developers and users to be responsible and accountable, and to design tests that are accessible and transparent.

While Bachman and Palmer (1996) have used the aspect of impact to encompass all issues relating to the social dimension of testing, the framework employed for this study sees the social dimension as being related primarily to the implementation/administration of the test. According to Weideman, test design leads to the production of a test (2009a: 247). The test is then implemented and administered to the (in this case) students. This administration "ties the technical instrument to its social context and use" (Weideman, 2009a: 247). In the framework used here, the concern for the way in which the test scores are used is considered under the juridical and ethical reflections within the leading technical aspect of a test, in the regulative ideas of transparency and accountability. In explaining the juridical analogy within the leading technical design function, Weideman states that

> the juridical analogies within the technical aspect of an applied linguistic artefact are evident, furthermore, in the theoretical and public justification for the intervention. The applied linguist needs to provide a defensible theoretical rationale for every design, which serves to enhance the legitimacy of the intervention. The more transparent the justification, the more accountable it should also be, to academics and non-academics alike (Weideman, 2007a: 601).

The ethical dimension for Weideman deals with concepts such as the love and care we have for our fellow human beings and the way that these find expression in our technical designs and artefacts (such as language tests). According to Weideman this should be evidenced in the technical artefacts we create. Of this he states that

> applied linguistic designs therefore find their meaning in the service (or disservice) that they will perform for other human beings. The care with which designs are made points to the love that we show for humanity.

> This love is evident in the technical artefacts that we create (Weideman, 2007a: 602).

The focus here is on ensuring that the applied linguistic design "promotes the interests of those who are affected by it" (Weideman, 2007a: 601).

The last test quality Bachman and Palmer (1996) make reference to in their work is practicality. They claim that practicality is different in nature from the other five qualities. While the other qualities pertain to the uses that are made of test scores, practicality pertains primarily to the ways in which the test will be implemented, and to a large degree, whether it will be developed and used at all. Practicality is defined "as the relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities" (Bachman & Palmer, 1996: 36). The writers present this in the form of the model below:

$$\textbf{Practicality} = \frac{\textbf{Available resources}}{\textbf{Required resources}}$$

(Bachman &Palmer, 1996: 36).

The resources referred to in the model above are classified into three general types. Bachman and Palmer list them as:

1. Human resources (e.g. test writers, scorers or raters, test administrators, and clerical support).
2. Material resources: *Space* (rooms for test development and test administration), *Equipment* (typewriters, word processes, tape and video recorders, computers), *Materials* (paper, pictures, library resources).
3. Time: *Development time* (time from the beginning of the test development process to the reporting of scores from the first operational administration), *Time for specific tasks* (designing, writing, administering, scoring, analysing) (Bachman & Palmer, 1996: 36).

What Weideman refers to as the economic dimension and Bachman and Palmer's sixth test quality of practicality cover similar aspects. Weideman is concerned with technical utility and frugality, while Bachman and Palmer

consider questions about the availability of resources as opposed to the required resources for the test.

Bachman and Palmer's notion of test usefulness is particularly important to this study, as one of the core questions is what social responsibility test developers have, not just to the test takers but to everyone affected by the test – supervisors, parents, test administrators and society at large. The model for test usefulness includes considerations of all these aspects. Like Messick (1989a; 1989b), Bachman and Palmer (1996) subsume all other considerations under one concept. In the case of Messick, as we have noted, it was construct validity. For Bachman and Palmer it is test usefulness. Much has been written about this in the literature on language testing, yet most of the research in the field has focused on asking and answering questions about statistical measures that are used to determine the validity and reliability of a test. This issue has been discussed in detail in chapter 1. What must be repeated is that all questions regarding test development and use cannot be answered using only statistical measurements and numbers. In addition, it is important to ask whether a consideration of the social dimension requires that everything be subsumed or lumped together under the umbrella of one concept.

3.7 Kunnan and the test fairness framework

While Messick (1989a; 1989b) has expounded the concept of construct validity and Bachman and Palmer (1996) that of usefulness, Kunnan argues that "the concept of test fairness is arguably the most critical in test evaluation" (2004: 27). He states that very often test developers and researchers have valued only statistical evidence and ignored or discounted other types of evidence. Test evaluation is conducted "narrowly and focuses mainly on validity and reliability" (2004: 31). According to Kunnan (2004: 28), many testing professionals hold the view that testing research has always focused on issues of fairness through the concepts of validity and reliability, indicating once again a concern only with empirical evidence and, preferably, that kind of

empirical evidence that can be stated in numbers. What Kunnan (2004) proposes is a test fairness framework that can be used for evaluating tests and testing practice. Kunnan sees the concept of fairness as critical in testing practice, influenced by his belief that "language-testing professionals need an ethic to support a framework of applied ethical principles that could guide professional practice" (2004: 33). He presents testing professionals with an "ethics-inspired rationale for the Test Fairness framework" (2004: 33). Kunnan presents a brief explanation of the three predominant ethical approaches: utilitarianism, Kantian and deontological systems, and virtue-based ethics. He explains that

> …utilitarianism, which emphasises good results as the basis for evaluating human action, has two main features: its teleological aspect or consequentialist principle and the hedonic aspect or utility principle. In contrast, Kantian or deontological ethics focuses on ideals of universal law and respect for others as a basis of morality and sees the rightness or wrongness of an act in itself, not merely in the consequences of the act. Virtue-based ethics calls persons to be virtuous by possessing both moral and non-moral virtues by imitation, even though there are no principles to serve as criteria for the virtues (Kunnan, 2004: 32).

According to Kunnan his framework is based on "Frankena's mixed deontological system, which combines both the utilitarian and the deontological systems" (Kunnan, 2004: 33). This system has two general principles:

Principle 1:	*The principle of justice*: A test ought to be fair to all test takers; that is, there is a presumption of treating every person with equal respect.
Sub-principle 1:	A test ought to have comparable construct validity in terms of its test-score interpretation for all test takers.
Sub-principle 2:	A test ought not to be biased against any test-taker groups, in particular by assessing construct-irrelevant matters.
Principle 2:	*The principle of beneficence*: A test ought to bring about good in society; that is, it should not be harmful or detrimental to society.

> Sub-principle 1: A test ought to promote good in society by providing test-score information and social impacts that are beneficial to society.
> Sub-principle 2: A test ought not to inflict harm by providing test-score information or social impacts that are inaccurate or misleading. (Kunnan, 2004: 34).

In attempting to understand and explain the notion of fairness Kunnan turns to explanations provided in the *Standards for Educational and Psychological Testing* (*AERA*, 1999):

> The first two characterisations relate fairness to *absence of bias* and to *equitable treatment of all examinees* in the testing process. The third characterisation of test fairness addresses the *equality of testing outcomes* for examinee subgroups defined by race, ethnicity, gender, disability, or other characteristics. The fourth definition of fairness relates to *equity in opportunity to learn* the material covered in an achievement test (*AERA*, 1999: 74).

Kunnan's Test Fairness framework views fairness in terms "of the whole system of a testing practice, not just the test itself" (Kunnan, 2004: 37). Kunnan's model has five main qualities: validity, absence of bias, access, administration and social consequences, as indicated in the table below:

Table 3.6 Test fairness framework

Main quality	Main focus
1. Validity	
Content representativeness/coverage	Representativeness of items, tasks, topics Representation of construct/underlying trait
Construct or theory based validity	Test score comparison with external criteria Stability, alternate form, inter-rater and internal consistency
Criterion-related validity	
Reliability	
2. Absence of bias	
Offensive content or language	Stereotypes of population groups
Unfair penalisation	Content bias based on test takers background
Disparate impact and standard setting	DIF in terms of test performance; criterion setting and selected decisions
3. Access	
Educational	Opportunity to learn
Financial	Comparable affordability
Geographical	Optimum location and distance
Personal	Accommodations for test takers with disabilities
Equipment and conditions	Appropriate familiarity
4. Administration	
Physical setting	Optimum physical settings
Uniformity and security	Uniformity and security
5. Social consequences	
Washback	Desirable effects on instruction
Remedies	Re-scoring, re-evaluation; legal remedies

(Kunnan, 2004: 46)

According to Kunnan (2004: 39), these five test fairness qualities (validity, absence of bias, access, administration and social consequences) when working together could contribute to fair tests and testing practices. An important point that he raises is that the concern for fairness in language testing cannot be raised only after the test is developed and the test administered. He states that the "concern has to be present at all stages of test development: design, development, piloting and administration and use (which includes analysis and research), although different fairness qualities may be in focus at different stages" (Kunnan, 2004: 39). This is a far cry from earlier practices that saw test validation as something conventionally carried out only once the test was developed and administered. In answering the question of who is responsible for fairness-testing practices, Kunnan's answer is that in his view all primary stakeholders should be held responsible for promoting fairness (2004: 40). Like Bachman and Palmer (1996), Kunnan sees the field of language testing as one in which both test developer and test taker have rights, bringing to the fore issues of accessibility, transparency and accountability. One is forced to agree with Kunnan's contention that if a test is not fair there is little value in a test having qualities such as validity and reliability of test scores (2004: 40).

While Kunnan's (2004) framework does not adequately answer the questions posed by this study, it does indicate, like the Bachman and Palmer model, a definite move away from more traditional methods of validating language tests. The models proposed by Kunnan (2004) and Bachman and Palmer (1996) build onto the framework proposed by Messick (1989a; 1989b), the first framework/model to consider questions related to test use, social consequences and test fairness, while not disregarding the importance of empirical or statistical evidence. What these researchers present us with is the possibility of and a discussion of the principles for designing, developing and administering fair and accessible language tests, thus ensuring that test developers are answerable for their designs, that test takers can empower themselves by asking questions about the test, and that the consequences of the uses of the test results

are considered even before the design and development of the test. These experts foreground these issues by proposing an 'overarching' concept such as usefulness or validity or fairness to unify the field of language testing. As pointed out by Weideman, there is "divergence and congruence" (2009a: 239) between these concepts. The divergence between them is quite clear – either usefulness or fairness or validity is seen as the key concept in language testing. The congruence between them, according to Weideman, has to do with the fact that all these concepts stress the importance of construct validity. So while they may differ, the shared emphasis on construct validity draws attention away from this difference. Equally important is Weideman's observation that, while the concepts may differ, for Messick (1989a; 1989b) and Bachman and Palmer (1996) "an overarching or unified view of language testing is both required and desirable" (Weideman, 2009a: 239). Rather than these overarching concepts, what the field requires is that test developers see the relationship between fundamental concepts in language testing. It is the argument of this thesis that these fundamental concepts are related to the technical dimension of our experience, and the way that it relates to other dimensions that are reflected in it.

3.8 Conclusion

An important part of this study is in telling the story of TALPS. The rest of this study is, therefore, focused on a discussion of these fundamental concepts as they relate to TALPS. As explained earlier in this chapter, these fundamental concepts can be separated into constitutive and regulative conditions. While this study is primarily concerned with the regulative conditions of transparency, accessibility and accountability, it is necessary to consider the constitutive concepts of validity and reliability. Chapter 4 therefore provides the reader with a detailed discussion of the validity and reliability measures of TALPS. Chapters 5, 6 and 7 will focus respectively on the regulative concepts of transparency, accessibility and accountability.

Chapter 4

The constitutive concepts underlying the design of TALPS

4.1 Introduction

The concepts of reliability and validity help to determine whether the test is a strong one or not, does what it is designed to do, tests what it is designed to test, whether it is a consistent test, and whether we can make inferences that are justified about the test takers, based on their scores (cf. Van Dyk, 2010). Clearly, a lot therefore depends on the reliability and validity of a test. Despite being an important and essential part of the narrative, reliability and validity, however, do not tell the entire story of the test, and while the focus of this study is to identify the other equally important parts of the tale, issues of reliability and validity is where the story begins. The focus in this chapter is on the validity and reliability of TALPS. This discussion will take the form of a validation argument, similar in approach to the one taken by Van der Walt and Steyn (2007).

4.2 Validity and the validation argument

The concept of validity is indeed a complex, multifaceted concept that has undergone different interpretations (Van der Walt & Steyn, 2007: 138), as can be seen from the discussion in chapter three. It goes without saying, though, that whichever way one interprets the concept, there is agreement about the validity question, which asks: Does the test measure what it is designed to assess? Providing an answer to this question requires one to engage in a process of validation, i.e. provide a multiplicity of sets of evidence to support the claims made about the test.

There is evidently a distinction between the concept of validity and the act of validation. Davies and Elder (2005: 799) make reference to this distinction when they speak of Messick's concern with "validity as a theoretical concept rather than with validation in its practical operation". In discussing this distinction

further, their claim is that validity is an abstract and essentially empty concept, that it is through "validation that validity is established, which means that validity is only as good as its validation procedures" (Davies & Elder, 2005: 795). Van der Walt and Steyn label validation "an activity: the collection of all possible test-related activities from multiple sources" (2007: 141). These "multiple sources" of evidence may include what are traditionally conceived as content and construct validity, concurrent and predictive validity, face validity, reliability, as well as consequential validity (Davies & Elder, 2005: 798). In a nutshell:

> The validation process involves the development of a coherent validity argument for and against proposed test score interpretation and uses. It takes the form of claims or hypotheses (with implied counter claims) plus relevant evidence (Van der Walt & Steyn, 2007: 142).

Valuable advice pertaining to the construction of a validation argument is provided by Fulcher and Davidson (2007: 20) in their articulation of the following principles that condition the process of validation:

- Simplicity: explain the facts in as simple a manner as possible.
- Coherence: an argument must be in keeping with what we already know.
- Testability: the argument must allow us to make predictions about future actions or relationships between variables that we could test.
- Comprehensiveness: as little as possible must be left unexplained.
(Fulcher & Davidson, 2007: 20).

What follows below is a brief discussion of the validation of the TALPS test. While the focus of this study is not on the validation of TALPS, as indicated earlier, the validation of the test nonetheless remains an important part of telling the story of the test.

4.3 A validation of TALPS

Claim 1: The test is reliable and has a low standard error of measurement score.

According to Kurpius and Stafford, reliability can be defined as the "trustworthiness or the accuracy of a measurement" (2006: 121). Bachman and

Palmer state that reliability can be considered to be a function of the consistency of scores from one test, and test tasks, to another (1996: 19). Another important point that Bachman and Palmer make is that reliability is an essential quality of test scores and that unless test scores are relatively consistent, they cannot provide us with any information at all about the ability we want to measure (1996: 20). Ensuring that the test designed is a reliable test and that information about the test is available to everyone affected by the test points to the responsibility that the test developers are willing to take for their designs, and demonstrates how the constitutive concept of technical consistency relates to the regulative idea of accountability, an idea returned to in a subsequent chapter.

Kurpius and Stafford (2006) identify four types of reliability: test-retest reliability, alternate forms reliability, internal consistency reliability and interrater reliability. The main reliability measure used for both TALPS and TALL is internal consistency reliability. Test-retest reliability requires the test to be taken twice by the same group of students, alternate forms reliability is used when "you want to determine whether two equivalent forms of the same test are really equivalent" (2006: 126), and interrater reliability is used when "two or more raters are making judgements about something" (2006: 129). The internal consistency reliability is obviously the most practical to use – it requires that students take the test once and reliability measures are calculated using statistical packages like TiaPlus, SPSS or Iteman.

The instrument/package that was used (TiaPlus: cf. CITO, 2006) provides two measures: Cronbach's alpha and Greatest Lower Bound (GLB). All three pilots of the test have rendered very impressive reliability measures as indicated in Table 4.1. The first pilot had a reliability of 0.85 (Cronbach's alpha) and 0.92 (GLB). The pre-final draft had measures of 0.93 (Cronbach's alpha) and 1.00 (GLB). The final version of the test had measures of 0.92 (Cronbach's alpha) and 0.99 (GLB). The measures for the final draft were based on the analysis

done on the combined group (North-West University and University of Pretoria).

Table 4.1 Reliability measures for the TALPS pilots

TALPS pilot	**1st pilot**	**2nd pilot**	**2nd pilot**	**3rd pilot**
Cronbach's alpha (reliability)	**0.85**	**0.93**	**0.93**	**0.92**

The low standard error of measurement score is another indication of the reliability of the test. The standard error of measurement is a "deviation score and reflects the area around an obtained score where you would expect to find the true score" (Kurpius & Stafford, 2006: 132). A test can never be one hundred percent reliable. At the same time a test score is not always an accurate indication of your abilities. It has been accepted and is expected that "no person's obtained score (X_o) is a perfect reflection of his or her abilities, or behaviours, or characteristics, or whatever it is that is being measured" (2006: 101). Therefore, the score a person has *obtained* is not looked at in isolation but in combination with a *true* score and an *error* score. The basic equation for this is: $X_o = X_t + X_e$, where

- X_o = the score obtained by a person taking the exam (referred to as an obtained score or observed score)
- X_t = a person's true score
- X_e = the error score associated with the obtained score

(Kurpius & Stafford, 2006: 103).

With reference to Table 4.2 below, the individual student's true score then would be the obtained score - (minus) 3.87 (2nd pilot done on University of the Free State students) or 3.82 (3rd pilot done on University of Pretoria and University of the Free State students), which in each case is the standard error of measurement. Kurpius and Stafford explain that a smaller standard error of

measurement reflects a smaller error score and that the goal in reliability is to control error (2006: 133). A higher reliability is therefore an indication of a small error of measurement. The 1st pilot of TALPS that has a reliability of 0.85, had a standard error of 4.30. When the reliability measures in subsequent pilots improved, the standard error of measurement dropped to 3.82 and 3.87, respectively, as can be seen in Table 4.2 below.

The mean or average for all the pilots totals 57.22. The variance around the mean is highest for the students at University of the Free State with a 15.13 standard deviation. Overall, the variance around the mean for all the pilots seems to be quite stable, suggesting a normal or even distribution of scores around the mean. In the TALPS final version the standard error of measurement for the combined group of North-West University and University of Pretoria students is 3.84, for North-West University students 3.83 and for University of Pretoria students 3.80.

Table 4.2 Descriptive statistics of the TALPS pilots

Pilot	Mean	St.Deviation	SEM.
1st pilot (UP)	58.28	11.10	4.30
2nd pilot (UP & UFS)	61.33	14.19	3.82
2nd pilot (UFS)	57.38	15.13	3.87
3rd pilot (UP & NWU)	51.88	13.32	3.84

One other set of empirical information about the reliability of the test yielded by the TiaPlus Test and Item Analysis is the Coefficient Alpha of the test if it had a norm length of forty items. TALL and TALPS are made up of a number of short subtests. The reliability of a test or in this case a subtest will be compromised by its length – the longer a test is, the more reliable it usually is. Kurpius and Stafford explain that when a test is too short, the reliability coefficient is suppressed due to the statistics that are employed. The *Spearman-*

Brown correction procedure can be used to make up for this (Kurpius & Stafford, 2006: 129). One example of this is the **Dictionary definitions** subtest in the TALPS first pilot. This section has five items and is one of the shortest sections in the test. It has a Coefficient alpha of 0.36 and GLB measure of 0.42. The *Spearman Brown correction* procedure indicates a Coefficient Alpha of 0.82 if it had a standard norm length of 40 items. The ideal then would be to design longer tests, thus ensuring higher reliability measures. But – and this is always the technical trade-off – such effectiveness may conflict with technical implementation constraints: there may not be enough time available to administer the test. The test developer has to weigh up the advantages and disadvantages of lengthening a test, and take a responsible design decision (see below, Section 4.4 A longer and more reliable test).

Claim 2: ***The inter-rater reliability measure of the writing section is of an acceptable level.***

The need to develop a reliable testing instrument dictates that the inter-rater reliability measure be considered as well. According to Huot (1990: 202), an inter-rater reliability measure of "at least .7" is an "acceptable standard" (1990: 202). Inconsistencies between markers will obviously affect the reliability of the results of the test.

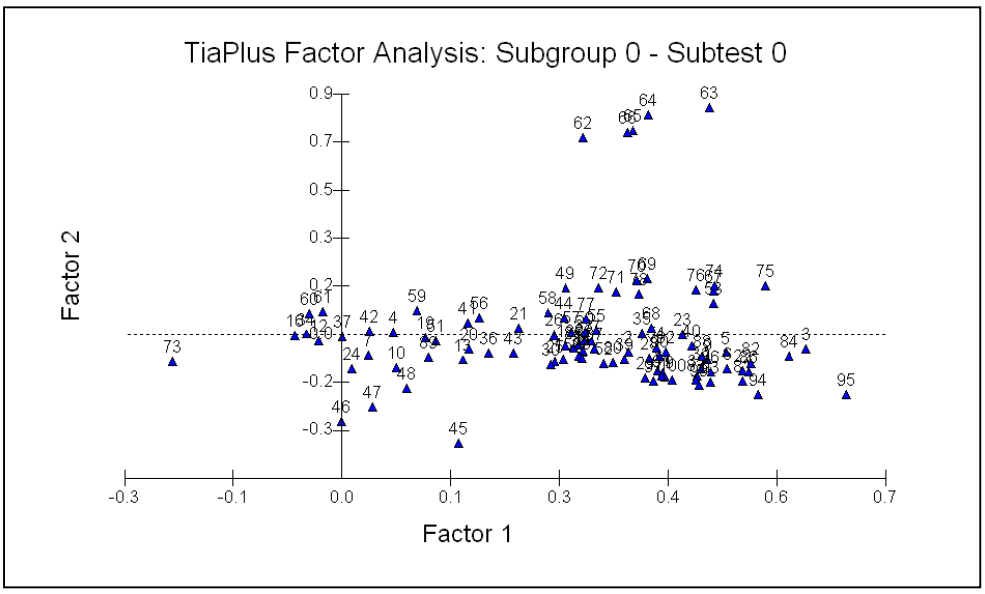
In determining the inter-rater reliability of the markers of the writing section of TALPS, the test designers followed a number of steps. In the first step of this process, two of the test designers, both of whom have extensive experience in the marking and assessment of student writing, took a number of writing sections that had been completed by the initial testees. These were marked using a rubric (**Appendix B**). After careful study of their marked papers, the markers found that there was no substantial discrepancy between their marking. The correlation between them was 0.8 – definitely a more than acceptable measure. When they then averaged their marks, they got an ideal mark for each student.

In the next step of this process a work session was organised, in which the initial markers acted as moderators. The main purpose of the session was to have a number of markers assess the same number of completed written sections of the test that had been pre-marked by the moderators, using the same rubric they had used before. The markers involved in this exercise are lecturers who teach in the Unit for Academic Literacy and are familiar with the marking of student assignments. The markers provided feedback on why they gave a specific score to a specific section using the rubric. The procedure was repeated a number of times with different students' written sections of the test. The data was then analysed with regard to the scores awarded to the same written texts by different markers, and again the inter-rater reliability was on an acceptable level i.e. well above 0.7. The conclusion of this session was that raters could successfully be trained to mark the writing section of TALPS, but needed to be monitored and moderated frequently.

Claim 3: The reliability measures of the test have not been compromised by the heterogeneous items in the test.

A factor analysis is used to determine whether the items in the test actually do measure just one construct or ability, in this case academic literacy. According to Ho (2005: 203) "the main aim of factor analysis is the orderly simplification of a large number of intercorrelated measures to a few representative constructs or factors." The workings of a measure of this nature and an example have been provided in chapter 3. The factor analysis for the TALPS first pilot appears below:

Figure 4.1 Measures of homogeneity/heterogeneity of TALPS first pilot



(Geldenhuys, 2007: 73)

According to Geldenhuys (2007: 73) the more heterogeneous items are, the less reliable the test can become. The factor analysis above indicates that in the case of TALPS there is a measure of heterogeneity: items 73 and items 62–66 are furthest away from the zero line. The test, however, still had a reliability measure of 0.85. The test designers chose to leave in these items, for reasons as explained in chapter 3.

Claim 4: The items on the test discriminate well between test takers.

One other statistical measure rendered by the package used is the average *Rit*-values (CITO, 2006) or the discriminative ability of the test items. One of the main purposes of a test is to be able to discriminate between the test takers. According to Kurpius and Stafford (2006: 115) a test cannot discriminate unless the items themselves discriminate between those who correctly answer the questions and those who do not. One of the main reasons to pilot a test is to determine which items discriminate well and which do not. Find below the average *Rit*-values for the TALPS pilots:

Table 4.3 Average Rit-values of the TALPS pilots

Pilot	Average Rit-values
1st pilot	0.25
2nd pilot	0.37
2nd pilot	0.40
3rd pilot	0.40

The average *Rit*-values for the first pilot are low, though this could be justified by the fact that the first pilot had one hundred items, some of which were shown to be weak items. Once these items had been excluded from the test, the measures rendered more acceptable *Rit*-values. The *Rit*-values for the 3rd and 4th pilots are relatively stable at 0.40, which is well above the 0.30 benchmark.

Claim 5: The test is based on a theoretically sound construct.

In terms of Kunnan's Test Fairness framework, construct validity is concerned with the representation of the construct/underlying trait (2004: 37) that is being measured. Bachman and Palmer (1996: 21) define a construct as the "specific definition of an ability that provides the basis for a given test or test task and for interpreting scores derived from this task." They explain further that construct validity is used to refer to the extent to which we can interpret a given test score as an indicator of the abilities or constructs we want to measure (1996: 21).

The construct for TALPS is based on the same construct as TALL – a detailed discussion of this can be found in chapter 1. TALL has in many ways been the sounding board for TALPS. Moreover, the success of TALL has in part been the justification for TALPS. TALL and TALPS are designed to test the same ability – the academic literacy levels of students: undergraduate in the case of TALL, and postgraduate in the case of TALPS. The construct on which TALPS is based should therefore be a valid one and has been discussed briefly in chapter 2.

Mention has also been made in chapter 1 and chapter 2 of the reasons for the switch from the ELSA PLUS to the present construct.

The discussion here of the construct validity of TALPS demands further discussion of the construct on which the test is based. Van Dyk and Weideman (2004a: 7) set out to answer the all important question of "what would a construct based on a theory of academic literacy look like?" In doing so they considered the work of Blanton (1994), Bachman and Palmer (1996) and Yeld (2000). Blanton's (1994) definition was important to Van Dyk and Weideman (2004a) because it "described what proficient academic readers and writers should do" (2004a: 7). Importantly, it was a move away from an emphasis on vocabulary and grammar towards what Weideman has referred to as an "open view of language" (2003a: 58). When turning to Bachman and Palmer's (1996) definition of language ability, Van Dyk and Weideman (2004a: 8) found that while it provided more detail, the "apparent seepage between categories" in the construct could be confusing. In addition, Bachman and Palmer (1996: 66) point out that the construct would have to be reinterpreted for each testing situation. Yeld (2000) has done exactly this in the design of the academic literacy test developed at the Alternative Admissions Research Project (AARP), and it is this construct that Van Dyk and Weideman (2004a) find most useful. Van Dyk and Weideman (2004a: 10) point out, however, that while the construct was useful, the AARP test was an admissions test, was part of a larger battery of tests, was a two and a half hour test and took more than three hours to administer. This would not be practical for the academic literacy test planned for the students at the University of Pretoria. What was needed was a "reconceptualisation" (Van Dyk & Weideman, 2004a: 10) of how the test was designed. After much rationalising, re-ordering and reformulating (2004a: 10), the result was a "streamlined version" (2004a: 10) that made possible the testing of academic literacy within a much shorter time frame.

The proposed blueprint for the test of academic literacy for the University of Pretoria requires that students should be able to:

- understand a range of academic vocabulary in context;
- interpret and use metaphor and idiom, and perceive connotation, word play and ambiguity;
- 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;
- interpret different kinds of text type (genre), and show sensitivity for the meaning that they convey, and the audience that they are aimed at;
- interpret, use and produce information presented in graphic or visual format;
- 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;
- 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;
- 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;
- 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 (Weideman, 2003a: 61).

Van Dyk and Weideman (2004a: 11) point out that the abilities in the blueprint echo strongly what it is that students are required to do at tertiary level. The construct has been discussed at seminars, at presentations and with other experts in the field (2004a: 11). There has been consensus about the fact that the elements identified in the blueprint constitute a number of essential components of what academic literacy entails (Van Dyk & Weideman, 2004a: 11).

Claim 6: The internal correlations of the different test sections satisfy specific criteria.

In addition to the discriminative power of items, test developers are also concerned with the internal correlations in a test i.e. determining how well subtests in a test correlate/depend or work with each other as well as the whole test. The table below is an indication of the internal correlation of the 2nd pilot of TALPS (University of Pretoria and University of the Free State students):

Table 4.4 Table of subtest intercorrelations (TALPS 2nd pilot)

	Subtest	Total test	Subtest(s) 1	2	3	4	5	6	7
Scrambled text	1	0.43							
Interpreting graphs	2	0.73	0.22						
Dictionary definitions	3	0.37	0.24	0.12					
Academic vocabulary	4	0.59	0.24	0.53	0.31				
Understanding texts	5	0.83	0.25	0.51	0.32	0.35			
Grammar & text relations	6	0.82	0.24	0.51	0.15	0.36	0.57		
Text editing	7	0.72	0.18	0.40	0.24	0.34	0.50	0.59	
Number of testees	:	117	117	117	117	117	117	117	117
Number of items	:	88	5	10	5	10	33	15	10
Average test score	:	61.33	2.66	6.56	4.19	7.60	23.75	8.96	7.62
Standard deviation	:	14.19	1.92	2.89	0.93	1.83	4.89	4.52	2.85
SEM	:	3.82	0.75	1.22	0.75	1.24	2.30	1.51	1.00
Average P-value	:	69.7	53.16	65.64	83.76	75.98	71.98	59.72	76.15
Coefficient Alpha	:	0.93	0.85	0.82	0.35	0.54	0.78	0.89	0.88
GLB	:	1.00	0.94	0.89	0.41	0.74	0.94	0.96	0.92
Asymptotic GLB	:	Na	Na	Na	Na	Na	Na	Na	Na

Davies *et al*. (1999) explain that a correlation coefficient is a value showing the degree to which two variables are related, that a coefficient of zero indicates that there is no relationship between the two variables, a coefficient of -1 indicates a perfect negative correlation, and a coefficient of +1 indicates a perfect positive correlation (1999: 36). In terms of the correlation between each pair of subtests, these should fall between 0,3–0,5 (Alderson, Clapham & Wall, 1995: 184).

Alderson *et al.* explain that the reason for having different test components is that they all measure something different and therefore contribute to the overall picture of language ability attempted by the test. These correlations should be fairly low, in the "order of +.3 - +.5" (Alderson *et al*, 1995: 184). If, however, these components correlate very highly (around +.9) we may wonder whether the two subtests are testing different traits or skills, or whether they are testing the same thing (1995: 184). Of the 21 correlations in the table above, 9 fall below 0.3. Should we adjust, in line with our experience with TALL, these levels to 0.2 and 0.5, then 15 of the 21 are between the acceptable parameters. With regards the correlation between each subtest and the whole test, this should be "around +.7 or more since the overall score is taken to be a more general measure of language ability than each individual component score" (Alderson *et al*, 1995: 184). Overall the average of the correlation between each subtest and the whole test, while not ideal, is an acceptable 0.64. The subtests that correlate best with the whole test are the **Interpreting Graphs**, **Understanding Texts** and the **Grammar and Text Relations** subtest. In the case of the TALPS final draft version (University of Pretoria and North-West University combined), the average of the correlations between each subtest and the whole test is 0.66, indicating that the subtests correlate well, and more acceptably, with the test.

Table 4.5 Table of subtest intercorrelations (TALPS final draft version) UP & NWU combined

	Subtest	Total test	Subtest(s) 1	2	3	4	5	6	7
Scrambled text	1	0.47							
Graphic & visual literacy	2	0.78	0.30						
Academic vocabulary	3	0.61	0.35	0.39					
Text types	4	0.37	0.26	0.22	0.19				
Understanding texts	5	0.81	0.17	0.64	0.39	0.20			
Grammar & text relations	6	0.82	0.32	0.54	0.41	0.23	0.54		
Text editing	7	0.77	0.32	0.49	0.42	0.20	0.54	0.60	
Number of testees	:	272	272	272	272	272	272	272	272
Number of items	:	76	5	10	10	5	21	15	10
Average test score	:	51.88	1.94	6.81	7.44	2.17	17.30	8.74	7.48
Standard deviation	:	13.32	1.57	2.77	1.89	1.24	4.19	3.96	2.75
SEM	:	3.84	0.78	1.21	1.26	0.87	2.31	1.62	1.05
Average P-value	:	64.84	38.75	68.13	74.38	43.38	69.19	58.28	74.78
Coefficient Alpha	:	0.92	0.76	0.81	0.56	0.51	0.70	0.83	0.85
GLB	:	0.99	0.88	0.89	0.69	0.72	0.85	0.92	0.89
Asymptotic GLB	:	Na	Na	Na	Na	Na	Na	Na	Na

Claim 7: The test displays content validity.

A factor analysis is also useful in determining the content validity of the test. According to the factor analysis above, not all items are related to a single construct underlying the test. As indicated earlier, it was the decision of the test developers not to exclude these outlying items, the reasoning being that "for an ability as richly varied and potentially complex as academic language ability, one would expect, and therefore have to tolerate, a more heterogeneous construct" (Weideman, 2009a: 237). Leaving out these outlying items would have increased the reliability of the test. The test, however, already has an excellent reliability measure of 0.85. The high reliability of the test allows the test developer the freedom to include these items without compromising the reliability of the test or the construct.

The one other method of determining the content validity of the test is to get "expert ratings of the relationship between the test items and the content domain" (Kurpius and Stafford, 2006: 147). These experts judge each item to determine "how well it assesses the desired content" (2006: 147). In the case of TALPS, members of the design team were already familiar with the design of a test of this nature, having been involved in the design of the TALL. In addition to this, drafts of TALPS were evaluated by other specialists within the academic institutions involved who were either interested in being involved in the process of design and development or were interested in using it on their students.

Claim 8: The face validity of the test meets the expectations of potential users.

The concept of face validity can be considered a problematic one in the field of language testing. In most of the literature in the field it is not included as one of the types of validity, experts believing that it is not really validity because it does not deal specifically with the test but with the appearance of the test. Bachman (1990: 287) points out that the term has been buried, that the "final internment of the term is marked by its total absence from the most recent (1985) edition of the 'Standards'". Despite this, the concept of face validity has made its mark in the field, as is evident from Bachman's observation that "even those who have argued against 'test appeal' as an aspect of validity have at the same time recognised that test appearance has a considerable effect on the acceptability of tests to both test takers and test users" (1990: 289).

If one were to accept unquestioningly Messick's concept of construct validity as the "be all and end all" of validity, then it would be easy to discredit or ignore the value of face validity. The move to redirect the field of language testing to one that is more 'inclusive' (Shohamy, 2004: 79) of all those involved in the testing process dictates that any and all information pertaining to the test be available to all those involved. A consideration for the social dimension of

testing means that the test takers are no longer external to the test but are an important consideration from as early as the design stage. One of the concerns with the concept of face validity, as pointed out by Bachman, is that "we may become complacent, accept the appearance of "real life", or 'face', for validity, and fail to discharge our duties and responsibilities, as test developers and users, for demonstrating the evidential bases for test interpretation and use" (1990: 288). Bachman makes a valid point. As already pointed out, the literature is filled with examples of test developers and users who have used tests in unethical ways. Responsible test developers, in their quest to ensure transparency of the test development process, and in their desire to be accountable for their designs, should know that face validity is not sufficient to validate a test, just as a high reliability does not necessarily indicate that the test is suitable for the purpose for which it was intended.

The point of this study is to show that no one aspect defines a good test, that a valid and reliable test is not necessarily a fair and socially acceptable test. Test development and implementation is a long process and every step of this process is an important one. All evidence collected about the test, during its design, development and implementation, results from its pilot/s. This means that the opinions of those involved and affected by the use of the test/test scores can be used to enhance the value of the test. It would be unethical to use only one type of evidence to validate a test. The framework employed in this study outlines the conditions that must be considered in the design and development of a test. These are a combination of the psychometric qualities of the test (validity and reliability) and the social and other considerations (articulation, implementation, utility, alignment, transparency, accountability, fairness and care), creating a more comprehensive set of test design principles.

In this view then, the concept of face validity does contribute to creating a whole and more complete picture of the test. Face validity can also be related to Bachman and Palmer's concept of authenticity, which calls for the alignment of

tasks and the use of language in real life situations. Test takers need to believe that the questions they are being asked are relevant to what they want to learn or are relevant to what they are being tested on. This can also be related to the concept of ‘alignment’ in the framework used in this study.

McNamara (2000: 133) defines face validity as the extent to which a test meets the expectations of those involved in its use; the acceptability of a test to its stakeholders. Davies *et al*. (1999: 59) explain that face validity is the degree to which a test appears to measure the knowledge or abilities it claims to measure, as judged by an untrained observer. They explain that while face validity is often dismissed as ‘trivial’ (1999: 59), failure to consider the face validity of a test may “jeopardise the public credibility of a test” (1999: 59).

Face validity does not stand alone and apart from other types of validity evidence. According to Butler (2009: 293), face validity can be related to content validity. He points out that for a test to have content validity, the items in the test should reflect the domain being tested. We can relate this content validity to face validity by determining whether the items in the test are “transparent to such an extent that, when evaluating its potential usefulness postgraduate supervisors will be able to recognise the relevance of what is being tested” (2009: 293).

The face validity of TALPS is therefore an important consideration. Students who will be writing the test are from different faculties and disciplines. Their supervisors are not experts in the field of language testing and academic literacy. In having their students write the test they will have to believe that the test looks right, that it looks like a test that is testing the academic literacy of their students. In attempting to “speculate responsibly” (Butler, 2009: 299) about the face validity of TALPS, Butler looked at supervisor perceptions of their students’ academic literacy, at students’ perceptions of their academic literacy abilities, and aligned this with the design of TALPS. The surveys carried out

with the supervisors and students were completed as part of a doctoral study by Butler and were discussed in greater detail in chapter 2. A summary of Butler's findings can be found below:

1. Supervisors are aware of the language problems of their students, especially problems with error correction. This issue is addressed in Section 7 of TALPS which is the 'Text editing' question. Students, on the other hand, rated themselves high in the functional literacy abilities. The face validity is crucial here because students do not believe they have problems with their academic literacy and will not be keen to undertake any intervention.

2. Supervisors believe that their students struggle with academic writing. Section 8 of the test requires students to produce a written text. If the test did not contain a section on writing supervisors would question the validity of the test. Students themselves see the importance of academic writing. Section 8 will therefore help students recognise the relevance of the test.

3. Supervisors agree on the importance of academic argumentation, that it is the most important type of text the student should be able to interpret and use. They also emphasised the importance of the use of relevant evidence in their writing and that specific referencing systems should be used. In terms of the face validity of the text, all texts used in the test are provided with the necessary referencing details. Students are also expected to make use of proper referencing in the argumentative text they write in Section 8. Argumentation and the cognitive processes associated with it are tested in sections 1, 3, 5 and 8.

4. While students acknowledge the importance of the correctness of their writing, they believe that they can edit their own texts. Evidence collected from a text they produced for the doctoral research indicates otherwise. What is important regarding the face validity of the test is that students see the relevance of an editing question in the test.

5. Supervisors also rated their students (except mother tongue speakers of English) as average to low on the functional abilities indicated in the definition of academic literacy. The implication for the face validity of the test is that supervisors would probably like to see some of these issues included in the test. (Butler, 2009: 291-300)

Butlers 'speculation' about the potential face validity of TALPS indicates that it does meet the "expectations of prospective users" (2009: 299).

4.4 A longer and more reliable test?

On the eight claims that were used above to test the validity of TALPS, the sum of the evidence points to the conclusion that a more elaborate and detailed validation might yield similarly positive results for TALPS. While these claims constitute a hypothetical basis for synthesising the evidence to validate this measuring instrument, we still need to refer to a final issue that was flagged above, namely how test length has in the case of TALPS impacted on its reliability.

The detailed discussion above regarding the reliability of TALPS is an indication of the importance of reliability measures in a test. The reliability of a test can be compromised by its length. It goes without saying, then, that the longer a test is, the higher its reliability measures could be. The first draft of TALPS included 173 items and took approximately two and a half hours to complete. While a test of this length may have high reliability measures, what are the implications of such a long test? In the case of TALPS, a test of this length would create a number of problems, the first being the time required to complete the test. If the test takes two and a half hours to complete it would require more complex and difficult logistical arrangements, and more expenses in terms of the cost of producing, marking and scoring a longer test.

In the framework employed in this study, the economic analogy within the technical is focused on issues of technical utility and frugality (Weideman, 2007a: 602). Weideman states that “it is no use, for example, that the test is utterly reliable, if its reliability is a function of its being five hours long. It may measure fairly, but the administration of the instrument consumes so much time that its *utility* is undermined” (Weideman, 2006a: 83). The choice that test developers are faced with, then, is to choose between a longer, more reliable test and a shorter, less reliable test. It is fortunate that, in the development of TALPS the shortening of the earlier versions in fact resulted not only in a shorter test, but also in a more reliable one, since the non-functioning items were culled.

There are limits, however, to such shortening, once the unproductive items have been discarded. Test developers are therefore always faced with a number of difficult decisions or trade-offs (Weideman, 2006a: 79). The trade-off discussed in chapter 3 had to do with the test developers choosing to leave in the test items that the factor analysis for TALL 2008 identified as outlying items, their reasoning being that the construct of academic literacy is a rich and complex one, and that test developers should therefore be willing to include these items despite the fact that they may compromise the reliability measures. Once again, with regards to the reliability of the test, developers can decide between a longer and more reliable test and a shorter test that maybe has a slightly lower reliability measure. Geldenhuys (2007: 71) points out that while tests can be valid and reliable or efficient, if they are too long, they may become less useful. The trade-off between reliability and utility is tied to the framework employed in this study, reliability being a constitutive condition and utility or usefulness being a regulative condition for test design. A detailed discussion of these terms (constitutive and regulative) has been done in chapter 3. Very briefly here, these terms derive from the theoretical framework employed in this study. In this framework Weideman uses the term ‘constitutive’ interchangeably with the term ‘necessary’ and the term ‘regulative’ interchangeably with ‘sufficient’ (Weideman, 2009a: 236). It is just one example of how important it is to consider all aspects in the design and development of a test, that developers need to recognise the ‘complementarity’ (Bachman & Palmer, 1996: 19) of the different aspects or qualities of tests rather than seeing them in isolation.

The process of the development of TALPS shows how the test was shortened from a 173 item test to a 76 item test, shortening as well the time needed to complete the test, from 150 minutes to 120 minutes. While the reason for excluding some of the items from the test was that they were weak items, Geldenhuys (2007: 74) points out that the test developers had originally planned a shorter test. The justification for this was that because of the broad range of postgraduate students at the University of Pretoria, the ideal would be to have a

short test that gave a "first, rough indication" (2007: 74) as to the academic literacy levels of the student, rather than waste time on testing students who may not need to be tested. Students who are shown to be at risk by a first, shorter test could then be required to take a longer test. The study by Geldenhuys looks at the possibility of shortening the test even further by including only a few of the subtests. To do this he uses the example of TALL (2007).

Table 4.6 Table of subtest intercorrelations (TALL 2007)

	Subtest	Total test	Subtest(s) 1	2	3	4	5	6
Scrambled text	1	0.46						
Graphic & visual literacy	2	0.67	0.25					
Text types	3	0.44	0.14	0.24				
Understanding texts	4	0.87	0.33	0.55	0.34			
Academic vocabulary	5	0.82	0.29	0.47	0.33	0.69		
Text editing	6	0.84	0.25	0.44	0.26	0.59	0.62	
Number of testees	:	3905	3905	3905	3905	3905	3905	3905
Number of items	:	65	5	8	5	22	9	16
Average test score	:	39.87	3.24	5.54	2.93	14.72	5.41	8.03
Standard deviation	:	13.35	1.86	2.19	1.28	4.57	2.61	5.15
SEM	:	3.34	0.73	1.11	0.83	1.93	1.16	1.44
Average P-value	:	61.33	64.79	69.25	58.59	66.90	60.11	50.18
Coefficient Alpha	:	0.94	0.85	0.74	0.58	0.82	0.80	0.92
GLB	:	0.97	0.93	0.81	0.77	0.85	0.82	0.96
Asymptotic GLB	:	0.96	0.93	0.80	0.77	0.84	0.82	0.96

(Geldenhuys, 2007: 80)

Geldenhuys explains that while the correlation between the subtests and the total test is satisfactory, three subtests show a higher correlation than the rest. These are **Understanding texts** (0,87), **Academic vocabulary** (0,82) and **Text editing** (called **Grammar and text relations** in TALPS) (0,84). If these have high correlations, can only these items, perhaps, be used to make up the test? What Geldenhuys points out is that this high correlation is to be expected because these three subtests make up a large part of the original test. While he goes on to prove that by carrying out partial correlations the three items mentioned above can "potentially constitute a shorter test" (Geldenhuys, 2007: 81), test

developers must question whether judgements like this should be made on empirical grounds alone. An empirically grounded test is not always the best test or the fairest test and decisions like this should be done 'judiciously' (2007: 81), after a consideration of a number of factors. Each subtest in the test assesses a component of academic literacy as outlined in the definition of academic literacy on which the construct of the test is based. Leaving out subtests means that students are now being tested on only a few of these abilities. Is the test, then, a fair and accurate reflection of the student's academic literacy? Geldenhuys advises that decisions made about what to include in a shorter test would require that the design of the test is changed or that the decision is made on other than only empirical arguments (2007: 81). While TALPS has been shortened, it still does include a number of subtests and a writing question, as indicated below:

Table 4.7 Table of subtests in drafts 1, 2 and final (TALPS)

Task type	Marks (1st draft)	Marks (2nd draft; pilot)	Talps Final
Scrambled text	15	5	5
Graphic and visual literacy	16	16	10
Dictionary definitions	5	5	
Academic vocabulary	40	27	10
Text type	5	5	5
Understanding texts	60	60	25
Grammar and text relations	22	22	15
Text edit	10	10	10
Writing section			20
Total	**173**	**150**	**100**

(Geldenhuys, 2007: 78)

What the above study indicates is that in the process of test development, developers are faced with decisions that cannot be made in isolation. Decisions or trade-offs like these must be done with all stakeholders in mind and should be

transparent to everyone involved. This transparency ensures that test developers become accountable for the designs and decisions about them.

4.5 Conclusion

A fair and responsible test is one that is not only valid and reliable but socially acceptable as well. While the focus of this study is a concern with the social dimension of testing, one cannot completely ignore any part of the testing process. Ignoring the empirical analyses of a test would be irresponsible – in the same way that the literature on language testing speaks volumes about irresponsible test developers/users that have ignored the social aspects. In order to document effectively the process of the development of TALPS, it is important to consider these facts. It should be the aim of test developers to design tests that are valid, reliable, accessible and transparent, by test developers being willing to be accountable for their designs. This chapter has focused on a discussion of the constitutive concepts (validity and reliability measures) of the test. By telling the story of how the test was developed, it has taken the first step in becoming accountable – in terms of openness about the theoretical claims and empirical analyses that have thus far supported the development of TALPS. The further challenge, of course, is how one might move from such theoretical accountability towards a public, and wider, defence of its design and uses. This we shall return to in chapters 5, 6 and 7 below, which will focus on the issues of transparency, accessibility and accountability. This is in keeping with the framework employed for this study, giving the readers as complete an analysis of TALPS as is possible.

Chapter 5

Transparency issues in testing academic literacy: The case of TALPS

5.1 Introduction

The focus of this study is an exploration of the concepts of transparency, accessibility and accountability as regulative conditions for a postgraduate test of academic literacy, the term regulative having been defined in detail in chapter 3. Much has been written in the literature on language testing about the need to be fair, ethical, responsible, and professional. The 'list' of terms calling for a concern for the social and other effects of language testing appears to be endless. What can be deduced from the available literature, as pointed out in chapters 1 and 2, is that there has been a positive move in the field to a consideration of the social dimension of language testing. Experts acknowledge that language testing cannot take place in isolation, without reference to context, or excluding the very people whose lives are most affected by the use of these test scores. Shohamy (1997; 2001), McNamara and Roever (2006), Bachman and Palmer (1996), and Hamp-Lyons (2000a; 2000b; 2001), among others, have pointed out the negative ways in which tests have been used, and the negative effects these have had on test takers. In addition, they have discussed what they believe test developers (and test takers) should do to ensure the development and use of fair and responsible tests. In light of these debates, this chapter will attempt to do the following:

- Provide a definition for the term 'transparency' with reference to its use in language testing;
- Consider the strategies used by the developers of TALL to ensure that the test is transparent, with a view to improving the transparency of TALPS;
- Propose the design of a website and a brochure for TALPS;

- Consider effective ways to promote the responsible use of TALPS.

5.2 Defining transparency

The concept of transparency has, in the recent past, become the watchword in government and politics, in the corporate world, in the media and even in the humanities and social sciences. Naurin (2007) states that transparency literally means that it is possible to look into something, to see what is going on:

> A transparent organisation, political system, juridical process or market is one where it is possible for people outside to acquire the information they need to form opinions about actions and processes within these institutions (Naurin, 2007: 2).

True transparency means that information is easily available to those who need it and that, importantly also, this availability of information allows an open dialogue between those within and those outside of the organisation. The urgency of the need for transparency across all fields has led to the creation of a global body called *Transparency International*, an organisation that "brings together relevant players from government, civil society, business and the media to promote transparency in elections, in public administrations, in procurement and in business" (*Transparency International,* 2010). Similarly, the importance of transparency has been highlighted through the formation of the Global Transparency Initiative, which is a society "committed to the idea of greater openness at the international financial institutions" (*Global Transparency Initiative,* 2006). This same organisation has developed a charter entitled "Transparency Charter for International Financial Institutions: Claiming our Right to Know". The charter is made up of nine principles: The right of access, automatic disclosure, access to decision making, the right to request information, limited expectations, appeals, whistleblower protection, promotion of freedom of information and regular review. While all nine principles may not necessarily apply across all fields, what comes through from the charter is the need for information to be freely or easily available.

The concept of transparency in the field of language testing has not been explored in great detail. While experts in the field of language testing have stressed the need for an open dialogue between test developers and test takers, for test takers to be able to ask questions about the tests and for test developers to take responsibility for their designs, this has not always been done. The first step towards achieving this begins with the concept of transparency because, realistically, one cannot ask questions about a process one knows nothing about, nor can one ask questions if one does not know who to direct these questions to. Weideman defines the transparency of a test as the "availability of information about its content and workings" (Weideman, 2006a: 82).

According to the framework employed in this study transparency and accessibility (Chapter 6) tie the technical qualifying dimension of the designed test to its juridical aspects (Weideman, 2009a: 249). This juridical analogy dictates, *inter alia*, that in addition to the applied linguist having to have a theoretical rationale for the design, the designer will need to publicly defend the design. According to Weideman

> in order to become more accountable, a first step is to ensure greater transparency, by making available as much information as possible about a test. Without transparency there is not enough information for others to assess the consequences of a test and to call the test designers to account. At the very least, a public description of the test should be available (Weideman, 2006a: 80).

Subsequent chapters and the final chapter explore further juridical echoes within the technical sphere, but the focus here is on how these relate to issues of transparency and openness to the public at large. Clearly then, transparency is a prerequisite to accountability. In order to ensure that the test process is transparent and open to scrutiny by everyone who might take a public interest in it, all information regarding the test must be easily available to all, i.e. in order for the entire testing process to be transparent, there must be accessibility. What, then, is accessibility in language testing, and how do we go about

ensuring that any test we design is transparent and accessible? These are some of the questions that the next chapter will attempt to answer.

5.3 The transparency of TALL

Once again we turn to the designers of TALL who present us with a number of ways they have proposed to make the test transparent and accessible. TALL and TAG are written during the academic orientation week which falls in late January/early February. At the University of Pretoria, students are made aware of the test from as early as May of the preceding year when over five thousand pamphlets on the academic literacy test are given out to prospective students in five faculties: Humanities, Economic and Business Management Sciences, Law, Natural Sciences and Theology. All first year students who have been accepted at the University of Pretoria receive a *Welcoming Day Brochure* and an *Academic Orientation Programme* by October of the previous year. Included in this brochure is a copy of the pamphlet (see below) on the compulsory academic literacy test. According to Weideman (2006a: 80) the designers print more than 17 000 brochures annually for distribution to every prospective student at the institution, both on application and at the university's Open Days.

Why do you have to write this test?

For many students with high academic potential, a low level of academic literacy can put their academic success at risk. The University of Pretoria therefore measures the academic literacy level of each new first-year student so that students can be placed in appropriate courses according to their level of literacy.

When should you write the test?

The test is taken during the academic orientation week (late January/early February). More information is published in the academic orientation week guide.

Where should you write the test?

The test is written on the main campus, or, if you are registered for courses there, at the Groenkloof or Mamelodi campuses.

What should you bring along?

Your student card or student number, a soft pencil (HB or B) and an eraser.

Can you prepare anything?

No, there is nothing that you can prepare. However, since the test is challenging, it is advisable to get enough sleep the night before.

What do we measure?

An example is available on our website (see **contact details** below). The ten things that we want to determine are:

- Is your academic vocabulary good enough?
- Can you interpret metaphors and idioms?
- Can you see how the parts of a text are linked together?
- Can you distinguish between different types of language or text?
- Can you interpret graphic information?
- Can you distinguish between main and supplementary information; cause and effect; fact and opinion?
- Can you classify issues and compare things?
- Do you know how to make deductions from information?
- Can you define a concept, argue a case and present evidence?
- Can you see what the greater meaning of something is that you have read?

In which language should you write?

You have a choice of writing in **Afrikaans or English.** However, students in the health sciences write only in English.

Where and when do you get the results?

The results, presented in one of five codes, will be posted on the notice boards next to the Humanities Building two days after the test.

How must you interpret the results?

Code 1: Extremely high risk
EOT 110 and 120 are compulsory.

Code 2: High risk
EOT 110 and 120 are compulsory.

Code 3: At risk
Two weeks after the test of academic literacy, you have the opportunity to be retested. If you then get a code 4 or 5, the conditions for these codes apply, and if you get a code 1 or 2, EOT 110 and 120 are compulsory.

Code 4: Low risk
You do not take EOT 110 and 120, but we recommend that you consider taking EOT 161 or EOT 162.

Code 5: Low risk to no risk
You do not take EOT 110 and 120, but we recommend that you consider taking EOT 161 (Academic reading); EOT 162 (Academic writing); EOT 163 (Legal discourse) or EOT 164 (Communication in organisations).

And if you obtained distinctions for your languages in matric?

There is not necessarily a correlation between the results of the academic literacy test and matric results in your language subjects (or other subjects). The test determines whether you will cope with language in the academic environment.

Where and when can you make enquiries?

Enquiries will be answered during the first three days after the results have been released.

Contact details:
Unit for Academic Literacy, Humanities Building,
Room 17-27, University of Pretoria, Pretoria 0002
Tel.: (+27) 012 420 2782 / 2334;
Fax.: (+27) 012 420 3682;
Website: A sample test is available through the UP website, on the departmental web page: www.up.ac.za/ual

unit for academic literacy
eenheid vir akademiese geletterdheid

University of Pretoria Pretoria 0002 South Africa
Tel: (+27) 012 420 3111 Fax: (+27) 012 420 4555
www.up.ac.za

The brochure is designed to answer questions that prospective students and parents have about the test. It includes an outline of the construct on which the test is based ("What do we measure?"). The brochure also informs students of how to interpret the results of the test. The results are not available as a 'Pass' or 'Fail', but are instead set out in five categories as indicated below:

> Code 1: Extremely high risk
> EOT 110 and 120 are compulsory.
>
> Code 2: High risk
> EOT 110 and 120 are compulsory.
>
> Code 3: At risk
> Two weeks after the test of academic literacy, you have the opportunity to be retested. If you then get a code 4 or 5, the conditions for these codes apply, and if you get a code 1 or 2, EOT 110 and 120 are compulsory.
>
> Code 4: Low risk
> You do not take EOT 110 and 120, but we recommend that you consider taking EOT 161 or EOT 162.
>
> Code 5: Low risk to no risk
> You do not take EOT 110 and 120, but we recommend that you consider taking EOT 161 (Academic reading); EOT 162 (Academic writing); EOT 163 (Legal discourse) or EOT 164 (Communication in organisations).

Van der Slik and Weideman (2005: 33) explain the reasons for releasing the test results in this way:

> In order to destigmatise the test results, the test administrators have devised several strategies. One is to make the results known not in two categories (pass or fail), but to grade results in terms of the measure of risk. In 2005, the results were published in five risk categories, from 1 (for very high risk) to 5 (little or no risk).

Interpreting the scores as a 'Pass' or 'Fail' is bound to de-motivate students who may already have a negative attitude to the test and to the intervention programme. A 'Pass' or 'Fail' also suggests that you either have the academic literacy proficiency required at this level or you do not have it. A student 'failing' the academic literacy test may believe one of two things: that the test

scores are incorrect or that there is nothing she or he can do to improve their academic literacy level. Failing suggests the end of the road. On the other hand informing students that they are “at risk” of failing because of their academic literacy levels indicates to them that this risk could potentially be eliminated by attending the compulsory academic literacy course.

Other important information in the brochure deals with the language in which students can write the test. At the University of Pretoria students can choose to write the test in English or Afrikaans. The intervention programme is also offered in English or Afrikaans. Students in the Faculty of Health Sciences do not have this choice and must write the test in English. Of all the other universities writing the TALL and TAG tests, it is only Stellenbosch University that requires all students to write the test in both English and Afrikaans. The brochure also answers the frequently asked question of whether there is any correlation between the students’ Grade 12 language score and the results of the academic literacy test. Full contact details of the Unit for Academic Literacy are included together with the website address. Students are also informed that there is a sample test available on the departmental web page.

The web page of the Unit for Academic Literacy is available through the University of Pretoria website. In addition to general information on what the unit offers, there is specific information on the test. The Academic Literacy Test link provides students with similar information to that which is in the brochure. It also provides students with a sample of the test they will be writing. While the test cannot be printed or downloaded, students can spend time studying the format of the test and, if interested, can complete the test or any number of items they might wish to attempt. Providing students with a sample of the test is one way of ensuring transparency. Very often what is most daunting about taking a test is the fact that you do not know what to expect. Allowing students to see a sample of the test means that they are offered the opportunity to familiarise themselves with the format and structure of the test.

Not knowing anything about what to expect in a test is very often its most frightening part. Certain questions that students and parents may have about the test can be answered by studying the sample test. It gives students an idea as to the way the questions are structured and what is required from them, how many questions there could be and how much time can and should be spent on each question. It is also helpful to encourage students who are in the intervention programme and have not looked at the sample test to do so to help in preparation for the tests that they write as part of the intervention programme.

According to Weideman (2006a: 80) further transparency is ensured through interviews with the test developers for newspapers and radio, popular articles in family magazines and talks and presentations to non-academic audiences. As stated already, every step of the process of the development of TALL has also been the subject of numerous academic articles. These have been published in accredited journals and presented at national and international conferences. In a sense, it is their findings that are often 'translated' for lay audiences after first being presented for expert scrutiny.

Using these strategies as a starting point, the rest of this chapter will attempt to outline the steps that must be taken to make information about TALPS as easily available and as accessible as information about TALL.

5.4 A web page for TALPS

An advance in technology in the last twenty years has ensured that one of the most valuable sources of information today is the World Wide Web, allowing one the luxury of accessing information needed at the touch of a button. It goes without saying then that the creation and development of a web page should be one of the first major considerations in terms of the transparency and accessibility of TALPS. The Unit for Academic Literacy already has available an informative website for students and those interested in finding out about the work done here. The University of Pretoria website, incidentally, is ranked at

No.2 of the top 100 universities in Africa and at No. 531 of the top 8000 universities in the world (*Webometrics Ranking of World Universities,* 2009). The web page for TALPS will be especially useful considering that the University of Pretoria has large numbers of postgraduate students from other parts of Africa and the world. Students anywhere in the world should be able to access information about the test. In addition to being able to access information about the test, the ideal would be to have an online test available. There have now been two successful experiments with the online version of TALPS at the University of the Free State (in 2010 and 2011), and it should become broadly accessible after 2012. Online tests, like the Test of English as a Foreign Language (TOEFL) and the International English Language Testing System (IELTS), have taken testing to new heights. Test takers now have the luxury of taking tests in the comfort of their homes. The TOEFL website is an interactive one, giving individuals every bit of information imaginable to successfully complete the test. While the focus of TALPS is entirely different to that of TOEFL, what can be drawn from TOEFL is the usefulness and importance of having an interactive, user friendly and accessible website.

TOEFL is a globally recognised test. Its reliability and validity measures are documented in the form of academic articles available on the website, yet many who take the test are not familiar with the meaning of or concerned with the importance of these measures. Many simply place their belief in the test because of its prestige. It is, as stated earlier, an internationally recognised test. One is led to question, then, how the test has achieved such status. The purpose in asking these questions is to determine how other tests, like TALPS, could begin to make inroads in the already competitive world of testing. In addition to the power of the TOEFL, two other factors have contributed to its success: 1) the promoting of the test, and 2) the website. The concern here is with the website. The issue of promoting the test will be dealt with later in this chapter. The test developers and creators of the TOEFL website have anticipated every

concern and query of the test taker, providing them with a wealth of information about the test.

The TOEFL website (*TOEFL*, 2008) has a page that deals specifically with the preparation for the internet based test. How to adequately prepare for the test is usually one of the key concerns for test takers. Test takers are provided with free sample questions, a 'tour' of the test that is designed to give test takers more information about the test and about its scoring, structure and format. In addition there are books, guidelines and programmes that can be bought to assist students. Clearly, TOEFL is as much a business as a testing agency. The success of TOEFL obviously lies in its ability to be sold as a business product. What is most appealing about the TOEFL website is its accessibility, both in terms of the information it makes available, and the ease with which the test can be taken. Test takers have the option of a paper based or internet based test. Being computer literate or having access to a computer are not prerequisites, nor is it necessary to travel far distances, as there are a number of test venues that are globally distributed.

While the intention is to create a web page for TALPS within the website of the Inter-Institutional Centre for Language Development and Assessment (ICELDA) (http://icelda.sun.ac.za), it is important that mention is made of TALPS on the University of Pretoria website or the website of any other university using the test. Information about TALPS can be made available under information for postgraduate students. A link from the university's website should take students to the TALPS web page within the ICELDA website. The discussion below details the information that should be available on the TALPS web page. See **APPENDIX C** for a sample of the TALPS Home Page.

5.4.1 The design of the TALPS web page

The TALPS web page should include the following information:

General Information about the test

About ICELDA

This should provide information about the consortium that is responsible for the development and standard administration of the test, for marking the test, and for assisting in the interpretation of their results.

What is TALPS?

This would include a general description of the test.

Information for students

Why do you have to take the test?

Here test developers can outline the purpose of the test as well as provide some background to the need for the test. One might profitably include information/research on the high failure rate at universities as a result of poor academic literacy levels. Students, parents and others not directly involved in academic literacy are sometimes quick to discredit support courses – this would be the ideal place to educate others about the importance of developing students' academic literacy. While the Unit for Academic Literacy website includes a discussion of what academic literacy is, the definition, especially, must be repeated here. There could also be a link here to the information on academic literacy on the TALPS web page.

How to prepare for the test

Students often ask what is the best way to prepare for the test. The advice given to students is that academic literacy is not an ability you can master overnight. Students cannot study for the test, but can prepare by having a good rest the night before the test. The other advice that can be added here is for prospective test takers to prepare by finding out more about the test

and to encourage them to look at and complete the sample test and the sample questions.

What do you need for the paper based test / online test?

In terms of the paper based test students need their student card, a soft pencil and eraser. Students will also need information about where and when they will write the test. In the case of the University of Pretoria, students can be asked to contact the Unit for Academic Literacy directly or this information can be communicated to them by the relevant departments. An online test would be the ideal as it helps avoid the administrative detail that must go into arranging such a test.

Information for lecturers and supervisors

Background to the development of the test

One of the main aims in proposing a web page for TALPS is to ensure that everyone affected by or interested in the use of the test has all the information that they need. The emphasis is on ensuring not just accessibility but transparency as well. By being transparent in our designs we are taking responsibility for the tests we create and the judgements we are making about test takers. Experts in the field have also pointed out the need for test developers to document the process of test design and development. Information here will not only be useful to students who want to find out more about the test, but to other test developers, teachers, lecturers, researchers, the testing community and the public at large seeking advice or information, not just on TALPS, but about the steps or process to be followed when designing or developing tests of academic literacy. The information included here will be a summary of chapter 2 of this study entitled “Telling the story of the test”.

The construct of the test

This can include a definition of what a construct is and a description of the construct of academic literacy on which the test is based.

Different sections in the test and what they test

As the heading states.

Reliability / validity of the test

Definitions of these terms already appear in the dictionary provided. Explanations here are more detailed, providing information about reliability measures of the test as well as about the validity of the test (construct, content and face validity). While this information may be helpful to students, it is also aimed at others interested in test development.

Definition of "Academic literacy" and other terminology

This provides a detailed definition of the term "academic literacy". It could also link to academic articles written on the subject of the TALL and TALPS as well as general articles on academic literacy written by experts in the field, as the examples available under the 'Research' tab of the current ICELDA (2011) website. This would be especially useful for researchers seeking additional information. It would be helpful to provide a language testing dictionary of sorts here, providing brief definitions and explanations of terminology that students and others need to understand. Including such terminology here makes it easier for the web page to be accessed. People simply typing in one of the terms on any search engine will be directed to our web page. Terminology that can be included here is: Academic literacy, Construct, Specifications, Blueprint, Reliability, Validity, Intervention and Support courses. We could also provide a recommended reading list should visitors to the site be interested in extending their knowledge or reading.

Interpreting the results of the test

This should include a discussion of what the test score means and the reasons for releasing scores in this way.

Information about the intervention programme

This would include the course outline/syllabus of any subsequent intervention that is on offer, including information on the tasks, assignments, mark schemes and other general information related to such a course. It would also highlight the fact that the test and the course are based on the same construct, showing some link between the two, that each is dependent on the other.

Sample test

At the moment there is only one TALPS test. Developing tests is an expensive and time consuming activity. There are plans to develop more tests. Once this is in place then an old TALPS test can be used as a sample test, or a sample test needs to be specifically developed. Like the sample test for TALL, students will be able to look at the test and even complete the test, but because of the copyright on the test they will not be able to download or print the test. With TALPS it would be ideal to have students complete the sample test. This has a number of advantages. One of the main problems students struggle with is the timing. Students complain that the test is too long. This is not the case. The real problem, more likely, is the reading speed of students, the other is that they do not use their time effectively, generally spending too much time on some questions. Completing the sample test will give students an idea as to how long they should spend on each question as well as an indication as to whether they need to work on their reading speed. The ideal would be to have a timer to indicate a start and stop time (as in the current online version of TALPS). The other advantage of having students complete the sample test is that students then have an opportunity to familiarise themselves with the different sections

and the structure of the questions. It has already been pointed out that the most daunting part of taking a test is in not knowing what to expect. There should also be a model answer that students should be able to access, or an automatic scoring system to allow students insight into where they have gone wrong.

Research articles about the test

Here interested parties should have access to articles specific to TALL and TALPS.

Test designers – information about them and their expertise

This would be one of the most important contributions to the web page. It has already been pointed out that test developers are often invisible in the testing process. How do we expect test takers and parents to see the test as credible if they have no information at all about the people who have designed the test? This would be a first important step in ensuring not just transparency and accountability, but also in opening up some kind of dialogue between everyone affected by the test. By proving this information the developers are not hiding behind their designs but are willing to publicly take responsibility for their work. Information here will be similar to what is already available on the website – photographs accompanied by brief CV's outlining our experience in testing, contribution to the test, and fields of research and teaching specialisations.

It is not enough that test takers know who the test designers are and a little about them, though that is a good starting point. More important is the open dialogue between test takers and test developers. This can be achieved by adding a Feedback page with a Comment Box facility.

Frequently asked questions

[These still have to be articulated]

Feedback Page/Comment box

This facility allows test developers to consider comments made about the test. Replies to these comments can be made via e-mail. This opens up a possibility of dialogue between test developers and the public. Test developers no longer disappear once the test has been designed, but are available to answer questions and defend their designs if there is a need.

Link to test reading speed

Link to academic vocabulary exercises

Link to Coxhead's Academic Wordlist

Link to a sample of the test

Accommodations for students with disabilities

The accommodation of test takers with disabilities has been touched on early in this chapter, but must be considered here as well. Websites are only useful if they are accessible to all users. An online test will be especially useful as test takers with disabilities will not have to travel distances to take the test. In terms of making the site accessible to all, expert webmasters have suggested the following:

1. Add alternative text to images: For users whose browsers do not support images, the alternative text is what they will see (or hear) instead of the image.
2. Add video and audio devices with text (written transcripts): Ensure that your website supplies written transcripts for all audio content, so that deaf web users can follow the audio clip.
3. Check that you can access all areas of your website without the use of a mouse: Make sure you can navigate through your website using just tab, shift-tab and backspace. If not, then neither can keyboard- and voice-only users. *(Grow your website*, 2009)

5.5 A brochure for TALPS (Appendix D)

In addition to the website, it is important, as has been shown with TALL, to make available to students a brochure giving students information about the test. These brochures could be handed out to students making enquiries about postgraduate studies. In fact, a brochure could be placed inside the front cover of every application/admission form handed to postgraduate students. Faculties and departments using the test or intending to use the test could also give these out to students making enquiries. It should contain the following necessary information:

What is TALPS?

TALPS is the Test of Academic Literacy for Postgraduate Students. It is a test that has been designed to test your academic literacy levels i.e. are you equipped with the academic literacy proficiency you will need at postgraduate level?

Why do you have to write this test?

A simple answer to this question is that your faculty or department requires you to write this test. Depending on the selection criteria in individual faculties/departments, this test could be used in one of two ways:

- to determine whether you gain access to the desired programme/field of study
  or
- it could be used to place you in a specific programme to help improve your academic literacy should the test show that you are at risk.

Where do you write the test?

Once you have applied for admission to the desired programme your faculty/department will inform you of the date, time and venue of the test. The test result will be sent to your department who will inform you of the outcome. No results will be released by the Unit for Academic Literacy. You can also take the **Online Test.** To access the online test go to: http://www.icelda.sun.ac.za. Before taking the online test, arrangements must be made with the Unit for Academic Literacy.

What do you need for the test?

All you need is your student card, a soft pencil (HB or B) and an eraser. If you are taking the online test you will need access to a computer and the internet for approximately two and a half hours.

Can you prepare for the test?

As this is a proficiency test, you cannot study for it. It is advisable that you access the TALPS website (http://www.icelda.sun.ac.za) where you can view a sample of the test. This sample test cannot be downloaded or printed out. You can, however, complete the sample test online. This will give you an idea as to what to expect in the test. The website includes other useful information and exercises related to TALPS.

What does the test measure?

The test has been designed to measure the academic literacy proficiency required from postgraduate students. It was designed to determine the following:

- Is your academic vocabulary good enough?
- Can you interpret metaphors and idioms?
- Can you see how the parts of a text are linked together?
- Can you distinguish between different types of language or text?
- Can you interpret graphic information?
- Can you distinguish between main and supplementary information; cause and effect; fact and opinion?
- Can you classify issues and compare things?
- Do you know how to make deductions from information?
- Can you define a concept, argue a case and present evidence?
- Can you see what the greater meaning of something is that you have read?
- Can you write an argumentative text using generally accepted academic writing conventions?

In which language should you write the test?

The test can for now only be written in English, as most postgraduate studies are conducted in this language.

Can you be exempted from writing TALPS?

If your faculty requires that you write the test you cannot be exempted from writing it, even if you have written other language proficiency tests or

completed language courses at other institutions. The test can only be written once, so if you have written it already at another institution make sure that you inform the relevant faculty or department.

How should you interpret your score?

CODE	INTERPRETATION
CODE 1 (0–33%)	**High risk**: EOT 300 is compulsory
CODE 2 (34–55%)	**Clear risk:** EOT 300 is compulsory
CODE 3 (56–59%)	**Risk**: EOT 300 is compulsory.
CODE 4 (60–74%)	**Less risk**: You do not need to enrol for EOT 300
CODE 5 (75+)	**Little to no risk:** You do not need to enrol for EOT 300

Where can you make enquiries?

It is advised that you first speak to your faculty/department. Any questions that cannot be answered by them can be directed to:

Avasha Rambiritch
Unit for Academic Literacy, Humanities Building,
Room 17-27, University of Pretoria, 0002
Tel: (+27) 012 420 4834/2334
Fax: (+27) 012 420 3682
Avasha.rambiritch@up.ac.za
Website: www.up.ac.za/ual

Can arrangements be made for students living with disabilities or who have special needs?

Yes. Should you be a wheelchair bound student, sight or hearing impaired, or a student with a reading or learning disability, arrangements can be made with the Disability Unit for you to write the test. Questions or queries can be directed to:

Mr. Juan Erwee
Disability Unit
Student Affairs Building
Room 2–13
(012) 420 2333/4281
www.up.ac.za/studentaffairs

5.6 Promoting the responsible use of TALPS

TALPS has until this point not been widely used or marketed at the University of Pretoria. One of the main reasons for this was the lack of staff to undertake the responsibility of administering TALPS at the University of Pretoria. As Butler's (2007) study found, large numbers of students would have to enrol for the postgraduate writing course. Accommodating that many students requires a larger staff component than is available at the Unit for Academic Literacy. Butler's (2007) study has, however, indicated a serious need for a test and an intervention of this nature, and institutions must make the necessary arrangements to allow students to be tested and to benefit from the intervention provided.

The TALPS web page and the brochure will go a long way in making as much information as possible available about the test. In addition to this, the test must be promoted effectively. A first step is to market the test within the institution. Many faculties/departments within the institution are unaware of the test but may want to use it for their students. Presentations could be made to these faculties about the value of the test and the intervention programme. Presentations could also be made to other institutions that may want to use the test. As was the case with TALL, the test developers must publish articles about TALPS in accredited journals as well as present papers/seminars at national and international conferences. By doing this, the test developers can get the opinions of other experts in the field. Expert opinion had already been sought in the design stage, especially when considering the face validity of the test. A test of this nature will constantly need refinement and this refinement is best done if it is assessed by others working in the same field. Presenting papers at conferences and publishing articles about TALPS is another way to advertise the test to other academics who may wish to use the test or, as in the case of TALL, become partners in the design and development of TALPS.

It is not enough that academics and experts be aware of the value of the test. True transparency means that information about the test is accessible to a non-academic audience as well. One of the focuses in ensuring transparency and accessibility is to ensure that information is available to all those affected by the use of the test results. This should not be limited to the test users and takers only. Scholtz and Allen-lle state that in the US the families and parents of prospective students are more conversant with the nature and implications of tests and test results (2007: 922). In discussing this point, Hamp-Lyons (2001) points out that tests exclude parents because of their 'technical' nature, that tests would be fairer to parents if tests were scored in ways that parents understood, and if test results/reports made sense to them.

One of the questions asked frequently by parents at the University of Pretoria Open Day deals with the need for a test like TALL. Many parents believe that their children are academically literate, basing this judgement on the fact that the child has performed well in English at high school level. What generally follows is an explanation by staff members of the Unit for Academic Literacy that the focus of the academic literacy test is entirely different to what students are tested on at school. Very often parents are not fully conversant with the term "Academic literacy" and this needs to be explained in detail. At the University of Pretoria, as at other universities, many students from previously disadvantaged backgrounds are accommodated. In conversations with students it is not unusual to find that that student is the first in their family to be enrolled for tertiary study, or to have parents that are illiterate or semiliterate. Parents need to be aware of the value of a test like TALL or TALPS. Both of these tests do have financial implications, as the intervention offered is an extra module that must be paid for. Parents must be informed about the test, the need for the test and the positive effects the test and the intervention can have. While the information will be available via the website, not all parents have access to or can use computers.

In addition to the website, it is important that the test be promoted in newspapers and magazines and on radio talk shows to ensure that the wider public become aware of its purpose. Advertising the test in this way may be expensive and cannot be carried out more than a few times, but even that may be sufficient to make information about the test available. The aim is to make as much information as possible available to all those affected by the test, in order to empower the test taker. The shift to recognising social realities in the field of language testing means that knowledge has become power. If information is available about the test, then test takers are equipped with the information they need about the test. They also know who to direct their questions to. To eliminate the most frightening part of taking a test, namely not knowing what to expect, it must be ensured that in the case of TALPS test takers need not fear the unknown. The test, its purpose, the developers, the construct of the test and the scoring system are not a closely guarded secret.

5.7 Conclusion

The TALPS website, the brochure handed out to students, the presentations made to academics within and outside the institution, as well as the marketing and promotion of TALPS, will go a long way in ensuring the transparency of the test. But is transparency enough? Does the implementation of these strategies ensure a channel of communication between test takers and test developers? Does ensuring transparency mean that accessibility is guaranteed as well? Before attempting to answer this question it is necessary to consider the definition of the term ‘accessibility’ so as to determine the role it will play in the field of language testing. Chapter 6 below focuses on the concept of accessibility and the ways in which test developers can make the test accessible to all those affected by the uses of the test results.

Chapter 6
The accessibility of TALPS

6.1 Introduction

In the epilogue of her book *The power of tests* Shohamy (2001) calls for the practice of "democratic testing". She explains that this requires "shared authority, collaboration, involvement of different stakeholders – test takers included" (2001: 161). In the same vein, Weideman points out that each test design reaches out to our fellow beings, that the design itself anticipates that human beings will use it, and that tests have consequences for real people (2006a: 84). If the testing situation requires such close linkages among all those involved, then the answer to the question posed in the conclusion to chapter 5, is that transparency is not enough. Transparency is simply the first step in becoming responsible test developers, and must be followed closely by the attention that will be given to accessibility. This chapter will attempt to answer two important questions:

1. What is accessibility in language testing?
2. How do we go about ensuring that we design and use tests that are accessible?

In order to do this, this chapter will do the following:

- Provide a definition for the term 'accessibility' with a focus on how it is used in the field of language testing;
- Consider the rights of the test taker with a view to determining whether these rights have been considered in the design and development of TALPS;
- Investigate the 'internal' accessibility of the test to determine how approachable the test is to test takers;

- Consider the data gathered from a questionnaire administered to students who took the test, in order to assess test takers' responses to questions about the accessibility of the test.

6.2 Defining accessibility

Like the concept of transparency, the concept of accessibility has not been the main focus of researchers in the field of language testing. A good starting point, in light of this, would be to look at a general definition of the term. According to the *Longman Dictionary of Contemporary English* (2004), access is defined as having to do with rights of entry and use. In one of the very few definitions of accessibility provided by language testers, Beddow, Kettler and Elliot (2008: 2) define accessibility in general as "the extent to which an environment, system, or product eliminates barriers and permits equal access to all components and services for all individuals". They define test accessibility as

> the extent to which a test and its constituent item set permits the test-taker to demonstrate knowledge of the target construct. Thus, an accessible test eliminates barriers; permits equal access to all components and features for all test-takers; and yields scores from which subsequent inferences do not reflect error that is the result of incomplete test-taker access (Beddow, Kettler & Elliot, 2008: 2).

As pointed out in the previous chapter, the regulative conditions of transparency and accessibility tie the technical qualifying dimension of the test to its juridical aspects (Weideman, 2009a: 249). The conditions of transparency and accessibility are very closely linked: Weideman (2007b: 43) explains that

> in providing a public defensibility of the test, one must begin with the notion of transparency, that the clearer we can articulate the purposes of the design for a larger public, and so achieve greater technical transparency, the easier it will be not only for others to access our designs and scrutinise them, but also for us to respond with integrity and honesty to public criticisms of them (Weideman, 2007b: 43).

The definitions above indicate that accessibility is about more than making available information to people. Accessibility is concerned with:

1. rights (in this case the rights of the test taker) of entry and use, and
2. ensuring that the test is approachable to test takers and others affected by the use of the test scores.

The link between 1 and 2 above, of course, is the issue of equality, and especially equality of access. This is already evident or implied in the above discussions, and will once more become evident in the debates that follow. The first part of this chapter will attempt to discuss the accessibility of TALPS with reference to these two concerns. Transparency is the first step when considering the social dimension of testing, but this first step is ineffective if not followed closely by the second, which is ensuring accessibility.

6.2.1 The rights of the test taker

Clearly, as can be inferred from the definitions above, the rights of the test taker are a first important consideration. According to *AERA* (1999), test takers have rights to the following: information about the test, the intended use of the test, the scoring criteria used for the test, the testing policy, confidentiality protection and information about the different test formats if there is a choice of format. *AERA* (1999) also stipulates that consent must be obtained from the test takers, scores kept confidential and test data files protected. In addition to this, test takers must be made aware of what constitutes dishonest behaviour, categories for score reporting should not be stigmatising, and score reports should be accessible to test takers (1999: 85–90).

The American Psychological Association (*APA*) (1998: 2) states that a test taker has the right to:

1. Be informed of your rights and responsibilities as a test taker.
2. Be treated with courtesy, respect, and impartiality, regardless of your age, disability, ethnicity, gender, national origin, religion, sexual orientation or other personal characteristics.

3. Be tested with measures that meet professional standards and that are appropriate, given the manner in which the test results will be used.
4. Receive a brief oral or written explanation prior to testing about the purpose(s) for testing, the kind(s) of tests to be used, if the results will be reported to you or to others, and the planned use(s) of the results. If you have a disability, you have the right to inquire and receive information about testing accommodations. If you have difficulty in comprehending the language of the test, you have a right to know in advance of testing whether any accommodations may be available to you.
5. Know in advance of testing when the test will be administered, if and when test results will be available to you, and if there is a fee for testing services that you are expected to pay.
6. Have your test administered and your test results interpreted by appropriately trained individuals who follow professional codes of ethics.
7. Know if a test is optional and learn of the consequences of taking or not taking the test, fully completing the test, or cancelling the scores. You may need to ask questions to learn these consequences.
8. Receive a written or oral explanation of your test results within a reasonable amount of time after testing and in commonly understood terms.
9. Have your test results kept confidential to the extent allowed by law. Present concerns about the testing process or your results and receive information about procedures that will be used to address such concerns *(APA*, 1998: 2).

Have the rights of the test takers, as outlined by *AERA* (1999) and the *APA* (1998), been considered in the design and development of TALPS? A main concern of both these bodies (*AERA* and *APA*) is that test takers be armed with all the information necessary about the test. As indicated in chapter 5, the brochure and the website for TALPS will provide test takers with as much information as possible about the test.

The information available to test takers includes information about the scoring criteria, something stressed by the *AERA* (1999) and the *APA* (1998). The sample test will, in addition, give them an indication as to the format of the test. Test takers will also receive information about the test and the use of the results of the test from the faculty or department in which they wish to register. There are contact details in the case of the University of Pretoria for the Disability Unit, should any disabled test taker have any concerns in this regard. Students

are informed of the need to take the test from as early as when they first apply to the university for postgraduate study. The test is administered and its results interpreted by trained individuals, as discussed in chapter 2. Every effort has gone into ensuring that test takers are treated with respect and that no test taker is discriminated against. While it would be ideal to be able to anticipate and answer every question a test taker may have, it is not always possible to do so. What has been done, to ensure that all questions/concerns are dealt with, is to provide test takers with information on how they can make contact with the test developers should they need to. This can also be done through the Comment Box facility on the website or by simply contacting the Unit for Academic Literacy, as indicated in the brochure.

To answer other questions related to the accessibility of TALPS, we turn once again to Kunnan's Test Fairness Framework (2004: 39), discussed in detail in chapter 3, where he identifies aspects of accessibility that could contribute to the practice and use of fair and responsible tests: financial, geographical, personal and educational access, and familiarity with test conditions and equipment. Financial access deals with the issue of affordability, geographical access with the access to test sites, personal access with the access of students with disabilities or impairments, educational access focuses on the opportunity to learn, and access to test taking equipment and test taking conditions focuses on whether students are familiar with these (Kunnan, 2000: 38). Each of these aspects will be applied to TALPS to determine its accessibility.

6.2.1.1 Financial access

Financial access, according to Kunnan (2004: 38) refers to whether a test is affordable for test takers. It will cost a student under R90 to write TALPS. In the case of some institutions, notably the University of the Free State, that amount is automatically included in all postgraduate application fees. In comparison to the cost of other similar tests, TALPS is very fairly priced. Some research around the cost of similar tests has shown TALPS to be the most

accessible to students in terms of cost. The TOEFL test can cost close to R900 (*TOEFL*, 2010), the International English Language Testing System (IELTS) test costs R1900 (*IELTS*, 2010), and the latest addition to online tests, Pearson's Test of English (Academic), costs R1300. The cost of the intervention programme, should students be required to enrol for, is borne by the students as for any other course or module. What is positive about the intervention programme is that, as in the case of TALL, it is built into the degree and is not an extra year or extended programme. All the student pays for is the extra module which will, in the long run, have positive benefits rather than having to complete an extra year. This has much more serious implications such as the payment of additional fees, and the possible loss of income for that year (what is conventionally termed "opportunity cost").

6.2.1.2 Geographical access

Geographical access, according to Kunnan (2004: 38), refers to whether a test is accessible to test takers in terms of distance. Should institutions/ faculties/departments choose to use the test for placement purposes, the ideal would be for the test to be taken in late January or early February of the year the student is to commence with the study programme. In this way foreign students and students living far distances from the university would, by this time, most likely have sorted out their accommodation and travelling. TALPS is administered at the Universities of Pretoria, Stellenbosch, Free State and the North-West. Students can make arrangements to write the test at the most convenient test site. Results will then be sent to the relevant institution. More test sites are now available, as well as an online version that has just been piloted, which relativises geographical access problems.

Students will be notified about the need to take the test from as early as when they first make an application to be admitted to postgraduate study. This is often as early as June of the year preceding the commencement of the degree.

This gives students adequate time to make arrangements to be available to write the test.

6.2.1.3 Personal access

Personal access, according to Kunnan (2000: 4), deals with the access that students with impairments and disabilities have to the test. The Unit for Academic Literacy has always shown concern for these students (wheelchair bound students, students needing extra time because of learning problems or students needing question papers with large print). A *Welcoming Day* brochure as well as a *Fact Finder* handbook refers all disabled students to the Disability Unit so that special arrangements can be made for these students to write TALL. Similarly, the TALPS brochure advises such students to do the same. *AERA* (1999) uses the terms 'accommodation' and 'modification' (1999: 101) to describe this. Accommodation refers to "any action taken in response to a determination that an individual's disability requires a departure from established testing protocol" (1999: 101). It explains that such accommodation may include a modification of the test administration processes or test content. It does not imply a change in the construct being measured (1999:101). *AERA* (1999: 103) identifies a variety of test modification strategies: modifying presentation format, modifying response format, modifying timing, modifying test setting, using only portions of a test and using substitute tests or alternative assessments. The purpose of all such accommodation and modification is, of course, the goal of ensuring equality, and equal treatment of all, as has been remarked above.

The two strategies that the Unit for Academic Literacy had to apply in the past were the modification of timing and the modification of the test setting. In terms of the modification of timing, students requiring extra time as a result of a learning/reading disability were allocated the required time. The Disability Unit provides students with documentation to explain the disability and the amount of extra time that should be given. This information is forwarded to the

Unit for Academic Literacy. A modification of the test setting is required for disabled students who cannot easily access the test venue. These are mainly wheelchair bound students. These students write at the Disability Unit, and the answer sheets are forwarded to the Unit for scoring. *AERA* also points out that:

> When modifying tests it is also important to recognise that individuals with the same type of disability may differ considerably in their need for accommodation. The modification should be tailored directly to the specific needs of individual test takers (*AERA*, 1999: 103).

The accommodation of students with disabilities and impairments in the Unit for Academic Literacy does not take place in isolation. Arrangements and decisions are made together with the help of the Disability Unit and the student concerned. Standard 10.1 has also been complied with:

> In testing individuals with disabilities, test developers, test administrators, and test users should take steps to ensure that the test score inferences accurately reflect the intended construct rather than any disabilities and their associated characteristics extraneous to the intent of the measurement (*AERA*, 1999: 106).

With reference to Standard 10.1, the scores of the test takers do reflect the construct being measured, which in this case is academic literacy. There has been no modification to the construct. In several respects, TALPS is a standardised test and remains so despite a modification of timing and setting. The disabilities of our students are not "directly related to the focal construct" (1999: 101), meaning that a student's academic literacy proficiency cannot be affected by their disability. Students requiring extra time are not simply allowed a "multiple of the standard time" (1999: 107). In keeping with the *AERA's* recommendation that

> professional judgment based on available evidence regarding the appropriate time limits given the nature of an individual's disability will be the basis for decisions (*AERA*, 1999: 107),

such decisions regarding the allocation of extra time are made by the professionals in the Disability Unit and complied with by the Unit for Academic Literacy. In order to further ensure accessibility of the test to

students with disabilities, test developers should heed the advice provided in the IMS Guidelines for Developing Accessible Learning Applications (*IMS Global Learning Consortium, 2009*) which states that (a) candidates are made aware of how they can request testing accommodations, (b) the requests for accommodations are examined by a panel of qualified experts, and (c) if the accommodation is approved, the test taker has adequate opportunity to become familiar with the accommodation before using it in the actual test (*IMS Global Learning Consortium,* 2009).

6.2.1.4 Educational access

TALPS has been designed to determine as early as possible whether postgraduate students' academic literacy skills will hinder success in their studies. The link between academic literacy levels and academic success has already been alluded to in chapter 2. Research has shown (Butler, 2007) that it is not just undergraduate students who struggle with academic literacy. Supervisors of postgraduate students have pointed out that poor writing skills, among others, seriously hamper students' academic success. Butler's (2007) study highlighted the need for a reliable testing instrument that could determine the academic proficiency of these postgraduate students. The test scores will not necessarily be used to deny a student access to a desired programme of study. It should be used in the first instance to determine whether the students' academic literacy ability will hamper success. If this is the case, the student could then be required to attend a year long intervention programme designed to develop the skills needed to cope with postgraduate study. According to Butler (2007) the intervention will address the following critical and learning outcomes:

- Identifying and solving problems in which responses indicate that responsible decisions using critical and creative thinking have been made;
- Working effectively with others as a member of a group;

- Communicating effectively in an academic environment using language skills; and
- Collecting, analysing, organising and critically evaluating information.

At the end of the course, students should be able to:

- Employ their heightened awareness of their own academic literacy and writing abilities in order to seek out actively opportunities and resources for the development of such activities;
- Apply adequately the knowledge and strategies with regard to the writing requirements of their specific disciplines in their own academic writing practice;
- Engage in writing activity as an integral part of academic literacy by making productive use of writing practised as a process; and
- Make productive (and continuous) use of opportunities for (or guidance as to what resources may be used in) the development of their basic proficiency in English (Butler, 2007: 226).

TALPS, like TALL, is being increasingly used by some departments as an access test. The decision of whether a test is to be used as a high stakes or medium to low stakes test is one that must be addressed even before the design stage. Shohamy (2001) highlights the many negative ways in which tests can be used, mainly by people in power, to deny or restrict access whether it is to a country as an immigrant, or to an institution, a career or a field of study.

South Africa's history means that we have firsthand experience of this. The demands of justice dictate that we attempt to undo the wrongdoings of the past. One way of doing this would be to open up opportunities to those who were previously disadvantaged. As discussed in chapter 1, unequal educational opportunity meant that after 1994 higher education institutions had to admit students who struggled with English as a language of learning. Intervention methods had to be applied to deal with this, since it was considered to be unfair to deny students access because of poor proficiency in a language.

The ideal would be for the test to be used for placement purposes, i.e. to help determine the academic literacy levels of students so as to provide an intervention should students require it, rather than for access. The developers of TALPS are of the opinion that should the test be used to determine whether a

student should be allowed access into a desired programme, the test should not be the sole determinant but should ideally be used in conjunction with other criteria. Writing the test could be an advantage to students and could provide them with the "opportunity to learn" that Kunnan (2004: 38) makes reference to when discussing educational access. Identifying problems early means that more effective solutions can be found before it is too late. The golden rule, however, is that students may not be disadvantaged by writing the test. Often, by the time they take the test, they may already have been accepted into a programme or degree. According to Weideman:

> The results of the test, which will be available within 48 hours for all candidates, will, however, have to be used with care. Since language cannot predict all of a candidate's future academic performance, decisions cannot be taken to exclude students on the basis of the results of a single academic literacy test. It is recommended, therefore, that those departments who wish to employ the results for decisions involving access, do not set the weight of the results of this test at more than is internationally agreed as appropriate for language tests, which is between 10% and 20%. A weighting of, say, 60% to prior academic performance is generally used, with language ability and several other criteria making up the other 40%. There is one possible exception to the rule of not excluding anyone purely on the basis of language ability. This is where the ability is so low (usually in the lowest 7½% of testees) that it raises ethical questions about allowing those in who so obviously fall short of requirements that they will waste their time and resources on a hopeless venture (Weideman, 2010).

6.2.1.5 Familiarity with test conditions and equipment

Kunnan's final point is based on the familiarity of the test administrators with test conditions and equipment. Here the concern is on whether administrators (and invigilators) have had prior access to test taking equipment and test taking conditions so that they are familiar with these (Kunnan, 2000: 4).

The administrators of TALPS (chief invigilators and invigilators) are normally required to undergo a compulsory training session before the administration of the test. They are also required to familiarise themselves with the "Standard procedures for test administration" (**Appendix E**) which outline procedures for

the administration of the test from start to finish. The test takers of TALPS do not have to be familiar with specific equipment, but it is important that they follow the instructions of the chief invigilator closely. The test is in a multiple choice format. Students are usually required to shade in the relevant answer on an answer sheet given. Section 8, which requires that students write an essay, is completed on the question paper. The answer sheet is an optical reader form that is computer scanned. Therefore it needs to be filled in very carefully. It follows that invigilators need to take students through each step of this process carefully. Invigilators are tasked, furthermore, to assist students who have questions or problems with the test administration/content. According to the *AERA*:

> Those responsible for educational testing programs should ensure that the individuals who administer and score the test(s) are proficient in the appropriate test administration procedures and scoring procedures and that they understand the importance of adhering to the directions provided by the test developer (*AERA*, 1999: 147).

This is adhered to by the test administrators.

The standard set of procedures for TALPS stipulates that it be written under examination conditions. Students are made aware of this. One would assume that postgraduate students would be familiar with the rules that apply when taking tests and examinations. Within the framework being used for this study, this administration of TALPS ties the "technical instrument to its social context" (2009a: 247). The test does not exist in isolation but regulates "interaction between test designers, test administrators, test takers (testees), administrative officials, lectures and others involved" (Weideman, 2006a: 83). They are all stakeholders in the testing process and they need to be fully aware of their rights and responsibilities. The testing process cannot be effective without the co-operation and commitment of these stakeholders and a sensitivity for the institutional environment of the test administration.

The administration of a test is one of the five qualities that make up Kunnan's test fairness framework. According to this, evidence regarding the following is collected:

> a) *Physical conditions*: This refers to appropriate conditions for test administration, such as optimum light and temperature levels and facilities considered relevant for administering tests.
> b) *Uniformity or consistency*: This refers to uniformity in test administration exactly as required so that there is uniformity and consistency across test sites and in equivalent forms, and that test manuals or instructions specify such requirements. Uniformity refers to length, materials and any other conditions (for example, planning time or the absence of planning time for oral and written responses) so that test takers (except those receiving accommodations due to disability) receive the test under the same conditions. Test security is also relevant to this quality, as a test's uniformity is contingent upon it being administered in secure conditions (Kunnan, 2004: 39).

TALPS is written in conditions (venues, lighting, temperature level, audibility) appropriate for tests and examinations. TALPS is a standardised test and is administered as such. The training session that all invigilators undergo ensures that there is uniformity across test sites. All test takers are subject to the same instructions and have the same amount of time to complete the test (except for those students with disabilities and impairments for whom special arrangements have been made). Strict security measures are adhered to regarding the test.

6.2.1.6 Protecting the rights of the test taker

For a final point regarding the rights of the test taker, we turn to a number of issues first broached by Shohamy (2004: 86–87), who identifies the following as ways in which the rights of the test taker can be protected:

(a) Questioning the uses of tests

Shohamy states that test takers should see it as their civil right to question tests, test results and test methods. With TALPS, questions from test takers and others affected by the use of the test are welcomed by the test developers. The

Comment Box facility on the TALPS website will allow direct communication between the test designers and the test takers.

(b) The right not to be tested

According to Shohamy, test takers should also have the right to refuse to be tested. She states that tests are sometimes in the hands of powerful institutions, making it difficult for the individual to refuse to be tested. In the case of TALPS, there is no valid argument for the test taker to refuse to take the test. If the test is being used for placement, it may not disadvantage the test taker but rather should be seen as an advantage, as it identifies problems early enough to be rectified. If the test is being used for access purposes, then a very low test score is an indication to students of their chances of completing their studies successfully.

(c) Privacy and confidentiality

Shohamy's point here is that there should be accountability with regard to the purpose of the test, its practice, and its methods. The concept of accountability will be discussed in chapter 7. What is important here is that the test results of TALPS will be available only to the test taker, the supervisor and the lecturer in the intervention programme. The results will not be used for any purpose other than to determine the academic literacy level of the student. Should the results of the test be used for research purposes, the statistical analysis will be done on the group as a whole and not on individual respondents. Butler points out how important it is to share the test results with students; students should not only read a "final score from a list" (2007: 227), but the results should be discussed with each student individually and in detail so that they are fully aware of how their specific difficulties can be addressed by the intervention programme.

(d) Alternative forms of assessment

According to Shohamy this refers to the fact that test takers should be granted the opportunity to be assessed in other ways besides the traditional test. She states that “such information can be used to counter evidence against decisions based on tests only” (2004: 87). In the case of TALPS, students write the test before the start of the degree/programme. At risk students may then be admitted to the compulsory intervention programme. This course follows a continuous assessment procedure. It also follows a task-based approach where students complete a number of smaller tasks as well as a major assignment, the marks for which constitute the mark for the course. The course is divided into two semesters and the final mark for the course is an average of the first and second semester mark. It appears that in this regard the rights of the test taker have, on the whole, been carefully considered in the design and implementation of TALPS and the intervention programme. Students who for whatever reason believe that they have been unfairly discriminated against in the test have a number of opportunities in the form of tasks and assignments to demonstrate their academic literacy ability. It is also recommended that students write TALPS again at the end of the course. This would be a good indication of whether the course has helped develop the academic literacy skills of students. While the marks for this test will not be used for assessment purposes, the data collected will be useful to both test takers and test developers. Test takers will be aware of exactly which abilities need further development and test developers will be aware of what changes, if any, need to be made to the test/intervention but more importantly, these results could be used to determine if the test and the intervention are doing what it should be doing i.e. assessing academic literacy proficiency and providing effective support. A point that could do with further elaboration at this stage in respect of TALPS is that of articulating exactly what diagnostic information might be gleaned from it, and aligning that insight with the instructional design of the subsequent course.

(e) Sharing discourse

Here Shohamy makes reference to the need for a "dialogue" (2004: 87) between the testing community and ordinary citizens. Equally important to her is the need for everyone affected by the use of tests to be aware of the "techniques and terminology of the testing community" (2004: 87). She states that "testing cannot remain a field that belongs only to testers, but rather test takers and the public at large need to be part of a mutual discussion" (2004: 87). In addition to ensuring that information about TALPS is easily available, the TALPS website provides the Comment Box facility which can help facilitate some discussion between test takers and test developers. The references in the previous chapter to the various strategies to ensure transparency for TALPS and its administration are again relevant here.

6.2.2 Internal accessibility

The concept of accessibility needs to be considered from as early as the design stage right through to the implementation of the test. While a detailed exposition of TALPS has been given in chapter 2, the focus here is on the internal accessibility of the test. A main concern in this regard is the question: How accessible is the test to test takers?

The cover page of TALPS (**Appendix F**) indicates to students the date, mark and duration/length of the test. It also includes general information about the test. This is also discussed by the invigilator. Students are required to date and sign that page to acknowledge that they understand what the purpose of the test is, and that they give permission for information gathered from the results of the test to be used for research purposes. This is repeated by the invigilators. The cover page includes the test instructions. Once again, this has been explained in detail by the chief invigilator but is available here in the event of questions/queries. There are trained invigilators walking around the venue to be of assistance should students require it.

Each section in the test is numbered and labelled according to the task type it is, e.g. **Scrambled text** or **Academic vocabulary**. Below this appears an instruction to students. Some of these instructions explain what students need to do e.g. “Study the following text and graph and then answer the questions that follow” or “Choose the best possible answer from the list of options”. Others are more detailed explanations of what students need to do and these generally include an example to assist students in understanding better what is required e.g.

> Section 4
> **Text types**
> The sentences below are examples of different text types, such as advertisements, instruction manuals, academic textbooks and the like. You must match an item from the first set (26–30) with an item from the second set (A–E). For example, if you think that the language of 26 comes from the same text type as B, then mark 26B as your answer on the loose sheet.
>
> **OR**
>
> Section 6
> **Grammar and text relations**
> In the first of these three texts that follow, some words have been deleted. The possible **places** where they may have been deleted are marked with a / sign. Select, from the places marked A/, B/, C/, D/, where you think the word is missing. The first two have been done for you as examples.

The mark allocation for each question and total marks per section are indicated. Graphs are labelled and sources are available for these as well as for the text used in the **Understanding texts** section.

‘Internal’ accessibility therefore goes hand in hand with careful, thoughtful design, and ties in with the deliberate choice of topic, theme and content, which we now consider below.

6.2.2.1. Choice of appropriate content and material

An important consideration in the internal accessibility of the test has to do with the kind of material or content that the test developer uses in the test. This

is referred to by *AERA*: “To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers” and “[e]xpert judges may be asked to identify material likely to be inappropriate, confusing, or offensive for groups in the test taking population” (*AERA,* 1999: 44). The *Code of fair testing practices* (*APA*, 1988: 2) states that “[t]est developers should: Review and revise test questions and related material to avoid potentially insensitive content or language”, while Nevo and Shohamy (1986: 115), in producing nine standards for testing, identify “Fairness standards” which are intended to ensure that a “testing method is conducted legally, ethically, and with due regard to the welfare of tested individuals” and that “tests are based on known and accepted subject matter and criteria”. Kunnan’s advice in this regard is that test developers should “review and revise test questions and related material to avoid potentially insensitive content or language” (Kunnan, 2000: 2).

The choice of appropriate material is therefore essential in ensuring a fair and accessible test (see Norton & Stein, 1998). The range of students writing TALPS is very broad. They include local as well as international students, students of different races, gender and age, from the Honours student having recently completed his/her undergraduate degree to the middle aged (and older) student whose last degree was completed some twenty years ago, and is now required by his/her employer to further his/her studies. Important, as well, is the fact that the students taking TALPS are from different faculties and fields of study. Using discipline specific material may not be possible, but the material selected should be suitable for this varied audience. The safest option would be to use generic material that is nationally as well as internationally relevant, current topics that would be of interest to all people irrespective of sex, age or gender, and with language that is non-judgemental, non-racist and non-sexist.

6.2.2.2 Text accessibility

The test developers of TALPS had one other consideration. The test was designed specifically for postgraduate students, and in addition to the content being relevant and inoffensive to all test takers, it had to be of an appropriate level in terms of difficulty. Text difficulty or text accessibility is, according to Fulcher (1997: 497), an important but neglected topic in applied linguistics. He points out that factors that make texts more or less difficult include poor linguistic structure, contextual structure, conceptual structure and unclear operationalisation of the reader-writer relationship (1997: 497). He warns teachers and testers alike to guard against using only a readability rating (like the Flesch formulae) to determine the accessibility of the text. He advises that in addition to this, testers should seek out expert rating on the text chosen (1997: 499).

The text finally chosen for TALPS fulfilled these requirements. The text was one published in a well known popular academic journal – this is particularly important as this is the kind of writing that students, especially postgraduate students, are encouraged to comprehend. The Flesch Reading Test is, moreover, helpful in determining the reading level of the text being used and is an easily accessible tool. The Flesch Reading Ease Test helps determine how easy or difficult a text is, using a score of between 0 – 100. The easier the text is, the higher the score. In the case of TALPS, the text used in the **Understanding Texts** section has a readability index of 36.4. This score can be translated to determine a grade level. According to Garger (2008) a score of about 65 correlates with the 8^{th} to 9^{th} grade level, a score of about 55 indicates a 10^{th} to 12^{th} grade level and scores between 0 – 30 represent graduate level readability. A score of 36.4 is appropriate for postgraduate students. In terms of seeking out expert ratings for the text chosen, it has already been pointed out in chapter 2 that though the test developers of TALPS were highly experienced in the field of test design, expert ratings were sought, and the text deemed suitable for the test.

6.3 The TALPS questionnaire

Issues of transparency and accessibility cannot be resolved without the voices of the people most affected by the test. True accountability, which is the focus of the next chapter, dictates that test developers do not carry out their work in isolation, but that there is constant dialogue between all those involved in the testing process. The last part of this chapter therefore takes a detailed look at the data gathered from a questionnaire (**Appendix G**) administered to postgraduate students who took TALPS. The purpose of the questionnaire was to elicit information, comments, questions and reactions from the testees about the test. This personal contribution of the testees is important to this study, since without it part of the social impact of a test cannot be determined.

A Likert scale (Likert, 1961) was used in this questionnaire to measure student responses. In explaining the workings of a Likert scale, McIver and Carmines (1981: 22–23) state that:

> A set of items, composed of approximately an equal number of favourable and unfavourable statements concerning the attitude object, is given to a group of subjects. They are asked to respond to each statement in terms of their own degree of agreement or disagreement. Typically, they are instructed to select one of five responses: strongly agree, agree, undecided, disagree, or strongly disagree. The specific responses to the items are combined so that individuals with the most favourable attitudes will have the highest scores while individuals with the least favourable (or unfavourable) attitudes will have the lowest scores (McIver & Carmines, 1981: 22–23).

The Likert scale uses either numerical (1–5) or alphabetical (A–E) values. In the case of this questionnaire numerical values were used:

1 – Completely Disagree
2 – Disagree
3 – Neutral
4 – Agree
5 – Completely Agree

The questionnaire also comprised five open-ended questions which allowed students to voice their opinions and give brief reasons for their choices.

The first page of the questionnaire outlines for students the purpose of the questionnaire and the aim of the research. A definition for the term “Academic literacy” is provided should students be unfamiliar with the term. The questionnaire presents students with a number of statements related to the concept of academic literacy, so it is important that students are aware of the abilities that encompass academic literacy. There is a letter of consent informing students that the research is part of a doctoral study, participation is voluntary, confidentiality is guaranteed and that ethical clearance for the research has been obtained from the university. Full contact details of the researcher was available. The researcher was available at all test sessions where the questionnaire was administered so as to assist students with questions or queries.

The statements in the questionnaire have been designed to elicit responses from the testees about different aspects of the test and the testing process: definitions of academic literacy, opinions about the accessibility of the test, the rights of the test takers, preparation for the test on the part of test takers, students’ feelings about being considered at risk and their familiarity with the content of the intervention programme they may have to take if the results of the test show that they may be at risk in terms of their academic literacy levels.

6.3.1 Participants

The questionnaire was administered to students who took the test between 2008 and 2010. Students were asked to complete the questionnaire after they had completed the test. A total of 150 questionnaires were administered to students applying for admission to Master’s degrees in the following fields: Environmental Studies, Taxation and Agricultural Economics. None of the three groups of students (Environmental Studies, Taxation and Agricultural Economics) that filled out the questionnaire were writing the test for selection purposes. At the time that the questionnaire was administered, TALPS was a

fairly new test and not widely used. This would explain the small number of students who were tested at the University of Pretoria during that period.

The researcher received a total of 98 completed questionnaires. Not all questionnaires administered were completed by testees. There could be a number of reasons for this. TALPS is a relatively long test (2 hours). At the end of a test of this nature, some students may have been physically and mentally exhausted. Another reason could be the time factor. Many of these students are working full time and had taken time off from work to write the test. An extra ten or fifteen minutes to fill out the questionnaire meant more time off work.

In the case of the students applying for admission to the MPhil: Taxation, they were here at the university for a block session of two weeks. During this block session they were expected to attend an orientation for the programme they would be registering for, take TALPS, as well as attend introductory lectures on the courses they would have to enrol for (African Tax Institute, 2010a). Their days were full and busy, with very little time in between. Many of these students chose not to fill out the questionnaire. All students who were enrolling for the MPhil: Taxation, were expected to take TALPS but were not required to enrol for the compulsory intervention even if they were shown to be at risk. The reason for this was that these were national and international students who would not be able to attend the lectures. The co-ordinator of the postgraduate programme did indicate that should the course be available as an online module they would be happy to have their students take it. It must be noted here that testing students who have already been admitted to a programme of study, but doing nothing to provide support, is a futile exercise, especially if the test results indicate that the student needs support. These students were informed of the need to take the test through information in the brochure handed out to prospective students.

The following information is provided to students enrolling for the MPhil: Taxation:

> Apart from the entrance examination written by some candidates only, *all* the MPhil: Taxation students who are eventually registered for the Masters Program in Taxation, will be required to write the University of Pretoria's English Language Proficiency Test on Wednesday 17 February 2010. (Note that the cost of this language proficiency test will be borne by the ATI.) Only students who can provide proof that they have recently passed a TOEFL test, will be excused (African Tax Institute, 2010b).

Students' requiring admission to the MSc: Agricultural Economics had been informed that a prerequisite to admission to the programme was that the students sit for TALPS. Should the results of the test show them to be at risk they would be required to enrol for the compulsory intervention. Prospective students were informed about the need to take the test at an information session held before the start of the semester. No other information about the test was given to students. With the exception of informing students about the compulsory intervention, not much information about the course was given.

The same applied to students enrolling for the M.A. and MSc: Environmental Studies. On enrolment, students had to agree to be tested. Should the results of the test show them to be at risk, they were required to enrol for a special bridging course which included the compulsory intervention (EOT 300) and STK 100 or BME 120 (Basic statistics). An information brochure handed out to prospective students provides the following information:

> Language proficiency and computer literacy
> For admission, applicants need to pass the standard language proficiency test applicable to all first-year undergraduate students at the University of Pretoria. Exemption is given to students who wrote this test as undergraduates. If a student does not pass the test, we recommend bridging training in order to improve your English reading and writing skills in a professional environment. This training involves the masters-level core module Environmental Paradigms (ENV 810) as well as the English language course EOT 300. On an individual basis we may recommend further topics at honours level for particular students.

> Students who have successfully completed the bridging training have automatic acceptance to continue with the masters-level course work for their degree (Centre for Environmental Studies, 2010).

As can be seen from the information above, only the most basic information regarding TALPS and the intervention programme were given out to students. The brochures did not provide the contact details of someone whom students could direct questions to regarding the test or the intervention. In the case of the students in the M.A./MSc: Environmental Management, the information provided was incorrect. Students would not be writing the undergraduate test (TALL) but TALPS (Centre for Environmental Studies, 2010). Also, it is not the "University of Pretoria's English Language Proficiency Test" (African Tax Institute, 2010b) as indicated above, but the Test of Academic Literacy for Postgraduate Students (TALPS).

6.3.2 An analysis and interpretation of the results of the questionnaire

The data collected from the questionnaire were analysed using the Statistical Package for the Social Sciences (SPSS) (Version 17.0). The questionnaire, which comprised of twenty-four statements, rendered a reliability of .736. (Cronbach's alpha). According to Santos (1999:1), Cronbach's alpha determines the internal consistency or average correlation of items in a survey instrument to gauge its reliability. He points out that a Cronbach's alpha of 0.7 is an acceptable reliability coefficient and that "lower thresholds are sometimes used" (Santos, 1999: 1). An alpha of .736 is therefore acceptable for a questionnaire.

The item-analysis output from SPSS provides an Item-total statistics analysis. One of the columns in this analysis is the "Alpha if item deleted" column. According to Gliem and Gliem (2003: 86), this column represents the scales for Cronbach's alpha reliability coefficient for internal consistency if the individual item is removed from the scale. In the case of the TALPS

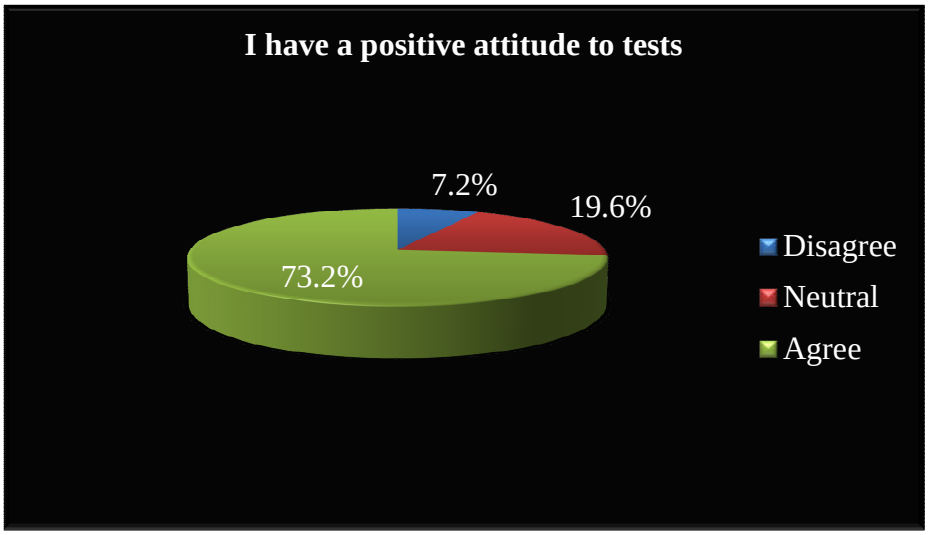
questionnaire, the analysis indicated that if the statement: *"I think that one needs to prepare specifically for all tests one has to write"* was deleted the Cronbach's alpha would go up to .746. If the statement: *"Test takers have little or no rights"* was deleted the Cronbach's alpha would improve to .750. The views of test takers for both these statements are important to this study. As indicated, a Cronbach's alpha of .736 is acceptable for a questionnaire as the questionnaire is not being used to make high stakes decisions about the testees. Rather, the purpose of the questionnaire was to elicit responses from the testees to see how the test developers could improve the accessibility of the test. The researcher chose to settle for a lower reliability rather than leave out these crucial questions.

The last part of the questionnaire includes five open-ended questions. The views and opinions of students drawn from these open-ended questions will be integrated into the discussion below. These views and opinions appear in the exact words of the student. Also, while the questionnaire used a 5 point scale, in analysing and interpreting the results of the questionnaire the numerical values 1 and 2 (Completely disagree and Disagree) were combined and 4 and 5 (Agree and Completely agree) were combined. Graphs therefore indicate 3 or 4 columns: Disagree, Neutral and Agree or Disagree, Neutral, Agree and Missing (where students did not answer the question).

6.3.2.1 The power of the test

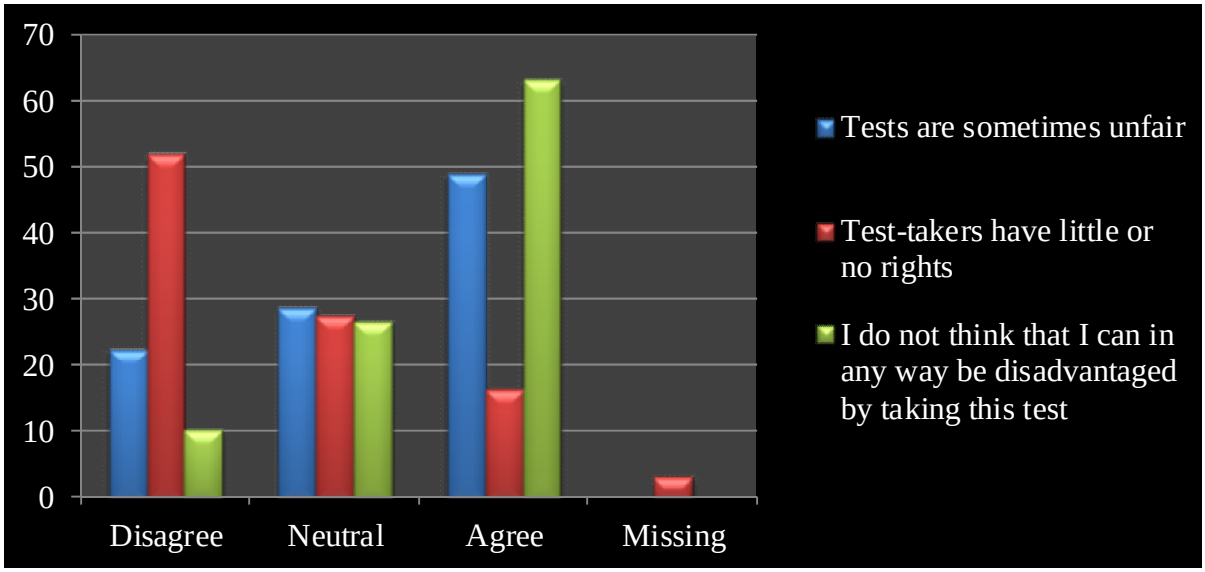
A first important consideration was to determine how these students felt about tests and the rights of test takers in general. Should the data have shown that these students have a negative attitude to tests and that they believe that test takers have little or no rights, their opinion of the test, the need for the test and the intervention would have been negatively affected by these feelings. The data show, however, that overall, these students do not have a negative attitude to tests (Figure 6.1):

Figure 6.1 Students' attitude to tests



Despite the majority (73%) of students having a positive attitude to tests, 49% of students agree that tests are sometimes unfair. Surprisingly, in light of this, is that only 16% of these students believe that test takers have little or no rights. Sixty-three percent of these students do not believe that they can be disadvantaged by taking this test (Figure 6.2). Student responses to this last statement could be motivated by the fact that these students had been informed beforehand that the results were being used for placement rather than selection. This could also be why these students do not believe that test takers have little or no rights. It is possible that had these students been informed that the test was being used for selection purposes, their responses may have been different. Overall, one can assume that these students do not have a negative attitude to tests, nor do they believe that, as test takers they do not have any rights. Less than half the students believe that tests are unfair.

Figure 6.2 Student perceptions of tests, test taker rights and TALPS



6.3.2.2 Defining academic literacy

It was also important to determine whether students understood the concept of academic literacy. When students were asked to define the concept of academic literacy (*What is your understanding of the concept of academic literacy?)*, only 6% of the students did not understand the concept. Their answers ranged from general statements like:

- It is the formal language that we use in school;
- The level of how you can cope with language used at university;
- The level of language used at tertiary level;
- Use of academic language appropriately;
- How academic reports and writing is done;
- Different from language skill. Used in an academic environment;
- Basic language understanding one needs to have for academic purposes at tertiary institutions;
- Language skills required to advance one's academic career;

to more specific definitions of what academic literacy entails:

- The ability to read, understand and write in such a way that it will yield positive results in my academic career;
- The ability to understand, analyse and critically evaluate academic literature;
- Is to be able to read, write and interpret academic information;

- Academic literacy is demonstrating proficiencies in reading, writing and doing simple calculations without the aid of a calculator;
- How you read and understand academic literacy and the way in which you relate this information;
- It means to be able to read and write on a professional manner that is acceptable.

Interestingly, only 28% of students identified reading, writing and the ability to critically analyse what was read, as important aspects of academic literacy. So while students have a general understanding of the concept, not enough students are aware of the abilities that encompass academic literacy. This indicates a need for further information about the concept of academic literacy to be available to students.

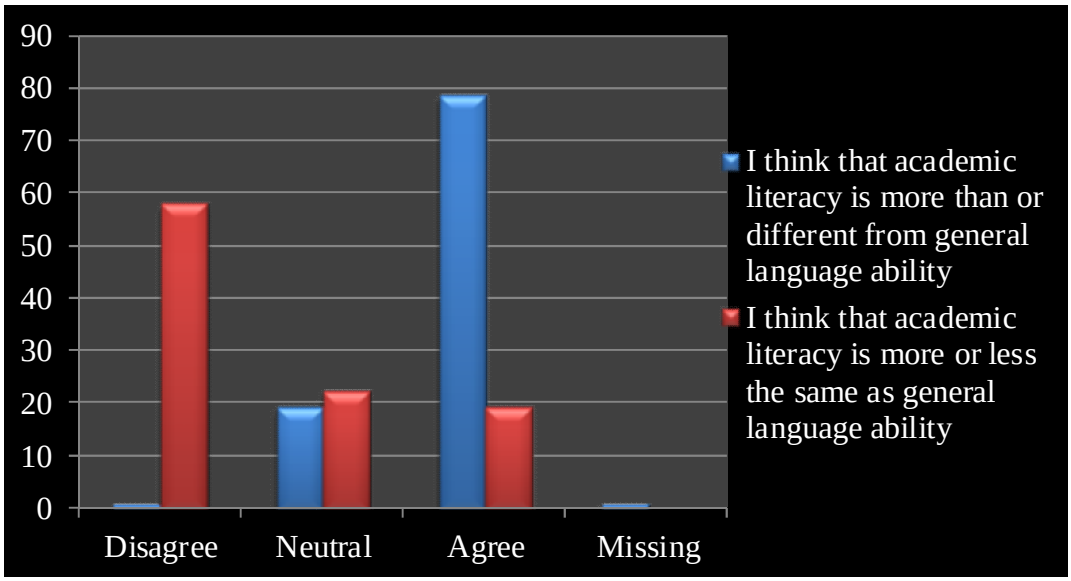
6.3.2.3 Academic language versus general language ability

While these students have a positive attitude to tests, and have a basic understanding of academic literacy, there seems to be a problem with their understanding of the difference between academic language and general language ability. As has been pointed out earlier in this study, students often believe that being proficient in the English language means that they are academically literate, that if they passed Matric or in this case an undergraduate degree, they do not have problems with their academic literacy levels. This misconception has negative effects. Students may be unwilling to accept a test score that shows them to be at risk in terms of their academic literacy levels, and may prefer to believe that the test score is inaccurate. Should they be expected to take an intervention programme to help them improve their academic literacy levels, this will be done reluctantly, with the student believing that she or he does not need the intervention because she or he does not have a problem.

The results of the questionnaire indicate that there is a very mixed response as to whether students are aware of the difference between academic literacy and general language ability (Figure 6.3). Students were first presented with one

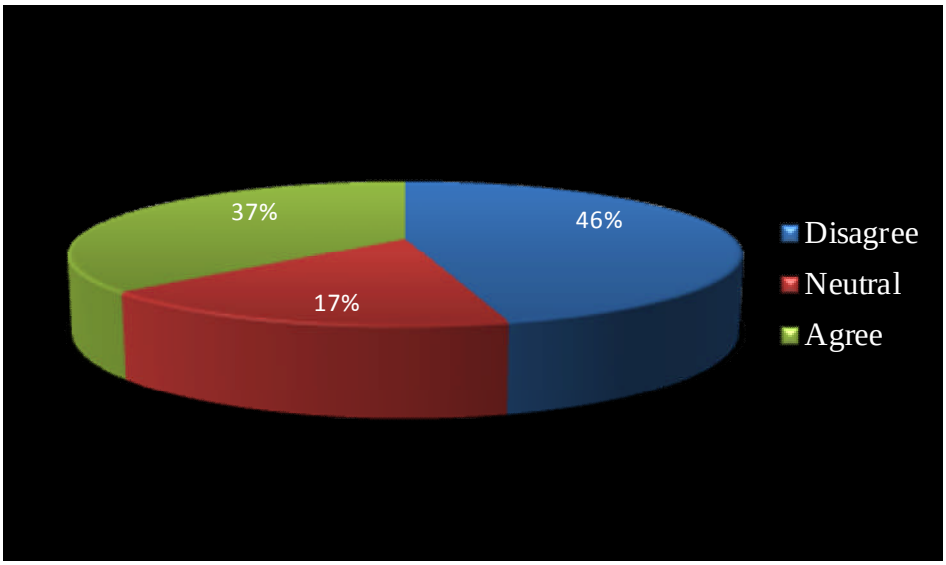
statement, phrased in two different ways so as to determine whether they understood the difference between the two:

Figure 6.3 Academic language versus general language ability



Seventy-nine percent of students agree that there is a difference between academic literacy and general language ability (*Academic language is more than or different from general language ability*). When the statement was phrased differently, presenting students' with the statement that *"Academic literacy is more or less the same as general language ability"*, only 58% disagreed with this statement. One would be tempted to say that students were simply confused with the phrasing of the statements (though these are postgraduate students who should understand the difference between these two simple statements). However, this explanation cannot be applied if we look at their responses to the next statement. When presented with the statement: *"If one is good at languages, one should have no problems coping with academic language"*, 37% of students agreed (Figure 6.4). In light of the fact that 79% of students agree that there is a difference between academic language and general language ability, one would have expected the majority of students to acknowledge that being good at languages does not necessarily ensure that one is good at academic language.

Figure 6.4 If one is good at languages, one should have no problem coping with academic language

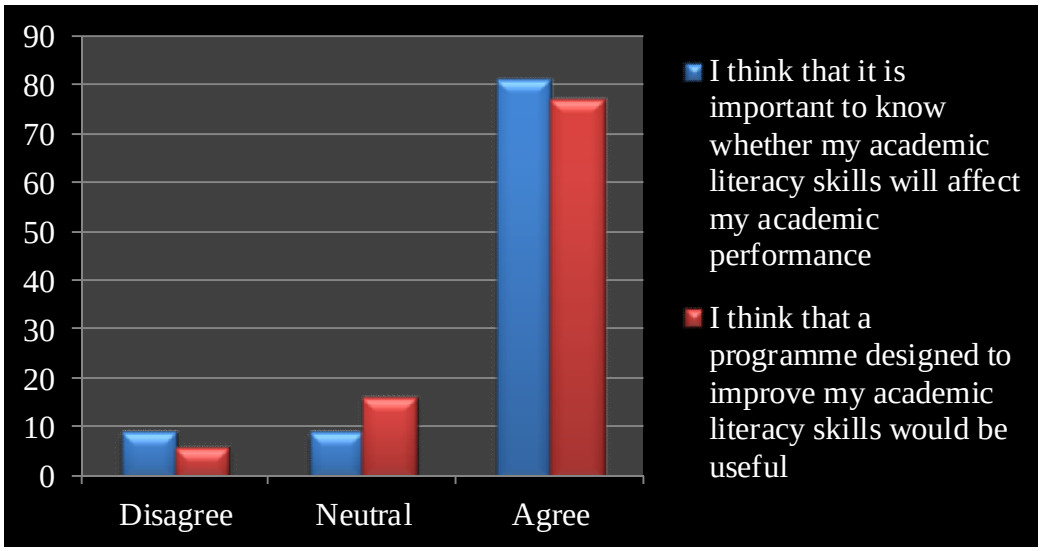


From this we can deduce that while students are aware that there is a difference between academic language and general language ability, they do not understand the difference clearly enough. It would have been expected that students who agreed with the first statement should have disagreed with the second and third statement. Instead there is a difference of 21% for the second statement and 33% for the third. One would assume that by postgraduate level students should be clear about the difference between academic language and general language ability. Unfortunately this is not the case.

6.3.2.4 Literacy skill and academic performance

Despite this, however, students are well aware of the importance of academic literacy as can be seen from their response to the next two statements:

Figure 6.5 Literacy skills and academic performance



Eighty-one percent of students believe that it is important to know whether their academic literacy skills would affect their academic performance (Figure 6.5) (*I think that it is important to know whether my academic literacy skills will affect my academic performance*) while 77% of students indicated that a programme designed to improve their literary skills would be useful (Figure 6.5).

6.3.2.5 Yes. I will take the test

When students' were asked, *"Do you think a test of this nature is at all necessary?"* 71% of students agreed that it was:

- Yes, people in South Africa still need to keep an internationally accepted standard; standards from schools are lower so it affects universities;
- Yes, it is. It shows whether one will be able to read and understand their research and also help in compiling a research report;
- Yes, to know if you need extra help in academic literacy;
- Yes, it highlights areas where an individual can improve as well as give a lecturer an understanding of the level of proficiency in his/her class;
- Yes, I think it is necessary to help those who do not have the skills required;
- Yes, if one is not academically literate a student will waste their time by possibly needing more time to complete a thesis or obtaining poor marks on a thesis.

Twenty-two percent of students did not believe that a test of this nature was necessary. Their justification for this ranged from the fear of being seen as a failure, to the idea that they had completed their undergraduate qualification at an English medium institution and were therefore deemed to be proficient in the language. Yet again we are presented with the belief, on the part of these students, that being proficient in English means you are academically literate:

- No, you might think you are a failure if you fail it;
- No, it makes me feel like my English is being tested and it makes me feel like I have to prove myself before the actual program starts;
- No, because by the time one reaches postgraduate studies one can be assumed to be proficient;
- No, I am a postgraduate student and at my level I understand the English language well;
- No, because some students are from the institutions that are English oriented and they already passed the languages and therefore specialised in their fields;
- No, English is English. If you passed it and did undergrad and graduated then you know and understand what is required from you;
- No, it is not necessary because in most African countries lessons are taught in English. It is good for non-English speaking students;
- No, because up to postgraduate level English has been the main language.

One way of dispelling these misconceptions that students have about the link between proficiency in English and academic success is by making available to students information about academic literacy, the test and the intervention.

6.3.2.6 Yes. I need some help

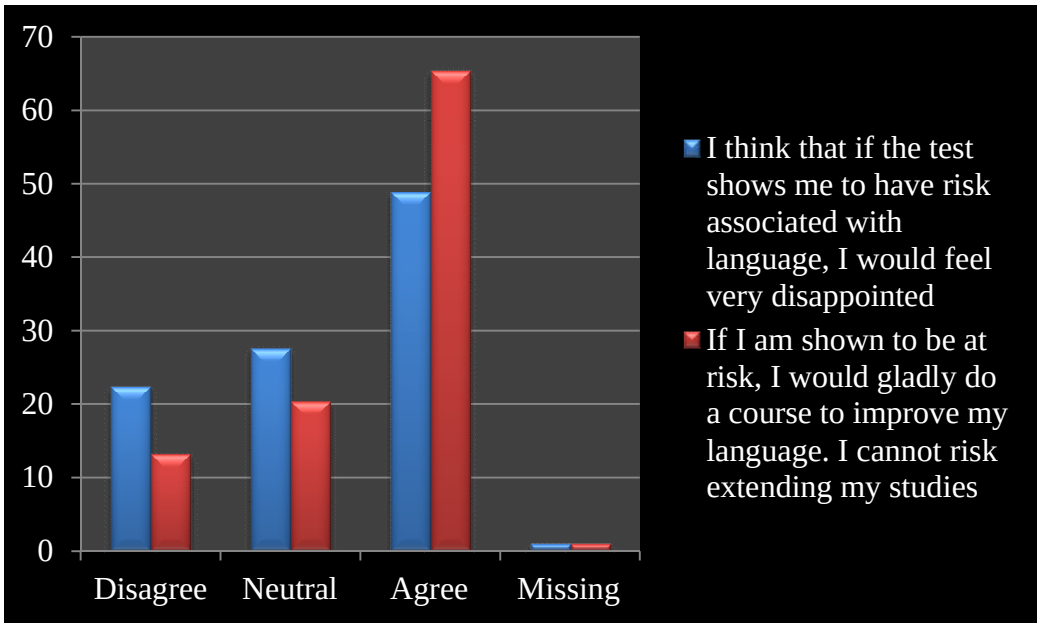
A pleasant surprise was students' indication that only 49% of them would feel disappointed if the test showed them to be at risk (*I think that if the test shows me to have risk associated with language, I would feel very disappointed*) (Figure 6.6). One would have expected a much higher percentage considering the stigma attached to being seen as not being academically literate. When presented with the open-ended question *"Would you feel in any way stigmatised if the results of the test show you to be at risk? Explain briefly"*, 57% of students indicated that they would not feel stigmatised. Their

willingness to know their academic literacy levels, and to be presented with an opportunity to improve, can be seen from the following statements:

- No, it will only help me in the future;
- No, I will have a chance to improve;
- No, it motivates me to work extra hard;
- No, I will follow the recommendation;
- Not really, the test has given me the opportunity to work on my weakness so I will be using my results to ensure I do that;
- No, it will allow me to improve where I am weak;
- No, it will be an eye-opener and an observation of my difficulties;
- No, because I am the one studying, it is my responsibility and if the course will help me then I will do it.

When students were presented with the statement, *"If I am shown to be at risk, I would gladly do a course to improve my language. I cannot risk extending my studies"* some 65% indicated that they would be (Figure 6.6). They acknowledge that they cannot risk extending their studies.

Figure 6.6 Student feelings about being shown to be "at risk"



As indicated in chapter 1, there is generally a stigma attached to such 'support' courses. Students do not want to be seen as needing support. Butler's study (2007) has highlighted the fact that while supervisors believe that students need

some intervention, students often believe otherwise. These students, however, seem to be aware of the importance of their academic literacy levels. The main reason for this could be that these are postgraduate students who may be well aware of how poor academic literacy levels can impact on their studies. When asked to discuss how they would feel about taking a compulsory academic writing course (*If you were shown to be at risk, how will you feel about taking a compulsory academic writing course?*), 55% of students' indicated that they would be happy to do so, as can be seen from some of their comments:

- I would gladly do it;
- Not bad, it will enhance my study abilities;
- That would be fair enough because I would be given another chance to improve my abilities;
- It would be to my advantage as it will help me with the writing of my dissertation;
- I would be grateful that the university has such courses to ensure that I am up to scratch in academic literacy to pursue my studies and successfully complete it;
- I think it will help a lot.

Thirty-five percent of the students were unhappy about taking the course. Of this 24% indicated time and finances as reasons for this:

- If the course is free I will take it but if I have to pay it is not possible;
- This will disadvantage me with regards to finances, time and travelling. I may have to pay for the course that I did not budget for and also attending classes at the different time to the programme I registered for;
- There is not enough time for me to do an extra subject.

These are realistic concerns. The intervention programme is another module the student would have to pay for. Also, the majority of these students are employed full time and will be attending lectures after hours. What students need to understand is that should the test show them to be at risk, the successful completion of their studies, and being able to do so within the given time frame, depends on them improving their academic literacy levels. As has already been pointed out, without this intervention students may fail or take

even longer to complete their studies. This would have more serious financial implications. The cost of the intervention at the University of Pretoria in 2010 was R2 240.00 as compared to the R16 800.00 it would cost to extend one's studies by a year, and the opportunity costs could be more than ten times as much. Twenty-seven percent of students indicated that they would be very unhappy to undertake the course:

- I would not accept it;
- No, I do not intend to undertake the course;
- No, this test is not actually the true reflection of my knowledge;
- I would not like to take it at all. It would make me feel depressed and daft. Not happy at all because the course would be irrelevant to my planned or envisaged studies;
- Devastated;
- I would decline it and pursue my career elsewhere.

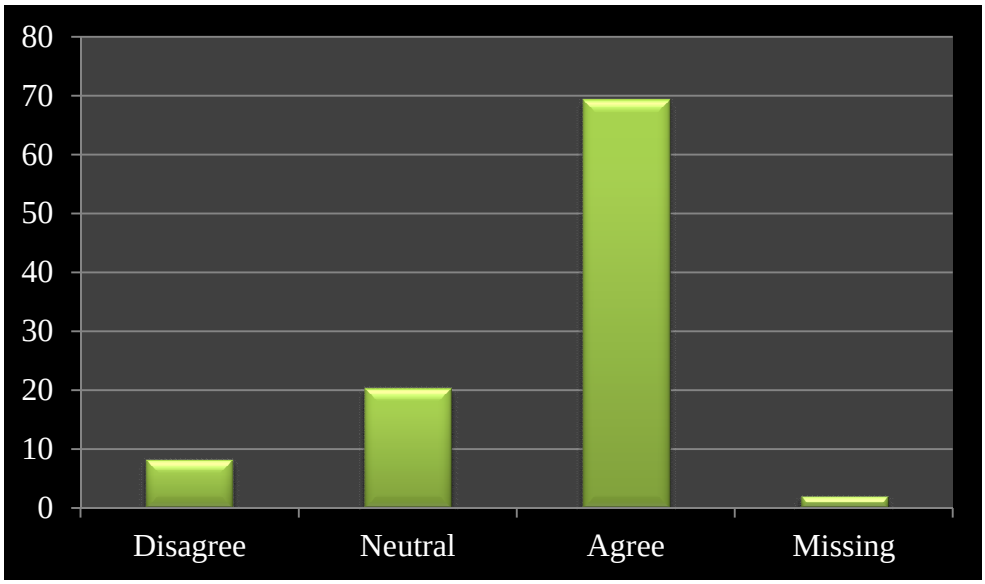
The main reason for students feeling this way could be that they did not have enough information about the test and the intervention. If students were well informed about both, they may see the necessity for the test and the value of the intervention. This is where the transparency of the test features – the brochure and the web page as well as the promotion of the test (to faculties/department within the university and to the general public) are crucial so as to give the test takers all the information they will need. A channel of communication between testers and test takers is necessary to help test takers answer questions they have about the test, the testing process and/or the intervention.

6.3.2.7 The accessibility of the test

While the first part of the questionnaire related to academic literacy and language tests in general, the second part of the questionnaire deals specifically with issues of accessibility as they relate to TALPS.

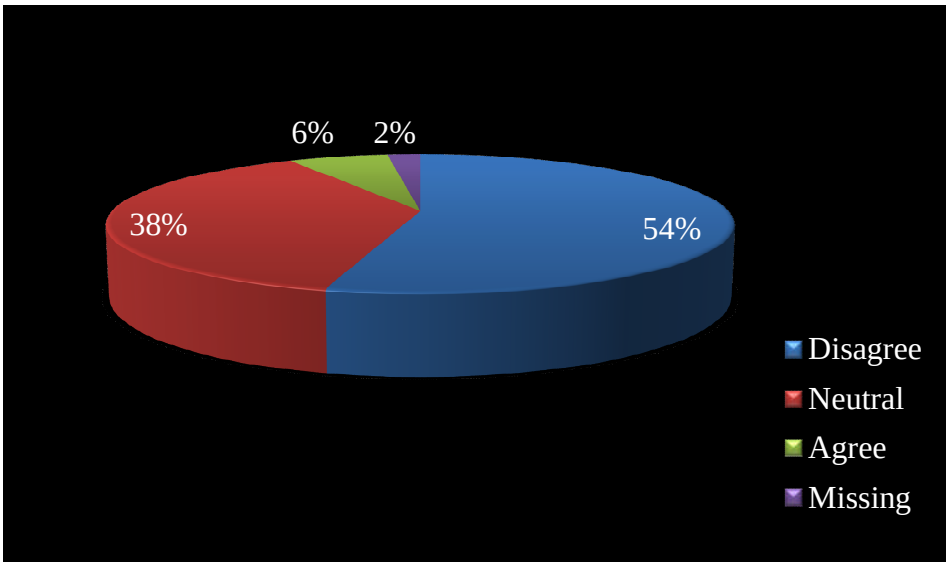
A first important statement deals with whether students were aware of the purpose of the test. Here 69% of students indicate that they were aware of the purpose of the test (Figure 6.7):

Figure 6.7 I am well aware of the purpose of the test



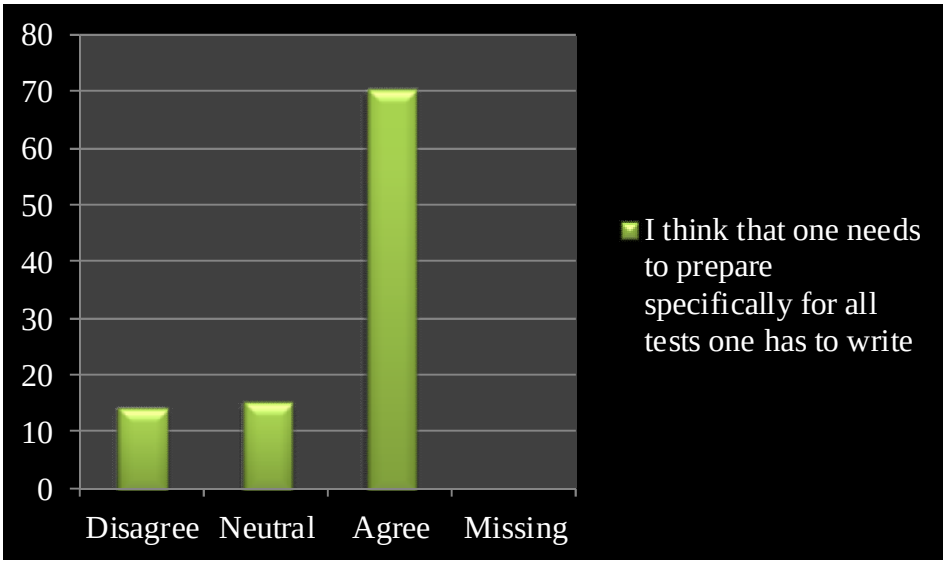
Despite being aware of the purpose of the test, however, 54% indicate that they were not well prepared for the test (Figure 6.8). If the majority of students were aware of the purpose for the test, why were they not prepared for the test?

Figure 6.8 I was well prepared for the test



Related to this is the indication by 70% of students that they believe that one needs to prepare specifically for all tests that one has to write (Figure 6.9).

Figure 6.9 I think that one needs to prepare specifically for all tests one has to write



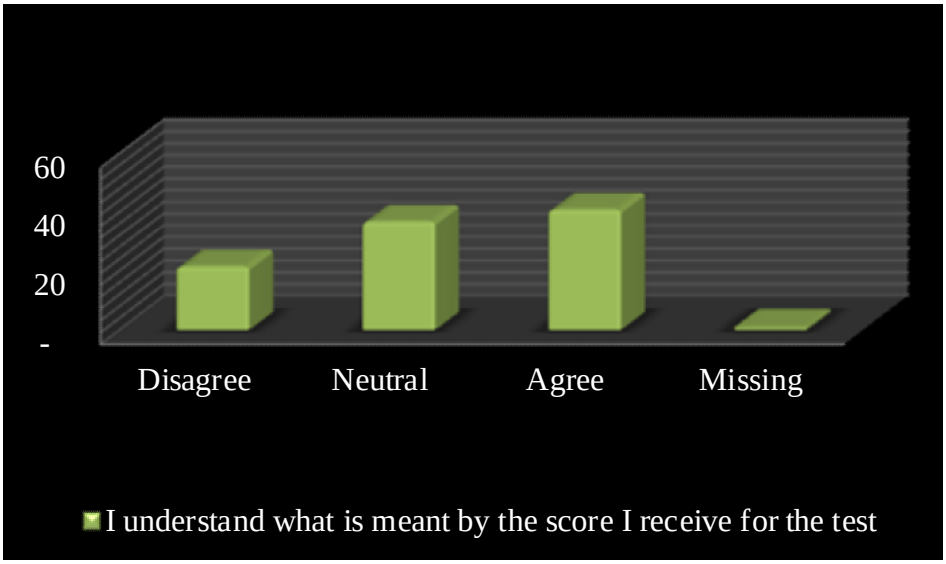
So while the majority of students think that preparation is important, why were more than half of the students not prepared? One of the reasons for this could be that despite students believing that preparation is necessary, there is also the belief that one cannot 'study' for tests of this nature. What is important here is the difference between the use of the word 'prepare' as opposed to 'study'. Yes, one cannot study for an academic literacy test, but one can prepare for the test by being aware of what the test is about and what it 'looks' like in terms of its format and structure. One can also prepare by ensuring that one is well rested before taking the test. This kind of preparation is essential for any test one writes. As pointed out earlier, sometimes the most daunting part about taking a test is the fact that one has no idea about what to expect. When students were asked, *"How do you think you could have best prepared for the test?"* students did indicate the need to have an idea of what to expect in the test as a means of preparation:

- If we had a slight idea about the scope it intended to test;
- Had it been communicated to me beforehand, the length and the duration it would have been better in the sense that I would have prepared myself psychologically;
- If we were given the format before writing the exam that could have helped me prepare for the test;

- If I at least had an idea of what to expect from the test;
- If a pre-test was available;
- Given a scope or prescribed book to read that could have prepared me better;
- If we at least saw one test from last year;
- As I was not aware of the questions that will be asked, I was not prepared at all;
- Get the previous questions to see the set up of the test;
- Be well informed about the content of the test. Also provide past exam questions so that candidates know the structure of the exam;
- By being fully informed what the test is for – the areas that the test aims to test.

The sample test that will be available on the TALPS webpage will go a long way in addressing these concerns.

Figure 6.10 I understand what is meant by the score I receive for the test

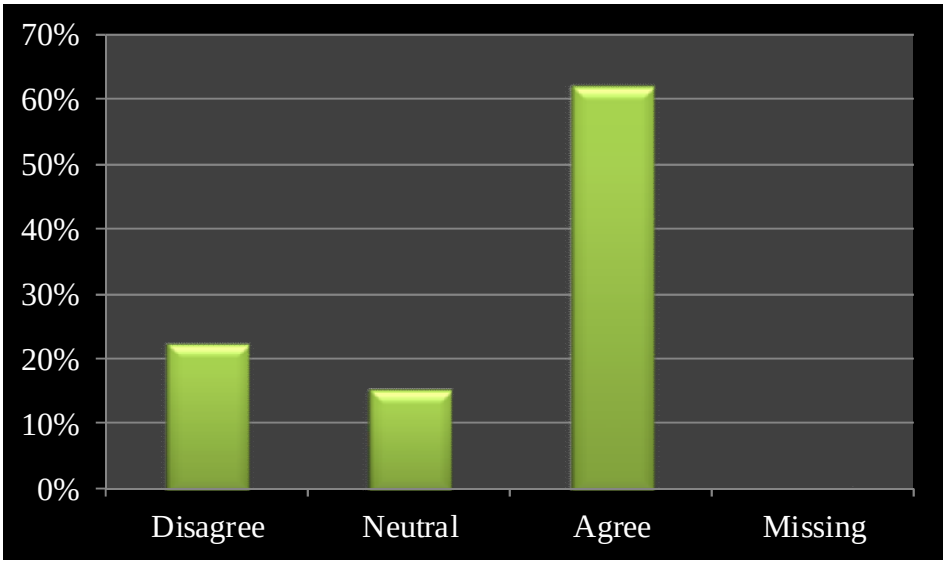


Only 41% of students indicated that they understood what is meant by the score they will receive for the test (Figure 6.10). In terms of ensuring accessibility to test takers, this is not a satisfactory percentage. Every student taking the test must understand what is meant by the score. The brochure that will be given to every student making enquiries about postgraduate studies at the University of Pretoria includes a description and explanation of the five codes that the test results are released in. This brochure will be available on the TALPS website and students will be able to download a copy. Information about the

interpretation of these codes will be available on the website under "Information for Students". In the case of TALL, chief invigilators are asked to explain this to students before the test. The same is done for TALPS.

Sixty-two percent of students indicated that they understood all the instructions (Figure 6.11). The instructions in the test are clear and not difficult to understand. Some of these instructions include an example to help students.

Figure 6.11 I understood all the instructions



Interestingly, while 62% understood the instructions, only 34% indicated that they understood all the questions. This could be an indication of the problems these students may have with the reading and the understanding of academic texts and/or vocabulary.

Figure 6.12 I understood all the questions

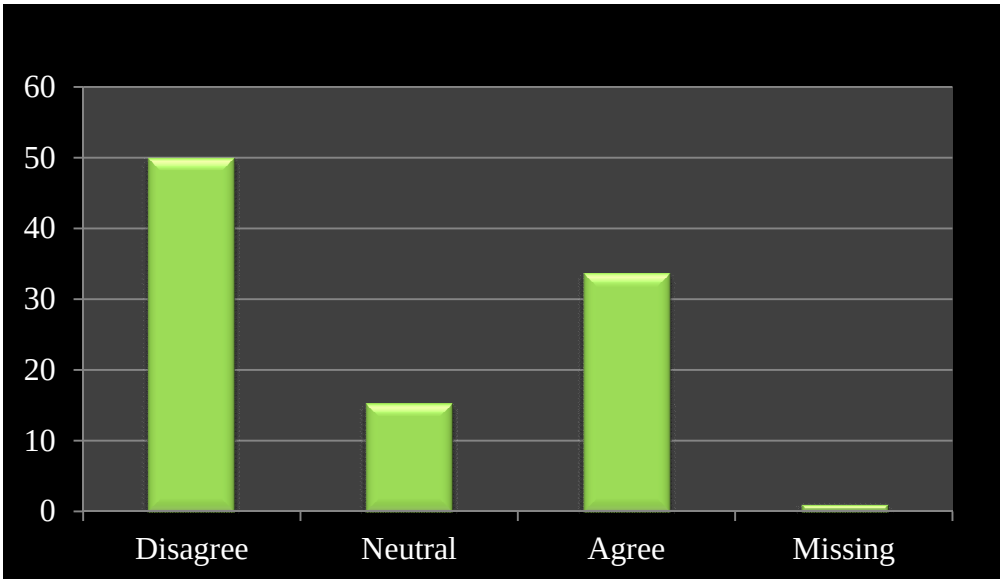
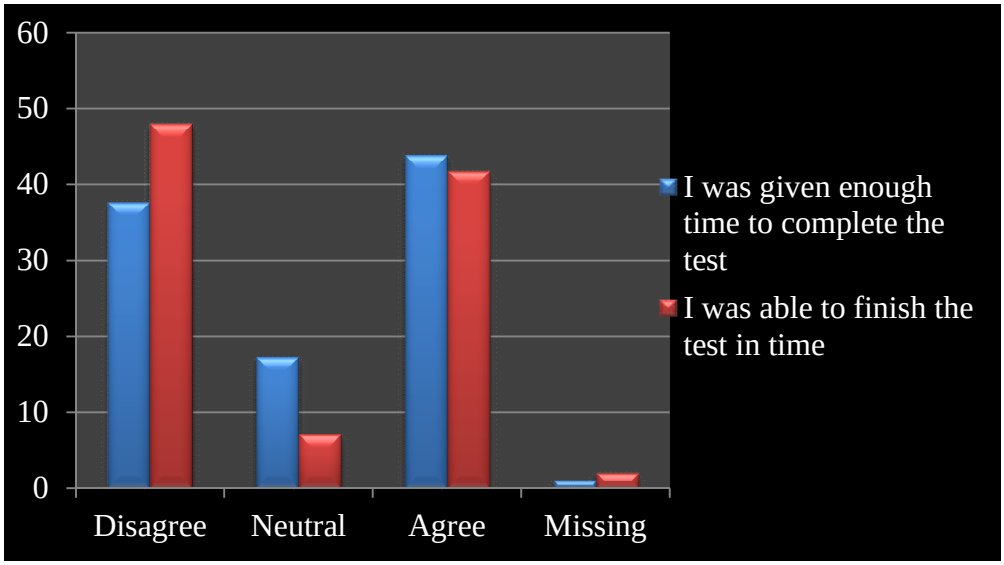


Figure 6.13 The time given to complete the test

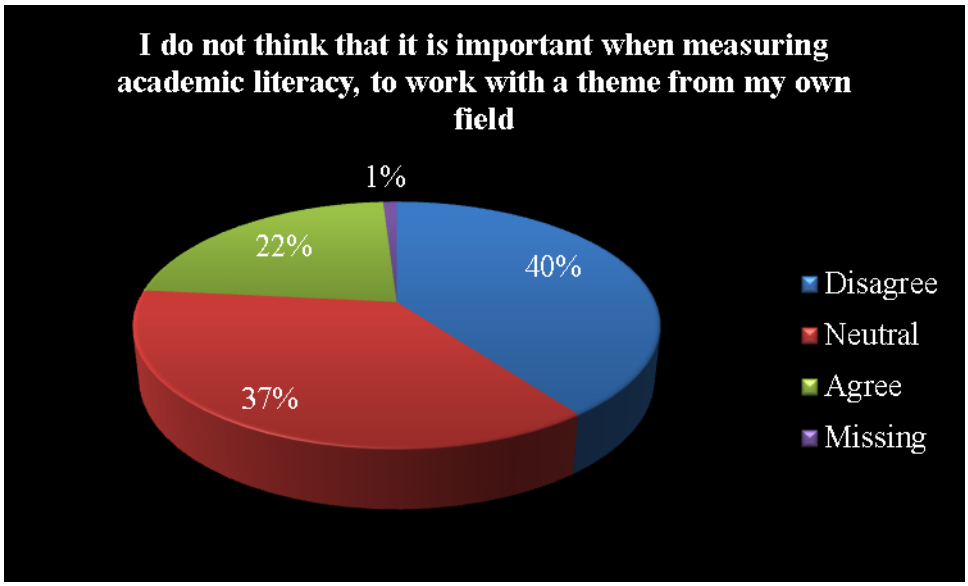


In terms of the time that students are given to complete the test, only 44% of students agreed that the time was sufficient while only 42% of students were able to finish the test in time. The main reason for students not being able to finish the test in time could be a result of students not reading fast enough. It is possible that students are not aware of their reading speed. It is for this reason that there should be a link available on the TALPS website allowing students to test their reading speed. It would also be helpful to provide students with

exercises to help them improve their reading speed. Allowing students to complete the sample test online may also give students an indication of how long it will take them to complete the test. It may help dispel the misconception on the part of students that the test is too long or that the time given for them to complete the test was not enough.

One other concern related to the accessibility of the test was the issue of the theme used in the test. One would assume that the vast majority of students would prefer a theme from their own field of study. Instead student responses to this were divided. When presented with the statement *"I do not think that it is important when measuring academic literacy, to work with a theme from my own field"*, 40% of students disagreed with the statement, indicating that they would prefer a theme from their field of study, 36.7% of students remained neutral while 22% agreed with the statement (Figure 6.14).

Figure 6.14 The importance of using a theme for TALPS

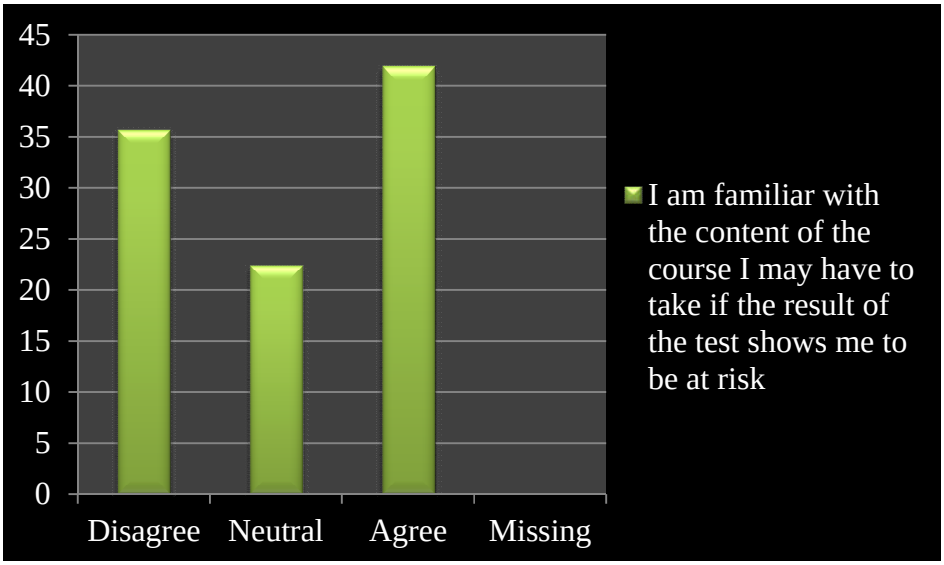


When students were asked "*How do you think you could have best prepared for this test?*" only 22% of students indicated that they need to know the topic/theme to adequately prepare for the test as can be seen from some of their responses:

- Reading, writing and more knowledge on global warming;
- If I was told to read more about global warming;
- If I was given the topic to research on as the topic of global warming is unfamiliar to me;
- Been informed upfront as the theme which seemed to me to be climate change and global warming.

In addition to questions about the accessibility of the test, students were asked about the intervention they may have to take. Surprisingly, 42% of students indicated that they were familiar with the contents of the course. Research pertaining to the information given out to these three groups of students indicates that while students were informed of the possible need to take the course, very little or no information about the content of the course was given out. It is possible that some students applying for admission to postgraduate study had, as part of their undergraduate degrees, taken the compulsory academic literacy course and were thus familiar with what a course like this would entail. Clearly, students do not know as much as they should about the intervention. Once again, information about the intervention should be available on the webpage. The brochure provides some information about the intervention but also directs students to the website as well as to a lecturer within the Unit for Academic Literacy who may be able to assist by answering questions students may have.

Figure 6.15 Students' familiarity with the content of the intervention



6.3.3 Discussion and conclusions

The analysis above can be summarised into a number of findings:

1. These students in general do not have a negative attitude to tests, nor do they believe that as test takers they have little or no rights. Less than half the students believe that tests are sometimes unfair.
2. While they are familiar with the concept of academic literacy, they do not have a clear enough understanding of the abilities it encompasses.
3. The majority of students acknowledge that literary skills affect academic performance, that a test of this nature is necessary and that they would be willing to take an intervention programme to help them improve should they need it. Less than half the students would feel disappointed if the test showed them to be at risk.
4. Students, in general, were aware of the purpose of the test but were not adequately prepared for the test. They indicated the need to have an idea of the 'scope' of the test as a means of preparation.
5. While 62% of the students understood the instructions, half of the students did not understand the questions. Students also indicated that the time given to complete the test was not enough and that they did not finish the test in time.

6. Students' opinion of the importance of using a theme from their field of study was divided – while close to 40% agreed that a theme from their field of study is important, 37% remained neutral and only 22% think that the theme used is unimportant.
7. Not enough students understand what is meant by the score they will receive for the test.

The data gathered from these students has given test developers valuable insight into the feelings and opinions of test takers. The questionnaire administered to this group of test takers has attempted to give a voice to a group often ignored in the entire testing process. The data collected has also presented a number of "pleasant surprises" and may also dispel some of the ideas we have about students' feelings towards testing, academic literacy and support courses.

There are, however, four concerns for the test developers of TALPS:

1. Students need a clearer understanding of the abilities that encompass academic literacy.
2. Students indicated that they were not prepared for the test.
3. Students felt that the time given to complete the test was not enough, nor did the majority of students finish the test in time.
4. Not enough students understand what is meant by the score they will receive for the test.

One way of addressing these concerns is by making available information to students. The web page and the brochure include a definition for the term "Academic literacy", outlining what it is that students should be able to do at university level. Seeing a sample test will provide students with the 'scope' they require. It will also make them aware of how long the test is and how much time they are given to complete the test. They may also have the option of testing their reading speed as well as information on ways to increase it. The

proposed brochure and the web page include information about the interpretation of the results of the test.

6.4 Conclusion

Overall, however, one can deduce that TALPS is, to a large extent, accessible to test takers. The strategies proposed here and in chapter 5 will ensure further accessibility and transparency, especially in terms of making information available about the test to test takers and others interested in the use of the test. It must be noted, however, that while ways of improving the accessibility of the test have been considered, TALPS is still a fairly new test, having been piloted in 2007. As the test is used, more research will be conducted on the test, and further strategies to improve the accessibility of the test will become apparent.

Chapter 7

Accountability

7.1 Introduction

The theory that underlies the framework of this study is based on the idea that applied linguistics is a discipline that provides solutions to language problems (Weideman, 2006a: 72). These designed ‘solutions’ (Weideman, 2006a: 72) take the form of language courses or language tests, and, as explained in chapter 3, such a course or test is in turn justified by a theoretical analysis (Weideman, 2006a: 72). Bygate (2005: 4) articulates this further when he states that the discipline of applied linguistics should go beyond just a “theoretical and descriptive study of the role of language in real-world problems, into the planning, design, and evaluation of potential responses” (2005: 571), and that the intention should be to address, and not merely describe, real-world problems (2005: 571). Those working in the field of applied linguistics, then, are not passive observers but active participants. If we are to accept the role of designers that provide solutions to language problems, then there needs to be interaction between us and those who rely on us to provide these solutions. Clearly, these solutions cannot be designed or developed in isolation, and the framework employed in this study is based on this premise.

This framework does not so much place emphasis on one central concept in language testing, but instead presents the reader or designer with a number of fundamental concepts in language testing (see chapter 3). It also does not separate the empirical analyses of a test from its social effects, nor does it see the constitutive concepts of validity and reliability as separate or superior to the regulative concepts of articulation, implementation, utility, alignment, transparency, accountability and fairness/care.

What is the role of applied linguists working within this framework? How do we apply these concepts in our designs? Should we be active participants or passive observers, hiding behind the 'scientific' (Weideman, 2006a: 80) justifications for our designs? Or are we, like members of other professions, responsible for the designs we create? If we are responsible for our work and to the people affected by it, how do we ensure that we undertake this responsibility with integrity, ethicality and professionalism? These are some of the questions this chapter will attempt to answer. In order to do this, this chapter will focus specifically on the aspect of the accountability of the test developer. This chapter will attempt to do the following:

- Provide a definition for the terms 'accountability', 'dual accountability', 'public accountability' and 'academic accountability' with particular reference to their use in language testing;
- Consider the all-important question of the limits of the responsibility of language testers;
- Investigate ways in which the test designers of TALPS can ensure that they are accountable or doubly accountable for their designs. As will become apparent below, such a discussion would be incomplete without a wider consideration of the teaching/intervention that follows the test.

7.2 Defining accountability

The term 'accountability', like the term 'transparency', features prominently in the literature of many disciplines: commerce, law, education, public management and human resources (see Norton, 1997; Beu & Buckley, 2004), to name just a few. Explained simply, accountability has to do with taking responsibility for your actions. Accountability, however, does not stop there but requires, in addition to accounting for one's actions, that one be willing to

face the consequences of these actions. According to Sinclair (1995: 220), accountability entails a relationship in which people are required to explain and take responsibility for their actions. Bovens (2005: 7) explains that accountability should be defined as a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct to the forum, which then becomes a platform that can pose questions and pass judgment, and even sanction the actor.

The same kind of relationship is echoed in other authorities. According to Frink and Klimoski (2004: 2), for example, definitions of accountability tend to

> revolve around two specific themes. One theme concerns the context, that is, who and what is involved in a given situation, and the second theme involves the notion of an evaluation and feedback activity in some form (Frink & Klimonski, 2004: 3).

Explained simply: "Accountability involves an actor or agent in a social context who potentially is subject to observation and evaluation by some audience(s), including oneself" (2004: 3). There are also

> standards, or expectations against which the agent's behaviour are compared, and the belief on the part of the agent of some likelihood that he or she may need to answer for, justify, or defend the decisions or behaviours. In addition, it is important that there are outcomes for the agent (i.e., sanctions, rewards, or punishments that can be explicit or implicit, and also objective or subjective) (Frink & Klimonski, 2004: 4).

In explaining his use of the term accountability, Weideman (2006a: 72) turns to the definition provided by Schuurman (2005). Schuurman's definition, too, stresses the need for one to be aware of his/her actions and to "give account of the same to the public" (Schuurman 2005: 42).

The next section turns, therefore, to the discussion of the following question: How does the concept of accountability, as defined above, relate to the field of language testing?

7.3 Understanding accountability in language testing

While a concern for issues related to accountability has not been completely ignored, the emphasis in the field of language testing seems to have revolved around two aspects of accountability: the need to 'professionalise' the field and the need for codes (ethics and practice). Davies makes reference to the creation of an "ethical milieu" (1997: 336) for language testers, the need for a "professional community" (1997: 336) which stipulates particular codes for behaviour and practices. An example of this is the Hippocratic Oath, which is a code of ethics adopted by members of the medical profession. Codes of ethics stipulate principles or standards or responsibilities that members of that profession should operate within. In general, they encourage ethical practice, providing guidelines to do this.

Codes of ethics in 'weak' (Boyd & Davies, 2002: 307) professions like language testing, however, are problematic. Members of these weak professions cannot be forced to abide by or uphold these codes or be forced to become part of the professional body that determines these codes. In these 'weak' professions, in other words, the sanctions that can be applied are themselves lacking in force. Why then, asks McNamara and Roever (2006: 139), should these 'weak' professions adopt a code of ethics if it cannot be enforced? The answer, they state, could lie in the fact that a code of ethics "gives members a moral guideline for action and helps them to resolve ethical conflicts" (2006: 139). In the case of the field of language testing, the concern with the need for a code of ethics or practice was highlighted by Kunnan (2006) who, at a meeting of the Association of Language Testers in Europe (ALTE) in 2006, outlined ten sets of standards and codes already in place.

What is immediately noticeable from the documents available to language testers (*ALTE*, 1994; *ALTE*, 2001; *ILTA*, 2000; *ILTA*, 2005) is that in some documents there is a distinction between codes of ethics and codes of practice,

while there are other documents like the Code of Fair Testing Practices in Education, (Joint Committee on Fair Testing Practices, 2004) that are, as pointed out by Shohamy (2001: 145), a combination of ethical codes and a code of practice.

A code of ethics is mainly concerned with issues of moral and proper conduct. But how does one determine what is really moral or proper? Different belief systems mean that we interpret different situations differently. What is moral for one may be immoral or unacceptable to another. Some religions are based on the ideal of peace and harmony, in which case taking another life would be considered immoral. On the other hand, the tenets of some other religions could be based on the ideal that taking a life in a war over religion is proof of your commitment to that religion. In that case, the taking of a life is considered just and proper. If, as Boyd and Davies maintain (2002: 307), "weak professions" cannot be forced to abide by these codes of ethics, how do we ensure that these "weak professions" at least operate within an acceptable standard of practice? Boyd and Davies point out that a code of ethics is not a statue or a regulation and it does not provide guidelines for practice – what it does is offer a "benchmark of satisfactory ethical behaviour" for those in the profession (2002: 306). Boyd and Davies (2002: 302) point out, also, that these codes of ethics are necessary but not sufficient (2002: 301). In their words, and in keeping with the theory that underlies this study, "codes of ethics are regulative ideals" (Boyd & Davies, 2002: 301).

The ideal may be to have, in addition to a code of ethics, a code of practice that will "identify the minimum requirements for practice in the profession and focuses on the clarification of professional misconduct and unprofessional conduct" (Boyd & Davies, 2002: 306). A code of practice may have more weight in ensuring acceptable standards than a code of ethics. McNamara and Roever point out that while such codes are important to the profession, these codes will never have the same 'bite' (2006: 147) as codes in other professions

(regulated by professional bodies whose membership is compulsory if one were to practice as a professional), that the acceptance of these codes rely on the individual member's conscience, which in turn is influenced by their moral and ethical standing (McNamara & Roever, 2006: 147). They state that while these codes provide a moral framework within which members should work, such a framework is "vague and has no built-in mechanisms to deal with violation" (2006: 147). In a similar fashion, Davies (1997: 336) points out that while the ILTA Code of Practice is a thorough and detailed document, it is lacking in two ways: it contains no reference to issues of ethics or morality and central place is given to validity, reliability and practicality (Davies, 1997: 336), and that possibly this is what ethics and morality means in language testing (Davies, 1997: 336). The same could be said for the *Standards for educational and psychological testing* (*AERA*, 1999) (commonly referred to as *Standards*), which is the one other document of importance to language testers in this regard. The document is divided into three parts: Test Construction, Evaluation and Documentation, Fairness in Testing, and Testing Applications (*AERA*, 1999: iii), but dedicates only thirty-five pages to the discussion of fairness in testing and test use and all of seven pages (pages 111–118) to the responsibilities of test users.

So, while codes (of ethics and practice) have been put in place to help regulate the profession and those associated with it, codes are just not enough. They may help satisfy the need for accountability to the profession, but might make no real contribution to public accountability. This helps explain Weideman's statement that the link between 'standards' for language testing and public accountability is not always "as productive as is conventionally thought" (Weideman, 2007b: 43). Because language testing is so closely linked to social issues, it is imperative that test developers also become publicly accountable for their designs. Often, however, language testers work in isolation. It is quite possible that there is little or no contact with the people who are most affected by their designs. Working in isolation or relative isolation means that it is much

easier not to be held accountable for your actions or designs. For real progress to be made in the field of language testing, language testers cannot, and should not, ignore the voices of the lay communities they may serve. In addition to this they need the input, advice and opinion not only of their peers, but also of those affected by the implementation of their designs.

In the field of language testing, Shohamy (1997; 2001) and others have stressed the need for dialogue between all those affected by the testing process. Professionals like lawyers, doctors, social workers, language testers and the like cannot function in isolation. They are accountable to the profession they belong to and to the people most affected by their practices. Bygate refers to this as being "doubly accountable" (2004: 19). He explains that applied linguists need to be accountable to the discipline within which they work, and to the communities that they serve. He also makes mention of the relationship between the "scholarly apparatus of the academy and the social reality which is under scrutiny" (2004: 7), pointing once again to the fact that those working within a particular profession or discipline cannot function effectively without consideration for the very people they claim to serve.

Weideman (2007b: 43) is in agreement with Bygate's contention that applied linguists need to be doubly accountable. He explains that our applied linguistic designs (Weideman, 2007b: 43) must be accessible not just to experts but to users and the general public, that we cannot only defend our designs by "reference to other expert opinion" (2007b: 43). He states that "the technical defensibility of a design which links the technical and the juridical" (2007b: 43) does not only depend on its theoretical defensibility; that in addition to being able to defend the theory on which the design is based, we need to publicly defend the design. The design should be accessible and the defence understandable to the expert, the user and the lay public. It is therefore not enough that the test is based on a theoretically sound construct (that of academic literacy). This "theoretical accountability" or "theoretical

justification" is only one part of the picture. Equally important is that this information be available, but more importantly, understandable to those affected by the use of the test results.

Clearly the concept of accountability, as used by Weideman (2003b; 2006a; 2007b; 2009a), moves beyond a concern with the need to account for something or account to someone. Within the framework on which this study is based the regulative condition of accountability relates the technical, qualifying dimension of the designed test to its juridical and ethical aspects (Weideman, 2009a: 249). Juridical, as used by Weideman (2009a), is not used in the traditional sense of simply referring to retribution or punishment, but rather in the positive sense of a concern with doing that which is right and fair – importantly that one's actions can be justified, defended or explained (fairly and openly) to those involved, interested in or affected by these actions. The ethical analogies referred to here generate a sense of care and concern for others. Accountability, then, according to Weideman (2006a; 2007b; 2009a) focuses on the element of responsibility without neglecting the need for fairness, care and concern for those who are affected by the use of the test results. What this means is that, in addition to ensuring that we design tests that are valid and reliable and based on theoretically sound constructs, our concerns should extend to the effects of our tests on test takers and others affected by the use of the test results. This lands the debate squarely within the realm of so called consequential validity or test impact (see next section).

Weideman has often stressed the need to work with integrity (2007b: 30), that we should demonstrate through our designs the love we have for others (2007b: 44), and that our designed solutions to language problems should relieve some of the suffering, pain, poverty and injustice in our world (2007b: 44). In proposing a responsible agenda for applied linguistics, Weideman states that applied linguists should "seek to become accountable by designing theoretically and socially defensible solutions to language problems"

(Weideman, 2007b: 29). The need to work with ethical considerations in mind and therefore with integrity is also highlighted by Schuurman (2010), who has had considerable influence on the work of Weideman. Schuurman (2010) defines ethics as a "theoretical discipline that reflects on the good or responsible actions of human beings" (2010: 111). He makes reference to "an ethics of responsibility that integrates ethos, intention, values, and norms in a coherent way" (2010: 122). In proposing eight principles to be applied to ensure responsible design in technology, he stresses the need for the consideration of harmony, justice, care, love and trust. It is, however, important to point out here that while it is essential for test developers to work with integrity and responsibility, it is equally important to be realistic about our limitations or the limitations of the field in which we work.

7.3.1 The limits of accountability

Concerns about the limits of the accountability of the language tester are well documented in the literature on language testing. Hamp-Lyons, in her discussion of the acceptance of responsibility on the part of the test developer, makes reference to the "acceptance of responsibility, within awareness that responsibility is complex, that all situations are multi-faceted, and that absolutes are dangerous" (2000b: 589). If "absolutes are dangerous", should test developers accept absolute responsibility for their designs or are there limits to their responsibility? Should one agree with Hamp-Lyons who states that "we must accept responsibility for all those consequences *which we are aware of*" (1997: 323) and with Davies (1997: 335) who states that it is not possible for the tester to take account of *all* the possible social consequences? Davies answers these questions when he states that an "ethical perspective for a language tester is necessary" (1997: 335) but that there is a limit to what is achievable and desirable, that the open-ended offer of consequential validity goes too far, and that it is not possible for a tester to take account of all possible social consequences (1997: 335).

Making reference to these concerns with issues of accountability, washback and impact as documented in the literature on language testing, Fulcher (1999: 3) asks if these concerns are a "new awareness that has suddenly blossomed in the last few years". His answer is that they are not a new concern, that "while these new studies are to be welcomed, they are generated by a sense of panic, or moral directionlessness"; that when "considering ethical issues as applied to practical testing issues, there is the sense that there is something fishy going on, but no real basis for investigating why something may be wrong". His advice to language testers is to go back to the "theoretical framework of Messick for guidance, rather than flailing about in the postmodern sea of uncertainty" (1999: 3). His reasoning is that the Messick framework incorporates ethics into the concept of construct validity (1999: 3), in the notion of consequential validity.

In this debate one must agree with Hamp-Lyons, who in response to Fulcher states that accountability, washback and impact studies and concerns with ethics and responsibilities "is not evidence of a sense of panic, or moral directionlessness" (2000b: 589) and:

> that the role to be played for discussion in professional meetings on questions of responsibility, ethics, morality, the demands of professionalism, the appropriateness of specific practices and decisions within actual social, cultural and political contexts, the courses of 'right conduct' open to the tester, cannot be over-emphasised (Hamp-Lyons, 2000b: 589–590).

Fulcher's suggestion that all answers can be found in the "elegant Messick framework" (1999: 3) is questionable as well. The shortcomings of Messick's framework have been discussed in detail in chapter 3. To recapitulate very briefly here: going back to Messick's framework would mean the accentuating of the concept of (construct) validity at the expense of other fundamental concepts like ethics, responsibility and accountability. Weideman (2009a) observes that when a test unintentionally disadvantages others, it is a breach of its ethical concern with the rights and interests of others rather than a threat to

the test's validity. It is only if one privileges the concept of validity at the expense of other fundamental concepts in language testing that one will see this as a threat to validity (Weideman, 2009a: 249). Popham (1997) is of a similar view. He states that "lumping our attention to the social consequences of test use with the concept of validity will not only muddy the validity waters for most educators, it may actually lead to less attention to the intended and unintended consequences of test use" (Popham, 1997: 13).

The framework employed in this study is beginning to be applied in understanding the relations between transparency and accountability in language test design (Weideman, 2009b: 66). Working within this framework means that test designers are conceptually prompted to consider issues of accountability. Importantly, the framework helps limit the test designer to what he or she is responsible for, because realistically one cannot claim responsibility for effects that are beyond our control or foresight. Hamp-Lyons's argument is a simple one, that while test developers are responsible for the designs created, there are limits to their responsibility and that "ultimately, each person will make a personal choice based on their knowledge, experience, values, constraints, priorities" (2000b: 590).

That said, however, the focus here is on the accountability of the language tester and how this can be put into practice. How do test designers, in becoming accountable for their designs, ensure that they work with integrity and that their tests do good and have positive effects?

A good starting point would be to look at the way the concept of accountability relates to these issues. As has been pointed out, in the field of language testing, at first glance, accountability has two dimensions: theoretical accountability or accountability to the profession, and public accountability. There is, however, one other kind of accountability that needs to be considered. It is the aspect of

the academic accountability that the language tester must consider that will also be discussed in this chapter.

7.4 Theoretical accountability

Theoretical accountability, as defined by Weideman (2007b: 43), refers to one being able to defend the theory on which the design is based. A test designer cannot claim to be truly accountable if theoretical accountability has not been considered. Theoretical accountability is synonymous with construct validity. Identifying construct validity as one type of accountability is an example of how each mode of experience within the framework employed here coheres or is related to every other mode or aspect (Weideman, 2009b: 66; Weideman, 2011: 7). Construct validity (discussed in detail in chapter 3) or theoretical accountability is a necessary or constitutive condition that lies in the foundational direction (see chapter 3). According to Weideman, the various notions of test validity or the effect or force of a test relates the technical instrument to the physical side of reality (2007b; 2009a). Within this conceptual framework, accountability is a regulative condition or "disclosing reference" (Weideman, 2009b: 66) that lies in the opposite direction and ties the technical design to its juridical and ethical analogies. According to Weideman (2009b: 66), the term "disclosing" refers to the opening up of one leading or qualifying aspect by its anticipation of dimensions that occur later in the sequential order of aspects: the constitutive condition of validity is therefore disclosed or opened up by the regulative condition of accountability.

Theoretical accountability is often the one type of accountability least neglected. Experts in many fields have, almost always, felt the need to be accountable to their peers – by publishing their research in accredited journals and/or by presenting their research at national and international conferences. This has, however, not always been the case with the field of language testing. Some test designers have worked and continue to work in isolation. Often the

attitude seems to be that a test ‘belongs’ to particular individuals or organisations and as such there is no real need to be accountable in the true sense of the word. Or tests are designed with the best of intentions, and locally implemented with no recourse for design conventions and principles. Other experts in the field and the general public are expected to “buy into” the value of the test without any real accounting on the part of the test developer.

Clearly, this is not accountability at all. A first step in accountability should be to one’s peers and other experts. In the case of the design of TALL, research regarding the construct, blueprint, piloting and refinement of the test was presented by the designers to other experts at conferences and in published research papers – presenting a forum for other experts to comment, question and provide valuable input or critique. With TALL this sharing of information led to other institutions choosing to become partners in the design and use of the test. The same has been done with TALPS, and should be done with other large scale tests when they are being developed. For an overview of such discussion and scholarly debate, the ‘Research’ tab on the website of the Inter-institutional Centre for Language Development and Assessment (ICELDA) directs one to more than two dozen studies on these tests (http://icelda.sun.ac.za).

One cannot deny that a first step in becoming accountable requires being accountable to those working within the profession. Theoretical accountability is crucial in the design process. It must, however, be followed closely by an accountability to the public who are affected by or interested in the use of the test.

7.5 Accountability to the public

The need for public accountability has been alluded to by many in the field. Boyd and Davies (2002) call for the profession of language testing to have

high standards, with members who are conscious of their responsibilities and open to the public (2002: 312), that it is not too late for language testers to "build in openness to its professional life" (2002: 312). Earlier in the same article, in a discussion of the relationship between doctor and patient, Boyd and Davies make reference to the "open dialogue with patients and the public" (2002: 300), pointing out that this would be the way to go in the medical field. The same should apply to other fields that require interaction between the profession and the individual and society. Rea-Dickins's (1997: 304) call is for there to be relationships between all stakeholders (learners, teachers, parents, testers and authorities) in the field of language testing. She states that "a stakeholder approach to assessment has the effect of democratising assessment processes, of improving relationships between those involved, and promoting greater fairness" (1997: 304).

As can be seen from the discussion above, the accountability of the language tester must extend to the public being served. Defining public accountability, however, is a fairly easy task, ensuring accountability to the public less so. Public accountability means exactly that – to be open to the public one serves, and thus allowing the "open dialogue" referred to above. It is not enough that test designers defend their designs to the experts or their peers in the field. Equally, if not more so, those affected by the use of the test scores must be well informed as well. This is where Bovens's point becomes important – that one must be aware of the kind of information that is made available. It is not enough that the information is made available. The information must be understandable to the very people who need to understand it most and not a "…monologue without engagement. To qualify as public accountability, there should be public accessibility of the account giving" (Bovens, 2005: 10).

In the case of TALPS, the website and the pamphlets distributed to interested students will go a long way in ensuring that the public is provided with information regarding the test. Importantly, the test designers have ensured that

the language used in both these mediums is understandable to the lay person. The point here is that care must be taken with the way information is dispensed to the public. What is available for the experts in the field may not be accessible to the lay person taking the test. The challenge, to a certain extent, is to translate technical concepts into more readily accessible, non-specialist language, while at the same time relating their theoretical meaning to real or perceived social concerns. All the while, it is incumbent on the test designer to be mindful of the limitations inherent in theoretical explanations, and in the technical measuring instrument (the test) that is being employed. For example, public explanations of how test results can be used must be quite open about the fact that tests cannot predict everything. Language tests lose their predictive value of future performance, for instance, with every subsequent year of a student's study at university.

Within the framework employed in this study, issues of public accountability/defensibility relate the technical, qualifying dimension of the test to its juridical aspects (Weideman, 2009a: 249) – the term juridical itself pointing to the fact that tests, like the law, for example, are not experienced in isolation but in conjunction with those affected by it. The presence of a jury in a courtroom is evidence of this. The sentencing of a criminal is not a private affair between lawyers and the judge. Every step of the process from the arrest and incarceration of the criminal to sentencing is documented, is transparent and open to scrutiny from the public. Everyone affected by the crime is aware of their rights and the process to be followed to exercise these rights. Information about the crime, the trial and the sentencing is accessible to the public. Lawyers and judges are bound by the codes of practice as stipulated by the professional associations they belong to. Should they act irresponsibly and be found guilty of unethical behaviour, the consequences are severe – such as having their licence to practice revoked and suffering private and public humiliation. Lawyers, like testers, provide a service to humanity and as a consequence of this should be concerned with the way their work affects

others. The law, like testing, can have dire consequences – testers, like lawyers, need to be accountable for their designs and should be willing to publicly defend their designs.

7.6 Academic accountability

Academic accountability, presented here, seems to imply that it is a new and separate type of accountability that needs to be considered. This is not so. Strictly speaking, academic accountability may very well be a subset of public accountability, both of which (public and academic accountability) can be classified as being a part of social accountability. The concepts of public and academic accountability are separated here, despite the fine line between them, to allow us to "separate from each what is conceptually distinct" (Weideman, 2009a: 249). The specific purpose here is to emphasise or highlight every aspect of accountability. What academic accountability has in common with certain other kinds of accountability is that it is an institutional kind of accountability. Why it is particularly relevant here, as will become clear below, is that it relates strongly to the context – in this case an institutional context – in which the test under discussion is being employed.

7.6.1 Defining academic accountability

The main focus of academic accountability, according to Dill (1999: 127), is to ensure that universities maintain or improve the quality of their teaching and learning. He explains that universities should become "learning organisations" (1999: 127) where the focus should be on "creating knowledge for the improvement of teaching and learning" (1999: 127). According to Kearns (1998: 140), academic accountability has to do with a "strong institutional commitment to quality teaching". He points out that this should provide students with the "prospect for gainful employment or other opportunities upon graduation" (1998: 140).

Academic accountability cannot be discussed or defined without a consideration of the concept of institutional accountability. Should we see academic accountability as referring only to the accountability of the professionals or academics within the institution we could then talk about the academic accountability of the language teacher or the mathematics teacher and so on. This would suggest that there is a divide between the institution and the academics working within it. In this scenario institutional accountability would focus on the organisational, administrative or structural components of the institution. Should we agree to adopt Dill's definition of academic accountability, then the divide between academic accountability and institutional accountability disappears. Dill (1999: 127) makes reference to the fact that universities, not institutions or academics, need to consider their quality of teaching and learning. Realistically, and practically so, it would be difficult and problematic to separate academic accountability from institutional accountability. Dill also makes reference to the fact that universities should become "learning organisations where learning is maximised" (1999: 129). In order to achieve this goal the academy and the institution should work towards a common goal. Using this justification, this study will therefore consider academic accountability as synonymous with institutional accountability. Academic accountability, as used here, refers, then, to the accountability of the language tester in respect of the teaching and learning that follows, or should follow, a test, with the specific aim of ensuring that this teaching and learning has some positive outcome.

Within academic accountability one needs to consider the 'public' versus 'private' aspect of accountability: 'public' referring to those outside of the institution, an aspect discussed earlier in this chapter, and 'private' referring to accountability within the institution. In terms of accounting to those within the institution there are two groups one needs to be concerned with: the one refers to the faculties, stakeholders and management of the institution. As mentioned in the chapter on transparency, it is important that any information regarding a

measuring instrument, in this case the TALPS, be shared with other stakeholders. An effective method would be through seminars, presentations and workshops where information as well as research conducted about the test is shared. Another would be the standard set of routine meetings within the institution where such matters might expectably form part of the agenda. The second group that needs to be considered are those who have the most at stake – our students who take the test. Chapters 2 and 3 have discussed in detail the strength of the test. Does our responsibility end here? If it does, then what have we achieved, except perhaps to have made supervisors and students aware of the fact that the academic literacy levels of their students place them at risk? How has testing these students really contributed to the care and concern for others that Weideman (2009a: 235) makes reference to? Is it acceptable to be satisfied that we have designed and administered a socially acceptable test, yet have done nothing to assist those students who are shown to be at risk? Can the test be considered socially acceptable if this is the case? Have we at all prepared them for the responsible experience and outcomes that Kearns (1998: 140) makes reference to? The reality is that testing the academic literacy of students but doing nothing to help them may be considered a futile exercise. Issues of accountability dictate that if we test students, we should do something to help them improve. The responsibility of ethical language testers extends into the teaching that follows.

This part of the study is aimed at determining the effects that the test may have had, if any, on the intervention and the teaching that follows the test. In terms of the accountability of the test developers, which is the focus of this chapter, the intervention programme which follows the test must be considered. The educational context in which TALPS takes place has already been discussed in chapter 2 where the need for TALPS was considered. In the case of TALPS, the intervention or the course came first and had been in operation for a while before the test was designed and implemented. The test came about as a result of the course – the course is not an effect of the test. Despite this, the

intervention provided to students who are shown to be at risk by the results of the test is still an important one here. Research has shown that testing (whether it comes before or after the intervention/course) causes people to behave differently or to do things differently (see Smith, 1991).

7.6.2 The Postgraduate Academic Writing Module (EOT 300)

The intervention that is relevant in this specific instance is the Postgraduate Academic Writing Module (EOT 300), which was developed by the UAL because of the need to assist postgraduate students with their academic writing problems. Butler's (2007) study highlights this and the fact that, in addition to the course, there was a need for a reliable testing instrument. The test and the course work hand-in-hand. The test is used to determine the academic literacy levels of postgraduate students. Students who are shown to be at risk may be expected by their faculties at the University of Pretoria to take the EOT 300 module. Having students take the test before the course means that students who are not at risk do not have to sit through a module they may not need. Already there are positive effects – without the test students may not be aware of their academic literacy levels. In addition to an awareness of their abilities, students who are required to take the course are provided with an intervention that may help them succeed in their studies. Poor academic writing skills are bound to hamper their studies, and an intervention designed to help develop these skills may mean the difference between success and failure.

Writing, especially in the academic context, however, cannot function in isolation and is dependent on other abilities the student should acquire. A student who is a poor reader, for example, cannot be a good writer. Good writing depends on a student being able to read critically, to be able to summarise effectively what was read, and to use what she or he has learned in the reading/research process to construct a logical, well argued stretch of academic writing. In addition to this, it is essential that students' writing be free

of spelling and vocabulary/grammar errors, that they know how to use a dictionary to avoid these very errors, that they are aware of the conventions of academic writing and that this be evident in their work. As a result, the writing process must be taught in conjunction with these other abilities that students need. Based on this, the designers of EOT 300 point out that the aim of the course is the "further development and transfer of academic literacy" and that the "skills acquired and developed during this course should be applied to the wider context of their studies" (Butler, Pretorius & Van Dyk, 2009: viii).

Butler's study (2007) is focused on a framework that should be employed when designing a writing course for tertiary level students. This section will concentrate on the design and implementation of the course, with a focus on determining what effects the course and the test have and whether there is alignment between the test and the course.

7.6.3 The design of a postgraduate academic writing course

Butler identified thirteen "requirements or conditions" (2007: 42) that function as principles for writing course design. These are:

1. Include an accurate determination of students' current levels of academic literacy;
2. Include an accurate account of the understandings and requirements of lecturers/supervisors in specific departments or faculties regarding academic writing;
3. Engage students' prior knowledge and abilities in different literacies to connect with academic literacy in a positive way;
4. Consider learners' needs (and wants) as a central issue in academic writing;
5. Create a learning environment where students feel safe to explore and find their own voices in the academic context;
6. Give careful consideration to the most important mode for teaching and learning academic writing;
7. Determine whether primary and additional language users should be treated differently in writing interventions;
8. Provide ample opportunity to develop revision and editing skills;
9. Acknowledge assessment and feedback as central to course design;

10. Provide relevant, contextualised opportunities for engaging in academic writing tasks that students feel contribute towards their development as academic writers in the tertiary context;
11. Include productive strategies that achieve a focus on language form;
12. Support and encourage the use of technology in writing;
13. Focus on the interrelationship between different language abilities in the promotion of writing (Butler, 2007: 42–55).

The conditions above do not function in isolation but are a combination of factors affecting the course designer, the students and the supervisors in different faculties and departments. The first requirement, according to Butler (2007), is to determine the academic literacy levels of students. This is where TALPS features.

In addition to the test that he used at the time (TALL, which did not have a writing section), Butler (2007) suggests that to determine the writing abilities of students they should be required to write an essay. He states that while this may not be as reliable as the empirical analyses from a test like TALPS, it entails a "more credible and appealing" (2007: 43) method. It is an excellent idea to combine both assessment types. Often students may take a test but not see or understand how this is related to the abilities they are expected to have, i.e. they may not see the correlation between the different sections in TALPS and how these are related to their academic literacy levels, especially their writing skills. These essays can be evaluated individually, in groups and with the lecturer and the supervisor concerned. This first writing exercise can generate discussion between the lecturer and students, and ties in directly with the need to create a learning environment where students are comfortable enough to voice their fears, struggles and concerns about their academic literacy, specifically academic writing. It also helps open up a dialogue about students' needs – if the lecturer knows what students need, it will be easier to help them.

The teaching and learning of academic writing is of course not limited to the classroom. The lecturer and the course designers accept that the students sitting

in their lecture room will eventually be writing for someone else, in a different department or faculty. Butler stresses, furthermore, the need to recognise the match between the texts that students produce and what their lecturers expect from such texts (2007: 43). He points out that it is important to be aware of the different conventions in different disciplines and to make students aware of this. It goes without saying that this requires a dialogue between the course designer and the supervisors in the different departments/faculties.

Another important factor to consider in assisting others in the development of their academic literacy to is to determine the most effective method to teach. Butler highlights a concern relevant to most teachers – large classes. As he points out, first year academic literacy classes are usually quite full and individual attention is often not possible. In the case of the University of Pretoria, there is a tutor system in place for undergraduate students and students do have the opportunity to have their work evaluated by a senior student. With regard to postgraduate students, Butler points out that classes are sometimes smaller and individual attention is more likely possible. Smaller classes may be the ideal but cannot always be guaranteed. Lecturers will have to "find creative ways of dealing with this issue" (2007: 47). One way might be to use a variety of assessment methods – individual, peer and group assessments. In addition, the course should be designed to have a variety of assessments: tests as well as presentations, draft work and short and long written tasks, for example.

The question arises of whether primary and additional language users should be taught separately. Once again, though this may for some be the ideal, it may never be a reality. Lecturers will have to find ways to deal with this, and one way is to exploit heterogeneity. Butler's advice is to have quicker learners assist struggling students (2007: 49). In terms of the writing course, there is a need to develop the revision and editing skill of students – this can be done by teaching writing as a process and encouraging students to revise their work as

well as the work of their peers. Condition nine (9) above emphasises the need for assessment practices to be ‘transparent’ (Butler, 2007: 51) so that students are aware of the requirements of a task. Also, Butler points out the need for, and the importance of the correct kind of feedback to students. He says that there is a “strong need to balance positive and negative feedback to students” (2007: 52), that lecturers should maintain a careful balance, and not just criticise a piece for its “inadequacies” (2007: 52). Another consideration in the design of the academic writing course is the question of whether to teach using discipline-specific/subject specific material to teach. Butler points out that in general students have a negative attitude to such remedial courses – students need to see that the course is in some way related to their field of study. Material used should therefore be seen by students as “contributing purposefully to their studies” (2007: 54). Other considerations focus on productive ways of including/teaching grammar, using technology in writing and the interrelationship between writing and other language abilities, like reading (Butler, 2007: 42–55).

Have these requirements been incorporated in the design of the course? To answer this question we need to take a closer look at the course and the tasks students have to complete and determine whether these are aligned with the test. The Postgraduate Academic Writing course (EOT 300) is divided into two themes. Theme one presents students with “An Introduction to Academic Discourse” (Butler, Pretorius & Van Dyk, 2009). The focus here is to ensure that students recognise the characteristics of academic writing, apply academic reading strategies, take effective notes, learn to deal with vocabulary difficulties, make functional use of a dictionary and recognise important principles of academic writing (Butler, Pretorius & Van Dyk, 2009).

Below is a table for each theme indicating the tasks students have to complete:

Table 7.1 Theme 1: An introduction to academic discourse

TASK	TOPIC
Task 1	Mind maps
Task 2	Componential structuring
Task 3	Interviews (to determine lecturer expectations are regarding students' academic writing
Task 4	Style and register
Task 5	Scrambled text (general text structure)
Task 6	Text type
Task 7	Text type
Task 8	Scrambled text
Task 9	Facts and opinions
Task 10	Logical connectors
Task 11	Referencing/Bibliography
Task 12	Interpreting graphs and visual information
Task 13	Text editing

(Butler, Pretorius & Van Dyk, 2009)

Theme two focuses specifically on the writing process. Tasks in this part of the course are aligned with the steps in the writing process.

Table 7.2 Theme 2: The writing process applied

STEP	TASKS
Step 1: Identifying a research problem (+ pre-writing)	Students are given a topic by the lecturer. Tasks here focus on pre-writing activities where students are asked to write down everything they know about the topic/theme. They are asked to write down questions they have about the topic and these are discussed in groups.
Step 2: Gathering information (+ pre-writing)	Research skills Structuring a bibliography
Step 3: Synthesising and structuring information	In-text referencing Integration of information using mind maps Developing criteria for quality academic writing
Step 4: Writing the first draft	Tasks in Step 4, 5 and 6 focuses on the writing, revision and editing of students drafts and those of their peers using the checklists / revision tables provided by the lecturer.
Step 5: Revision (+ subsequent drafts following from revision)	
Step 6: Editing and writing the final draft	

(Butler, Pretorius & Van Dyk, 2009)

The table below highlights the alignment between the sub-tests in TALPS and the tasks students have to complete in EOT 300:

Table 7.3 Aligning TALPS and EOT 300

SUB-TESTS IN TALPS	WHAT EACH SUB-TEST TESTS	RELATION TO EOT 300
1. Scrambled text	Recognising different parts of a text, forming a cohesive whole.	***Task 5, 8***
2. Academic vocabulary	Testing students knowledge of words used in a specific context.	***Theme 1 and 2***
3. Graphic and visual literacy	Interpreting information from a graph, summarising the data, doing numerical computations.	***Task 12***
4. Text type	Identifying/classifying different genres/texts types.	***Task 4, 6, 7***
5. Comprehension	Reading, classifying and comparing, making inferences, recognising text relations, distinguishing between essential and non- essential information.	***Task 1, 2, 9***
6. Grammar and text relation	Sentence construction, word order, vocabulary, punctuation.	***Task 8, 10, 13***
7. Editing	Correction of errors in a text.	***Task 13***
8. Writing	Argumentative writing, structuring an argument, recognition of sources.	***Task 3, 11 and Theme 2***

Especially the tasks students are expected to complete, that were outlined above, demonstrate that there is alignment between the intervention and its outcomes, the tasks students have to complete, and the test. Important, also, is that the test and the course are based on the same definition of academic literacy (see chapter 2). The abilities that are tested by the test are the same ones the course is designed to develop. The EOT 300 course is obviously a well thought out, deliberately planned course, designed to assist postgraduate students in adequately developing their academic literacy and writing ability. It strives to develop, in as much detail as is possible in one year, the academic literacy abilities one would need to cope at postgraduate level.

The course follows the principle of continuous assessment. The final mark for the course is a combination of the mark for the tasks completed and the major assignment. The major assignment is to be discussed personally with the lecturer after it has been marked. The lecturer uses a list of correction symbols and students are to familiarise themselves with these to help facilitate discussion of their work. The study component outlines the focus of the course and points out that while the course will address all four language abilities: listening, reading, writing and speaking, the emphasis is on developing "effective listening, reading and writing in an integrated manner in a postgraduate academic environment" (Butler, Pretorius & Van Dyk, 2009: vi). The course has been designed to provide a number of different ways of learning – individual, small groups and one on one interaction with the lecturer, providing ample "opportunity not only to share their opinions and ideas, but also to evaluate one another's ideas" (Butler, Pretorius & Van Dyk, 2009: vii). The workbook and the course are designed to be interactive. There is constant communication and discussion between the lecturer and the students.

The workbook begins with the basic abilities students will need, and slowly works its way towards the real target problem – writing. In the course of this journey, students have adequate opportunity to identify and address problems they have with academic literacy, specifically writing. One of the main problems of support courses like this is having students acknowledge that there are problems. As has already been discussed in earlier chapters and in Butler's (2007) study, there is often conflict between what students and lecturers think. It is not unusual to find that students who have serious problems with their academic literacy believe otherwise. Their justification is that if they have reached university, or in this case postgraduate study, they must be academically literate. Having students not just acknowledge, but identify the problems they have is a first major step towards progress, allowing more effective help from the lecturer and the course.

The variety of assessment methods and different kinds of tasks provides students with ample opportunity to make progress with the course. At postgraduate level, the classes may be smaller, allowing effective interaction with the lecturer concerned. Students meet with the lecturer once their assignment has been assessed. Such a meeting is valuable in helping the student understand where problems lie and what can be done to resolve them. It is not enough to address these problems generally when teaching. Because the academic literacy levels of students depend on individual students' abilities in language, as well as their background, schooling, family life, race or region, these problems are best addressed individually. While this is not possible at undergraduate level, it is certainly desirable at postgraduate level. The focus here, however, is not to critique the course, but to determine if the test, which is written before the course but was developed as a result of the course, is aligned with the intervention.

Clearly, the test and the intervention have been shown to have positive effects. It must be pointed out here, as well, that in addition to EOT 300, there are other courses also at higher level than first year, that are designed to develop the academic literacy, specifically writing ability, of students. One such example is the "Essay writing course for students of History", designed and developed by Carstens (2009) as part of her doctoral study. The module was a semester course aimed at students with History as a major in their second year of study (Carstens, 2009). Like Butler (2007), Carstens (2009) found that students responded positively to the intervention and "indicated that their personal needs had been more than adequately addressed" (Carstens, 2009: 228).

Testing students makes them aware of their academic literacy levels, and providing them with an effective intervention designed to help them improve means that they may be able to graduate in the required time, that may not have been possible without the intervention. Weideman (2007b) sums this up effectively when he states that

> our designs are done because we demonstrate through them the love we have for others: it derives from the relation between the technical artefact that is our design and the ethical dimension of our life. In a country such as ours, the desperate language needs of both adults and children to achieve a functional literacy that will enable them to function in the economy and partake more fully of its fruits, stands out as possibly the biggest responsibility of applied linguists (Weideman, 2007b: 53).

7.7 Conclusion

This chapter began with a discussion of the concept of accountability, asking, also, the all important question of how test designers can ensure that they become accountable for their designs. The detailed discussion of the different types of accountability has attempted to answer this question. In a nutshell, however, test designers can do this by:

- designing fair tests that can be justified, explained and defended publicly;
- being transparent and opening up a dialogue between all those involved in the testing process;
- designing tests that do good and that have positive effects;
- being committed to the test takers we serve and by ensuring that our responsibility does not end with a score on a sheet, but is followed by effective teaching and learning that will eventually have potentially far-reaching, positive consequences for the society these test takers live and work in.

This chapter has considered the concept of accountability as it relates to the test designers and the process they followed in the development of TALPS. Clearly, accountability, as defined here, has many facets. Each of these, as explained, is a vital consideration in ensuring accountability. But does a concern for issues related to accountability mean that the responsibility of test

designers ends here? This study is focused on the concepts of transparency, accessibility and accountability. The burning question, then, is whether this is enough. If it is not, then, what is the way forward? This is the question the final chapter of this study will attempt to address.

Chapter 8
Regulative conditions for test design

8.1 Introduction

This study has attempted to articulate the application of an emerging theoretical framework for applied linguistics to the design, development and implementation of a postgraduate academic literacy test. The main focus has been on investigating whether the regulative conditions of transparency, accessibility and accountability could be incorporated in the design of one test and theoretically accounted for in terms of a framework.

As explained in chapter 3, the questions raised in this study could not be addressed by the existing frameworks of Kunnan (2000; 2004), Bachman and Palmer (1996) and Messick (1989b). While these frameworks address important concerns related to language testing, each proposes an 'overarching' concept such as usefulness or validity or fairness to unify the field of language testing. Rather than these overarching concepts, what the field requires is that test developers see the relationship between fundamental concepts in language testing. If there is any unifying concept in language testing, it would probably derive from the technical design features that characterise the measuring instrument we call a language test. The value of the theoretical framework employed here is that it sees the relationship between fundamental concepts, and it does so neither by separating the empirical analyses of a test from its social effects, nor by considering the technical qualities of validity and/or reliability as superior to aspects relating to the social dimension of a test. The framework challenges test developers to consider questions related to every aspect of the design, development and implementation of the test, bringing to the fore issues that would otherwise be ignored or relativised.

The main questions this study set out to investigate have been answered in the course of this investigation:

- Construct and other conventional forms of validity are not enough in the conventional sense, to validate, or to ensure that a language test conforms to responsible design principles. What is needed, in addition, is a detailed look at issues of transparency, accessibility and accountability as well as other conditions as reflected in the framework utilised here;
- The empirical analyses of a test are important - they do assist in taking decisions about the social, ethical and related dimensions of tests, and all these dimensions are therefore complementary and mutually dependant;
- Transparency and accountability are essential in keeping the design and production process of a test, as well as its administration, open to public scrutiny;

Chapters 5, 6 and 7 have focused on an in-depth discussion of the regulative concepts of transparency, accessibility and accountability in the design and development of TALPS. These regulative concepts tie the technical, qualifying dimension of the designed test to its juridical and ethical aspects (Weideman, 2009a: 249). The juridical analogies within the technical give rise to a condition or test design requirement that dictates that such a test must be transparent, fair and publicly defensible, while the ethical analogies generate a design condition that calls for a sense of care and concern for others (Weideman, 2009a: 249). What must be remembered is that these conditions (transparency, accessibility and accountability) do not function in isolation, but are tied in closely to each other and to a number of other conditions reflected in the framework employed in this study (see chapter 3). Weideman explains that “all of these connections between uniquely different, but related, modes of reality echo the idea that while there are unique dimensions to our experience,

none of them is absolute and each of them is necessarily related to all the others" (2009a: 243).

Importantly, this framework

> makes possible applied linguistic concept-formation; the various analogical technical concepts (labelled retrocipatory and anticipatory moments) lie at the basis of applied linguistic concept-formation (Weideman, 2009b: 66).

What this means is that each 'concept' in the framework is a condition or design principle that must be considered since "each yields a normative moment, i.e. an injunction about what the test designer should do if he or she were to be a responsible test developer" (Weideman, 2006a: 84). In light of this, then, and in order to give a more complete account of the design and development of TALPS, this final chapter will attempt to do the following:

- Consider the link between transparency, accessibility and accountability;
- Focus briefly on the conditions in the framework, or aspects of these conditions, that are not the main focus of this study but which, nonetheless, contribute to ensuring the design of a fair and socially acceptable test;
- Conclude by establishing whether the key questions posed in this study have been answered.

8.2 The link between transparency, accessibility and accountability

Before attempting to consider the link between transparency, accessibility and accountability, it must be remembered that one of the values of the framework employed in this study is in its "separating out what is conceptually distinct" (Weideman, 2009a: 249). While the concepts of transparency, accessibility and

accountability fall under the umbrella of “regulative conditions”, it is important that we not resort, to use Messick’s (1989a; 1989b) words, to “blurring the distinction” between these concepts. Each concept is distinct in what it brings to the design of fair and ethical tests. While not every concept is sufficient, all concepts are necessary to design fair tests. A number of questions arise in this regard:

1. What is the link between these three concepts (transparency, accessibility and accountability)?
2. Does a concern for transparency and/or accessibility automatically ensure accountability?
3. On the other hand, can there be accountability without transparency or accessibility?
4. Further still, are transparency, accessibility and accountability enough? If not, how can they be augmented as design criteria, and what arguments can be made for employing them as such?

To answer the questions posed above, while these concepts are at least to a certain degree linked, a concern for one does not ensure the presence of the other. Making information available to test users and others, or ensuring that the test designed is accessible in terms of its language, content, instruction to test takers, choice of reading texts and theme, is no guarantee that the test developer can, or will, be held accountable. Making information available to test takers is a first and definite step in the right direction, but transparency and accessibility do not automatically ensure accountability.

Firstly, a lot depends on the kind of information made public. As pointed out in the previous chapter, information presented to the public must be valuable and informative, presented in language that the public understands. It should not be a ‘monologue’ (Bovens, 2005: 10) but should present opportunities for the public to engage in a dialogue with those involved. Issues of accountability

should also not be limited to the development of codes or the imposition of sanctions and punishments, but should instead focus on doing that which is right and fair. The need to act responsibly should also not depend on some authority forcing one to comply but should rather be one of the standards that test developers choose to work within.

On the other hand, one cannot have accountability without transparency and accessibility. Information must be available for people to ask questions and mechanisms must be in place to facilitate dialogue between all involved. In considering the issue of accountability, it must be remembered that, as pointed out in the previous chapter, all facets of accountability need to be considered (theoretical, public, academic). In answer to the final question raised above, transparency, accessibility and accountability (theoretical, public and academic/institutional) alone are not enough to ensure the design, development and implementation of a fair test. However, should the concepts of transparency, accessibility and accountability be considered together with the other conditions in the framework, we would have a greater chance of developing such a test.

8.3 Designing a fair test

The second part of this chapter is focused on the conditions in the framework, or aspects of these conditions not touched on in this study, but which are nonetheless essential if one were to design tests that are fair and socially acceptable. To do this we look once again at a diagrammatic representation of the framework on which this study is based:

Table 8.1 Constitutive and regulative moments in applied linguistic designs

Aspect / function / dimension / mode of experience	**Kind of function**	**Retrocipatory / anticipatory moment**
numerical	constitutive	unity within a multiplicity of sets of evidence and conditions for (test) design
kinematic		internal consistency (technical reliability)
physical		internal effect / power (validity)
organic		technical differentiation
feeling		technical perception and intention
analytical	founding	design rationale (construct validity or theoretical defensibility)
technical	qualifying / leading function (of the design)	
lingual	regulative	articulation of design in a blueprint / plan
social		implementation / administration
economic		technical utility, frugality
aesthetic		harmonisation of conflicts, resolving misalignment
juridical		transparency, public defensibility, fairness, legitimacy
ethical		accountability, care, service

(Weideman, 2007a: 602)

While the focus of this study has been on the regulative conditions of transparency, accessibility and accountability, in order to give a full and complete picture of the design and development of TALPS, several other sets of functions have been discussed in detail: the *constitutive* concepts generated by the retrocipatory analogies, which include the *founding* function of a test design, its design rationale, often called its construct validity. In terms of the

set of constitutive functions, chapters 3 and 4 offered a detailed description *inter alia* of the empirical analyses of TALPS in the form of a validation argument. These same chapters covered, in as much detail as possible, information about the theoretical defensibility or construct validity i.e. the foundational function of the measuring instrument. In terms of the *regulative* conditions, transparency, accessibility and accountability operate as analogies, within the technical function, of the juridical and ethical modes of experience and have been discussed in chapter 5, 6 and 7. The rest of this chapter will touch briefly on analogies of the lingual, social, economic and aesthetic modes of experience within the leading technical mode of test designs as they relate to the design and development of TALPS.

8.4 The lingual analogy

Weideman (2009a: 247) points out that the *lingual* dimension of test design finds its expression in the form of the technical articulation of a blueprint or a set of specifications for the test. Briefly, this means that a test cannot be responsibly designed and developed if its construct is not expressed or overtly signalled, both for initial design and for future control. This has been discussed in chapter 3. Related to this lingual analogy within the technical is, furthermore, the "appropriate technical interpretation of the test scores" (2009a: 247), an aspect that will be discussed below.

8.4.1 Interpreting the results of the test

It is the responsibility of test designers to stipulate how to interpret the results of a test. Because test results almost always have effects (positive and negative) on test takers, it is imperative that a test is administered and scored according to the test developers' instructions (*AERA*, 1999: 61). With regards to the interpretation of the results of the test, the test designers of TALPS had two important considerations: determining the cut score (or cut-off point) and

providing advice to test users and test takers on how to interpret the results of the test.

8.4.2 Cut scores

AERA (1999: 175) defines a cut score as "a specified point on a score scale, such that scores at or above that point are interpreted or acted upon differently from scores below that point". In determining the cut-off point for TALPS, the test designers turned to the historical data they had gathered when using and administering the ELSA Plus, TALL and TAG. According to Van der Slik and Weideman (2005: 25–26) the cut-off point for one administration of the undergraduate English test (TALL) was set at 58 points, while the cut-off point for the Afrikaans version (TAG) was set at 52 points. They explain that

> historically, i.e. on the previous test (ELSA Plus), between 24% and 27% of students who wrote the Afrikaans test were identified as being at risk, while between 31% and 38% of students who took the English test failed. Based on past experience, on average about 31% of the total number of students is identified as being at risk (Van der Slik & Weideman, 2005: 26).

Their main argument for determining a cut-off point in this case has to do with "the comparative performance of the candidates who write a specific test on a specific day, and compares their performance with that of their peers on the same test" (Van der Slik & Weideman, 2005: 26). The cut-off point for TALL and TAG is conventionally about 10% under the average performance of the test, i.e. a student is at risk when his/her performance is 10% under the average of his/her peer group (Van Dyk, 2010). Van Dyk (2010) points out that in the case of TALL and TAG, provision has been made for potential misclassification. These misclassifications lie between 0.15–0.29 around the average – this means that everyone between 0.075–0.145 (Standard deviation) may be given a second chance or borderline test. This test usually takes place around two weeks after the first test.

In the case of TALPS, the test designers chose to use the same method to determine the cut-off point, in light of the fact that the students writing TALPS, though academically more mature, were very similar to the group of students who write TALL and TAG. Like TALL and TAG, these decisions are based on the most responsible arguments available. The cut-off point for TALPS, provisionally 60 points, still needs to be confirmed or modified with reference to sufficient empirical data. That is the subject of another study (Du Plessis, 2012), and will be a necessary step in ensuring the adequate and appropriate interpretation of TALPS results.

8.4.3 TALPS scoring scale

An interesting observation by Butler in his study (2007: 9) was that students rate themselves high in terms of their academic literacy abilities. Evidence collected by Butler (2007) shows that supervisors feel differently: “A large majority of respondents believe that their postgraduate students’ academic literacy levels range from average to poor” (2007: 125). What is important is what students think. If they believe that they are equipped with the academic literacy ability required, then they will not believe that the test or the intervention would be useful. Students who are shown to be at risk will have a negative attitude to the test and the intervention, which will in turn impact on their performance. A first step in convincing students of the value of the test would be to ensure the face validity of the test. By looking at the test, students should believe that it tests what is essential for them to know at tertiary level. One other way would be to ensure that they understand what the test score means. Allocating a ‘Pass’ or ‘Fail’ or a test score that was difficult to interpret would further stigmatise the students and the test as discussed in chapter 2.

Find below an example of a score scale that does not indicate a ‘Pass’ or ‘Fail’ but rather indicates at what level the student is at or what intervention is best suited to a particular score. The table presents the general guidelines for

interpreting the final test scores for the Standardised Assessment Test for Access and Placement (SATAP):

Table 8.2 Guidelines for interpreting the test scores for the SATAP

Less than 30 per cent	These students might have serious difficulties with academic work
30 – 40 per cent	It is advised that these students be placed on a foundation programme or extended curriculum
40 – 55 per cent	These students might need targeted assistance on the mainstream curriculum
Over 55 per cent	These students are unlikely to experience difficulties with academic work

(Scholtz & Allen–lle, 2007: 924)

This indicates a move away from the way in which students were scored previously, the focus now being on de-stigmatising the results. The test developers of TALL and TALPS use a similar scale as discussed in chapter 2. It is a scale that does not distinguish between a ‘Pass’ or ‘Fail’ but rather indicates the level of risk the student has. Find below the scoring scale for TALPS as well as advice to students on how this should be interpreted:

Table 8.3 Guidelines for interpreting the test scores for TALPS

CODE	INTERPRETATION
CODE 1 (0–33%)	**High risk: EOT 300 is compulsory.**
CODE 2 (34–55%)	**Clear risk: EOT 300 is compulsory.**
CODE 3 (56–59%)	**Risk: EOT 300 is compulsory.**
CODE 4 (60–74%)	**Less risk: You do not need to enrol for EOT 300.**
CODE 5 (75+)	**Little to no risk: You do not need to enrol for EOT 300.**

8.5 The social anticipation within the technical

The *social* mode of experience is reflected within the technical in the idea that we have of the implementation and administration of the test, an aspect covered in chapter 6. Important also, when considering the social analogy, is the aspect of the use of the test and the impact of the test.

8.5.1 The use of the test

One of the key considerations of the test designers when designing a test of this nature is the question of how the results of the test will be used. The advice that should be given about the use of the test has already been discussed in detail in chapter 6, but will be touched on briefly here. According to *AERA*:

> When educational testing programs are mandated by school, district, state, or other authorities, the ways in which test results are intended to be used should be clearly described. It is the responsibility of those who mandate the use of tests to monitor their impact and to identify and minimise potential negative consequences (*AERA*, 1999: 145).

In a responsible conception of what test design and use involves, there is a shift in responsibility from those who use tests to the test designers themselves. Once the need for TALPS had been established, the next important consideration was, therefore, the question of how the results of the test would be used. The use of tests to deny access has been well documented in the literature, as the introduction to this study attempts to show. Low language proficiency among students has its roots in the history of this country. The focus in education and testing today, however, should be on granting rather than denying access (Fulcher & Davidson, 2007: 412). A test like TALPS can be used to do exactly that – create access. Without the intervention nurturing the development of academic literacy that should follow the administration of the test many students may not successfully complete their studies. In discussing the SATAP (The Standardised Assessment Test for Selection and Placement), Scholtz and Allen-lle (2007) state that an academic literacy test is

essential in providing "insight into the intellectual profile and academic readiness of students" and that subsequent

> interventions have positive and financial implications: the individual becomes economically productive, it improves through-put rates and subsidies for institutions, and contributes to economic advancement in South Africa (Scholtz & Allen-lle, 2007: 921).

Tests like TALL and TALPS have moved away from the stereotyped tests that have negative consequences and disadvantage the test taker. The test developers of TALPS insist that should users choose to use the test for access rather than placement, they should use at least three other criteria or instruments to measure students' abilities rather than rely solely on TALPS. According to Albert Weideman (Executive Head of ICELDA, that owns TALPS):

> The results of the test, which will be available within 48 hours for all candidates, will, however, have to be used with care. Since language cannot predict all of a candidate's future academic performance, decisions cannot be taken to exclude students on the basis of the results of a single academic literacy test. It is recommended, therefore, that those departments who wish to employ the results for decisions involving access, do not set the weight of the results of this test at more than is internationally agreed as appropriate for language tests, which is between 10% and 20%. A weighting of, say, 60% to prior academic performance is generally used, with language ability and several other criteria making up the other 40%. There is one possible exception to the rule of not excluding anyone purely on the basis of language ability. This is where the ability is so low (usually in the lowest 7½% of testees) that it raises ethical questions about allowing those in who so obviously fall short of requirements that they will waste their time and resources on a hopeless venture (Weideman, 2010).

This is in keeping with *AERA's* (1999:146) advice that,

> In educational settings, a decision or characterisation that will have a major impact on a student should not be made on the basis of a single test score. Other relevant information should be taken into account if it will enhance the overall validity of the decision *(AERA,* 1999:146).

The test developers of TALPS have considered, from as early as the design stage, to what use the results of the test would be put. This concern about the

consequences of the use of the test, and the results of the test to make judgements about the test takers, points directly to issues of responsibility on the part of the test developers. Messick (1980: 1020) states that when determining whether a test should be used for a particular purpose, one should weigh the potential social consequences of the proposed test use against the potential social consequences of *not* testing at all. The test developers of TALPS have already determined that:

1) the test can be used for the purpose for which was designed. The construct of the test is based on an appropriate definition of academic literacy;
   and
2) there would be negative effects on students (and the institution, parents and the country) if students' academic literacy levels were not tested and if adequate support is not subsequently provided.

8.5.2 The impact of TALPS

Tests have effects that extend far beyond the confines of a classroom or a single institution. They affect even those who are not directly a part of the testing/teaching situation. This part of the chapter moves towards a discussion of impact, which according to Hawkey (2006), "is a term used to describe studies which investigate the influence of language programmes and/or tests on stakeholders *beyond* the immediate learning programme context" (2006: 7). He states that an impact study may investigate the effects of a programme or test on school heads, parents, receiving institution administrators, and high stakes test providers (2006: 7). In the case of TALPS these stakeholders include parents, supervisors, test administrators, university officials and the general public. Test developers are accountable to these stakeholders and thought must be given to the way the test impacts on them as well.

A first step towards accountability, as stated previously, is ensuring transparency. These stakeholders need as much information about the test as is

available. Saville (2003) states that one area of impact concerns “promoting the public understanding of assessment and related pedagogical issues worldwide” (2003: 74). He explains that this can be achieved by providing public information, research and advisory services, and that the aim should be to achieve greater understanding of the purposes and procedures of testing and the proper uses of examination information (results, grades, etc.) (2003: 74). This has been discussed in detail in the chapters on transparency and accessibility.

Saville (2003) also points out procedures that need to be implemented once the test is used. These procedures will help determine the kind of impact created:

- Who is taking the examination (i.e. a profile of the candidates);
- Who is using the examination results and for what purpose;
- Who is teaching towards the examination and under what circumstances;
- What kinds of courses and materials are being designed and used to prepare candidates;
- What effect the examination has on public perceptions generally (e.g. regarding educational standards);
- How the examination is viewed by those directly involved in educational processes (e.g. by students, examination takers, teachers, parents, etc.);
- How the examination is viewed by members of society outside education (e.g. politicians, employers, managers)
  (Saville, 2003: 75).

Some of the information is already available to the developers of TALPS (information about candidates, test users, course and course materials). However, in order to evaluate effectively the kind of impact TALPS has, extensive research must be carried out, especially to answer the last three questions (cf. Du Plessis, 2012). One of the points stressed in the literature on the effects of testing is that washback and impact studies must be carried out. This is also true for TALPS. Information gained from a study on impact can be used to make improvements on the test and could be a valuable source of information to test developers, users, test takers, parents and the general public. Reports emanating from this kind of research can be put onto the website and/or circulated to interested parties. As Hamp-Lyons states: “Impact studies

are extremely difficult and time-consuming to do well, but they are a necessity if testers are to understand the full nature of the work to which they have committed their professional lives and its consequences" (2000b: 587).

8.6 Technical utility

According to Weideman the connection between the technical function of a test and the *economic* dimension becomes evident when we consider the idea of utility (2009a: 248). The aspect of utility has been touched on briefly in chapter 4 where the trade-off between a longer test and the reliability of TALPS was considered. To add to that, TALPS satisfies a number of conditions in terms of its utility: it has already been shortened from 170 items (150 minutes long) to 76 items (120 minutes long). The test is two hours long, not an unreasonable length considering it is a postgraduate test. In addition to the 76 multiple choice items, the last section requires students to write a short essay. This writing section is necessary, especially because writing is one the skills postgraduate students must master. In terms of the face validity of TALPS (see chapter 4) students need to 'see' the link between the test and what is expected from them at postgraduate level. The test is not expensive to administer, nor does it place unreasonable demands on the student, as it costs under R90 to write the test.

8.7 Technical alignment and harmonisation

The *aesthetic* analogy, according to the framework above, finds expression in the "harmonising of conflicts, resolving misalignment" (Weideman, 2007a: 602). The reference to 'harmony' could be seen as the test designers' concern with ensuring alignment between the test and the intervention that follows (academic/institutional accountability). This is preceded by an alignment of the test tasks and the construct of the test: a technical misalignment here is undesirable in terms of responsible test design.

A detailed analysis of the alignment of TALPS and EOT 300 has been done in the previous chapter. Based on this analysis, it can be concluded that the test can possibly be followed by effective teaching and learning in the form of a well designed intervention module focused on helping students improve their academic literacy levels. Further evidence of the aesthetic anticipatory moments within the qualifying technical aspect of the applied linguistic design (Weideman, 2007a: 601) can be found in the various "trade-offs" the test designers have made in the design and development of TALL, TAG and TALPS (see chapters 3 and 4). According to Weideman (2007a: 601), "each such trade-off generates a need to weigh or assess, harmonise and justify a tough and responsible technical design decision".

8.8 Conclusion

This final discussion relating to the fundamental concepts in language testing, as they relate to TALPS, has presented us with a more complete account of the design and development of the test than we have had before. In doing this we have answered the question posed in the conclusion of chapter 7: that issues of transparency, accessibility and accountability are not enough to ensure fair, socially acceptable tests. This chapter has, therefore, focused on the other conditions for test design, as reflected in the framework, in relation to TALPS. In doing this, we have answered a key question posed in this study: have we, as test designers, succeeded in designing a socially acceptable, fair and accessible test? The application of the framework to the design, development and implementation of TALPS has presented us with sufficient evidence to answer this question. In terms of the analysis done in this study, we have to conclude that TALPS satisfies the requirements necessary to be considered a socially acceptable, fair and accessible test. Importantly, our investigation has demonstrated that these conditions for test design can be theoretically accounted for in terms of a framework.

If this framework is valuable in terms of ensuring the design and development of socially acceptable tests, we can assume that it could be applied, similarly, to other applied linguistic designs such as curricula and courses, and even to institutional language plans, but these fall outside the scope of this study. The degree of alignment of TALPS with the principles of responsible test design has also, however, been shown not to be perfect in every respect. Impact and washback studies are indicated as necessary augmentations of current measures of TALPS conforming to conditions for test design. This should not be surprising: tests need to be scrutinised and re-subjected to scrutiny all the time. They are never perfect. In this respect Messick (1989b) is correct: the validation process is never over. This study has shown, however, that it is not only the validity of a test that should be scrutinised for each new context in which it is administered, but all of its technical design features. Rather than ongoing validation, what we seem to need is an ongoing examination to determine whether or how a test conforms to principles or conditions for responsible test design.

A further possible limitation of this study is that it might not yet have fully uncovered all regulative conditions for test design. For example, when the technical dimension of experience anticipates the juridical, the ideas that emerge are possibly not limited to accountability, but may extend further to notions, as yet unexamined, like the technical ratification of test use, i.e. whether a test does full justice to the measuring of language ability and is accepted as a fair instrument despite prior challenges to it. Has it withstood the "test of time"? Does it have legitimacy? TALPS is too young and it may be too early to tell. This hangs together with the technical commitment that a test demonstrates, *inter alia*, to responsible design as well as to the care and concern for others. How trustworthy a measurement is it, and how general/specific is the trust we can place in it? While some of this has been investigated, further exploration is needed.

References

African Tax Institute. 2010a. *Information brochure: Masters program in taxation.* Pretoria: University of Pretoria.

African Tax Institute. 2010b. *Guidelines regarding the MPhil: Taxation entrance examination.* Pretoria: University of Pretoria.

Alderson, J. 1995. *Language testing in the 1990s*. Hertfordshire: Macmillan.

Alderson, J.C. & Wall, D. 1993. Does washback exist? *Applied linguistics*, 14(2): 115 – 129.

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

American Educational Research Association (AERA). 1999. *Standards for educational and psychological testing.* Washington: American Educational Research Association.

American Psychological Association (APA). 1988. *Code of fair testing practices in education*. [Online]. Available: http://www.apa.org/science/programs/testing/fair-testing. (Accessed 2010/10/07).

American Psychological Association (APA). 1998. *The rights and responsibilities of test takers: Guidelines and expectations.* [Online]. Available: http://www.apa.org/science/programs/testing/rights.aspx (Accessed 2010/10/07).

Association of Language Testers in Europe (ALTE). 1994. *The ALTE Code of Practice*. [Online]. Available: http://alte.org/cop/index.php (Accessed 2009/09/07).

Association of Language Testers in Europe (ALTE). 2001. *Principles of good practice for ALTE exams.* [Online]. Available: www.alte.org/cop/principles.php (Accessed 2009/09/07).

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

Bachman, L.F & Palmer, A.S. 1996. *Language testing in practice*: *Designing and developing useful language tests*. Oxford: Oxford University Press.

Barry, D. 2002. Language equity and assessment in South African education. *Journal for language teaching*, 36(1&2): 105 – 117.

Beddow, P.A., Kettler, R.J. & Elliot, S.N. 2008. *Test accessibility and modification inventory*. Nashville, TN: Vanderbilt University.

Beu, D.S. & Buckley, M.R. 2004. Using accountability to create a more ethical climate. *Human resource management review*, 14: 67 – 83.

Blanton, L.L. 1994. Discourse, artefacts and the Ozarks: Understanding academic literacy. In Zamel, V. & Spack, R. (eds.), *Negotiating academic literacies: Teaching and learning across languages and cultures*. New Jersey: Lawrence Erlbaum Associates. 235 – 319.

Borsboom, D., Mellenburg, G.J. & Van Heerden, J. 2004. The concept of validity. *Psychological review*, 111(4): 1061 – 1071.

Bovens, M. 2005. Public accountability: A framework for the analysis and assessment of accountability arrangements in the public domain. In Ferlie, E., Lynne, L. & Pollitt, C. (eds.), *The Oxford handbook of public management.* Oxford: Oxford University Press. 1 – 36.

Boyd, K. & Davies, A. 2002. Doctors' orders for language testers. *Language testing*, 19(3): 296 – 322.

Bureau for Institutional Research and Planning (BIREP). 2011. Pretoria: University of Pretoria.

Butler, H.G. 2007. *A framework for course design in academic writing for tertiary education.* Unpublished doctoral thesis. Pretoria: University of Pretoria.

Butler, H.G. 2009. The design of a postgraduate test of academic literacy: Accommodating student and supervisor perceptions. *Southern African linguistics and applied language studies*, 27(3): 291 – 300.

Butler, H.G. & Van Dyk, T.J. 2004. An academic English language intervention for first year engineering students. *South African journal of linguistics*, 22 (1&2): 1 – 8.

Butler, H.G., Pretorius, R.E. & Van Dyk, T.J. 2009. Unit for Academic Literacy EOT 300. Unpublished class notes. Pretoria: University of Pretoria.

Bygate, M. 2004. Some current trends in applied linguistics: Towards a generic view. *AILA review*, 17: 6 – 22.

Bygate, M. 2005. Applied linguistics: A pragmatic discipline, a generic discipline? *Applied linguistics*, 26(4): 568 – 581.

Carstens, A. 2009. *The effectiveness of genre-based approaches in teaching academic literacy: Subject-specific versus cross-disciplinary.* Unpublished PhD thesis. Pretoria: University of Pretoria.

Centre for Environmental Studies, The. 2010. *Information brochure*. Pretoria: University of Pretoria.

CITO. 2006. *TiaPlus, Classical test and item analysis* ©. Arnhem: Cito M. & R. Department.

Cliff, A., Crandall, J., DeKadt, E. & Hubbard, H. 2003. Report of panel: Departmental evaluation of the Unit for Language Skills Development. Pretoria: University of Pretoria.

Council on Higher Education (CHE). 2004. *Higher education monitor: South African higher education: past, present and future.* Pretoria: Council on Higher Education.

Council on Higher Education (CHE). 2006. *Eternal (and Internal) tensions? Conceptualising public accountability in South African higher education*. Pretoria: Council on Higher Education.

Coxhead, A. 2000. A new academic word list. *TESOL quarterly*, 34(2): 213 – 238.

Cronbach. L.J. & Meel, P.E. 1955. Construct validity in language tests. *Psychological bulletin*, 52(4): 281 – 302.

Davidson, F. & Lynch, B.K. 2002. *Testcraft.* New Haven: Yale University Press.

Davies, A. 1990. Principles of language testing. In Crystal, D. & Johnson, K. (eds.), *Applied language studies*. Cambridge: Basil Blackwell.

Davies, A. 1997. Demands of being professional in language testing. *Language testing*, 14(3): 328 – 339.

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

Davies, A. & Elder, C. 2005. Validity and validation in language testing. In Hinkel, E. (ed.), *Handbook of research in second language teaching and learning*. New Jersey: Lawrence Erllbaum Associates. 795 – 813.

Department of Education. 1997. Education white paper 3: A programme for the transformation of higher education. Pretoria: *Government gazette* No. 18207, 15 August.

Dill, D.D. 1999. Academic accountability and university adaptation: The architechure of an academic learning organisation. *Higher education*, 38: 127 – 154.

Du Plessis, C. 2012. *The design and reception of a test of academic literacy for postgraduate students.* Bloemfontein: MA dissertation (in preparation).

Frink, D.D. & Klimoski, R.J. 2004. Advancing accountability theory and practice: Introduction to the human resource management review special edition. *Human resource management review*, 14: 1 – 17.

Fulcher, G. 1997. Text difficulty and accessibility: Reading formulae and expert judgement. *System*, 25(4): 497 – 513.

Fulcher, G. 1999. Ethics in language testing. TAE SIG Newsletter – Special Conference Issue, 1(1). [Online]. Available: http://taesig.8m.com/news1.html. (Accessed 2009/09/23).

Fulcher, G. & Davidson, F. 2007. *Language testing and assessment: An advanced resource book.* New York: Routledge.

Fulcher, G. & Davidson, F. 2008. Tests in life and learning: A deathly dialogue. *Educational philosophy and theory*, 40(3): 407 – 417.

Garger, J. 2008. Determining readability: The Flesch reading ease test. [Online] Available: http://www.brighthub.com/education/languages/articles/15302.aspx. (Accessed 2009/08/03).

Geldenhuys, J. 2007. Test efficiency and utility: Longer or shorter tests. *Ensovoort*, 11 (2): 71 – 82.

Gliem, J.A. & Gliem, R.R. 2003. Calculating, interpreting, and reporting Cronbach's alpha reliability coefficient for Likert-type scales. Unpublished paper delivered at the 'Midwest research to practice conference in adult, continuing and community education', 8 – 10 October, Columbus: 82 – 88.

Global Transparency Initiative. 2006. *Transparency charter for international financial institutions: Claiming our right to know.* [Online]. Available: www.ifitransparency.org. (Accessed 2010/09/23).

Grow your website. [Online]. Available: http://web.up.ac.za/sitefiles/file/weboffice/workshops/Grow-your-website-Presentation.pdf. (Accessed 2009/08/03).

Hamp-Lyons, L. 1997. Ethics in language testing. In Clapham, C.M. & Corson, D. (eds.), *Language testing and assessment: Encyclopaedia of language and education* 7. Dordrecht: Kluwer Academic. 323 – 333.

Hamp-Lyons, L. 2000a. Fairness in language testing. *Studies in language testing*, 9: 30 – 34.

Hamp-Lyons, L. 2000b. Social, professional and individual responsibility in language testing. *System*, 28: 579 – 591.

Hamp-Lyons, L. 2001. Ethics, fairness(es), and developments in language testing. *Studies in language testing*, 11: 222 – 227.

Hawkey, R. 2006. *Impact theory and practice: Studies of the IELTS test and Progetto Lingue 2000*. Cambridge: Cambridge University Press.

Ho, R. 2006. *Handbook of univariate and multivariate data analysis and interpretation with SPSS.* Boca Raton: Chapman & Hall.

Huot, B. 1990. Reliability, validity, and holistic scoring: What we know and what we need to know. *College composition and communication,* 41(20): 201 – 213.

IMS Global Learning Consortium. *Guidelines for developing accessible learning applications*. [Online]. Available: http://www.imsglobal.org/accessibility/accessiblevers/index.html. (Accessed 2009/07/28).

Inter-institutional Centre for Language Development and Assessment (ICELDA). 2011. [Online]. Available: http://icelda.sun.az.za. (Accessed 2011/10/10).

International English Language Testing System (IELTS). *How much does it cost?* [Online]. Available: http://www.ielts.org/test_takers_information/test_takers_faqs/registering for_the_test.aspx#Howmuchdoesitcost. (Accessed 2010/10/11).

International Language Testing Association (ILTA). 2000. *Code of ethics for ILTA.* [Online]. Available: http://www.iltaonline.com. (Accessed 2009/09/18).

International Language Testing Association (ILTA). 2005. *ILTA: Draft code of practice: Version 3.* [Online]. Available: http://www.iltaonline.com. (Accessed 2009/09/18).

Joint Committee on Testing Practices. 2004. *Code of fair testing practices in education.*[Online]. Available: http://www.apa.org/science/programs/testing/fair-code.aspx. (Accessed 2012/1/13).

Kane, M. 2011. Validating score interpretations and uses. *Language testing,* 29(1): 3 – 17.

Kunnan, A.J. (ed.). 2000. *Studies in language testing 9: Fairness and validation in language assessment*. Cambridge: Cambridge University Press.

Kunnan, A. 2004. Test fairness, in Milanovic, M. & Weir, C. (eds.). *Studies in language testing,* 18. Cambridge: Cambridge Univesity Press. 27 – 45.

Kunnan, A.J. 2006. *ILTA code of ethics and beyond*. [Online]. Available: http://www.alte.org/further_info/sofia/ak_1_1106.pdf. (Accessed 2009/09/18).

Kearns, K.P. 1998. Institutional accountability in higher education: A strategic approach. *Public productivity & management review*, 22(2): 140 – 156.

Kurpius, S.E.R. & Stafford, M.E. 2006. *Testing and measurement: A user-friendly guide.* California: Sage Publications.

Likert, R. 1961. *New patterns of management*. New York: McGraw-Hill Book Company.

Longman Dictionary of contemporary English. 2004. Harlow: Pearson Education Limited.

McIver, J.P. & Carmines, E.G. 1981. *Unidimensional scaling*. Beverly Hills: Sage Publications.

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

McNamara, T. 2005. The social turn in language assessment. In McNamara, T., Brown, A., Grove, L., Hill, K. & Iwashita, N. (eds.), *Handbook of research in second language teaching and learning.* New Jersey: Lawrence Erlbaum Associates, Publishers. 775 – 778.

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

Messick, S. 1980. Test validity and the ethics of assessment. *American pathologist*, 35: 1012 – 1027.

Messick, S. 1981. Evidence and ethics in the evaluation of tests. *Educational researcher*, 10(9): 9 – 20.

Messick, S. 1989a. Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, 18(2): 5 – 11.

Messick, S. 1989b. Validity. In Linn, R.L. (ed), *Educational measurement* (3rd ed.). New York: American Council on Education & Macmillan. 13 – 103.

Messick, S. 1996. Validity and washback in language testing. *Language testing*, 13(3): 241 – 256.

Ministry of Education. 2001. *National plan for higher education*. Pretoria: Department of Education.

Naurin, D. 2007. Transparency, publicity, accountability – the missing links. Unpublished paper delivered at the CONNEX-RG 2 workshop on 'Delegation and mechanisms of accountability in the EU', 8 – 9 March, Uppsala.

Nevo, D. & Shohamy, E. 1986. Evaluation standards for the assessment of alternative testing methods: An application. *Studies in educational evaluation*, 12: 149 – 158.

Norton, B. & Stein, P. 1998. Why the "Monkeys passage" bombed: Tests, genres, and teaching. In Kunnan, A.J. (ed.), *Validation in language assessment*. Mahwah: Lawrence Erlbaum Associates. 231 – 249.

Norton, B. 1997. Accountability in language assessment. In Clapham, C. & Corson, D. (eds.), *Language testing and assessment: Encyclopaedia of language and education* 7. Dordrecht: Kluwer Academic. 313 – 322.

Popham, W.J. 1997. Consequential validity: right concern – wrong concept. *Educational measurement: issues and practice*, Summer: 9 – 13.

Rea-Dickens, P. 1997. So, why do we need relationships with stakeholders in language testing? A view from the U.K. *Language testing*, 14(3): 304 – 314.

Santos, J.R.A. 1999. Cronbach's alpha: a tool for assessing the reliability of scales. *Journal of extension*, 37(2): 1 – 5.

Saville, N. 2003. The process of test development and revision within UCLES EFL. In Weir, C. & Milanovic, M. (eds.), *Continuity and innovation: Revising the Cambridge Proficiency in English examination 1913–2002.* Cambridge: Cambridge University Press. 57 – 120.

Scholtz, D. & Allen-lle, C.O.K. 2007. Is the SATAP test an indicator of academic preparedness for first year university students? *South African journal of higher education*, 21(7): 919 – 939.

Schuurman, E. 2005. *The technological world picture and an ethics of responsibility: Struggles in the ethics of technology.* Sioux Center, Iowa: Dordt College Press.

Schuurman, E. 2010. Responsible ethics for global technology. *Axiomathes*, 20: 107–127.

Shepard, L.A. 1993. Evaluating test validity. *Review of research in education*, 19: 405 – 50.

Shohamy, E. 1997. Testing methods, testing consequences: are they ethical? Are they fair? *Language testing*, 14(3): 340 – 349.

Shohamy, E. 2001. *The power of tests: A critical perspective on the uses of language tests*. London: Longman.

Shohamy, E. 2004. Assessment in multicultural societies: Applying democratic principles and practices to language testing. In Norton, B. & Toohey, K. (eds.), *Critical pedagogies and language learning*. New York: Cambridge University Press. 72 – 92.

Shohamy, E. 2008. Language policy and language assessment: The relationship. *Current issues in language planning* , 9(3): 363 – 373.

Sinclair, A. 1995. The chameleon of accountability: Forms and discourses. *Accounting, organisations and society*, 20(2/3): 219 – 237.

Smith, M.L. 1991. Put to the test: The effects of external testing on teachers. *Educational researcher*, 20(5): 8 – 11.

Spolsky, B. 2008. Language testing at 25: Maturity and responsibility. *Language testing*, 25(3): 297 – 305.

Taylor, L. 2009. Developing assessment literacies. *Annual review of applied linguistics*, 29: 21 – 36.

TOEFL. [Online]. Available: http://www.ets.org/toefl/
(Accessed 2008/12/03).

TOEFL. *How much does the TOEFL test cost?* [Online].
Available: http://www.toeflgoanywhere.org/content/9-how-much-does-toefl test-cost.
(Accessed 2010/10/11).

Transparency International. 2010. *What is transparency international*? [Online]. Available: http://www.transparency.org/about us.
(Accessed 2010/02/11).

Unit for Academic Literacy. 2007. *Departmental self-evaluation*. Pretoria: University of Pretoria.

Van der Slik, F. 2006. Language proficiency and fairness. Keynote address, Southern African Applied Linguistic Association, Durban, 6 July.

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

Van der Slik, F. & Weideman, A. 2008. Measures of improvement in academic literacy. *Southern African linguistics and applied language studies*, 26(3): 363 – 378.

Van der Slik, F. & Weideman, A. 2009. Revisiting test stability: further evidence relating to the measurement of difference in performance on a test of academic literacy. *South African linguistics and applied language studies*, 27(3): 253 – 263.

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. & Weideman, A. 2004a. Switching constructs: On the selection of an appropriate blueprint for academic literacy assessment. *SAALT Journal for language teaching*, 38(1): 1 – 13.

Van Dyk, T. & Weideman, A. 2004b. Finding the right measure: From blueprint to specification to item type. *SAALT Journal for language teaching*, 38(1): 15 – 24.

Van, Dyk, T.J. 2005. Towards providing effective academic literacy intervention. *Per linguam*, 21(2): 38 – 51.

Van Dyk, T.J. 2010. *Konstitutiewe voorwaardes vir die ontwerp van ʾn toets van akademiese geletterdheid.* Unpublished PhD thesis. Bloemfontein: University of the Free State.

Van Rensburg, C. & Weideman, A. 2002. Language proficiency: Current strategies, future remedies. *SAALT Journal for language teaching*, 36 (1 & 2): 152 – 164.

Webb, V.N. 2002. English as a second language in South Africs's tertiary institutions: A case study at the University of Pretoria. *World Englishes*, 21(1): 49 – 69.

Webometrics ranking of world universities. [Online]. Available: http://www.webometrics.info/. (Accessed 2009/03/08).

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

Weideman, A. 2003b. Towards accountability: A point of orientation for post-modern applied linguistics in the third millennium. *Literator*, 24(1): 83 – 102.

Weideman, A. 2006a. Transparency and accountability in applied linguistics. *Southern African linguistics and applied language studies*, 24(1): 71 – 86.

Weideman, A. 2006b. Assessing academic literacy: a task-based approach. *Language matters*, 37(1): 81 – 101.

Weideman, A. 2007a. The redefinition of applied linguistics: modernist and postmodernist views. *South African linguistics and applied language studies*, 24(1): 589 – 605.

Weideman, A. 2007b. A responsible agenda for applied linguistics: Confessions of a philosopher. *Per linguam*, 23(2): 29 – 53.

Weideman, A. 2009a. Constitutive and regulative conditions for the assessment of academic literacy. *South African linguistics and applied language studies*, 27(3): 235 – 251.

Weideman, A. 2009b. Uncharted territory: A complex systems approach as an emerging paradigm in applied linguistics. *Per linguam*, 25(1): 61 – 75.

Weideman, A. 2010. Draft announcement: administration of TALPS, e-mail to A.Rambiritch. [Online], 18 August. Available e-mail: weidemanAJ@ufs.ac.za.

Weideman, A. 2011. Straddling three disciplines: foundational questions for a language department. *Acta varia*, (1): 1 – 23.

Weideman, A. & Butler, G. 2006. Project Proposal: Postgraduate academic literacy initiative. University of Pretoria: Pretoria.

Weideman, A. & Van der Slik, F. 2008. The stability of test design: Measuring difference in performance across several administrations of a test of academic literacy. *Acta academica,* 40 (1): 161 – 182.

Yeld, N. 2000. The construct of the academic literacy test (PTEEP). Mimeograph. Cape Town: Alternative Admissions Research Project, University of Cape Town.

ANNEXURE A

ACADEMIC LITERACY TEST
CONTROL FORM

FACULTY: _____

VENUE: _____

DATE:			
TIME			

NUMBER OF TESTS RECEIVED:	
NUMBER OF TESTS USED:	
NUMBER OF TESTS WITH DEFECTS:	
NUMBER OF UNUSED TESTS:	

NAME AND SIGNATURE OF INVIGILATOR:

_____ _____

NAME **SIGNATURE**

APPENDIX A TALPS PROJECT PROPOSAL

Universiteit van Pretoria
University of Pretoria

Postgraduate academic literacy initiative

Project proposal

Project leader

Prof. Albert Weideman
Unit for Academic Literacy
Room 17-15
Human Sciences Building
University of Pretoria
0002 Pretoria

albert.weideman@up.ac.za

Tel. (012) 420 4957
Fax: (012) 420 3682

Project officer

Mr Gustav Butler
Unit for Academic Literacy
Room 17-16
Human Sciences Building
University of Pretoria
0002 Pretoria

gustav.butler@up.ac.za

Tel. (012) 420 2334
Fax: (012) 420 3682

August 2006

unit for academic literacy
eenheid vir akademiese geletterdheid

Executive summary

The Postgraduate academic literacy initiative will address the urgent institutional need to develop the academic literacy levels of the increasing numbers of postgraduate students of the University of Pretoria who do not have English as a first language, and may be at risk academically, both in terms of delivering quality writing, and in the completion of their studies.

The project will have two main deliverables:

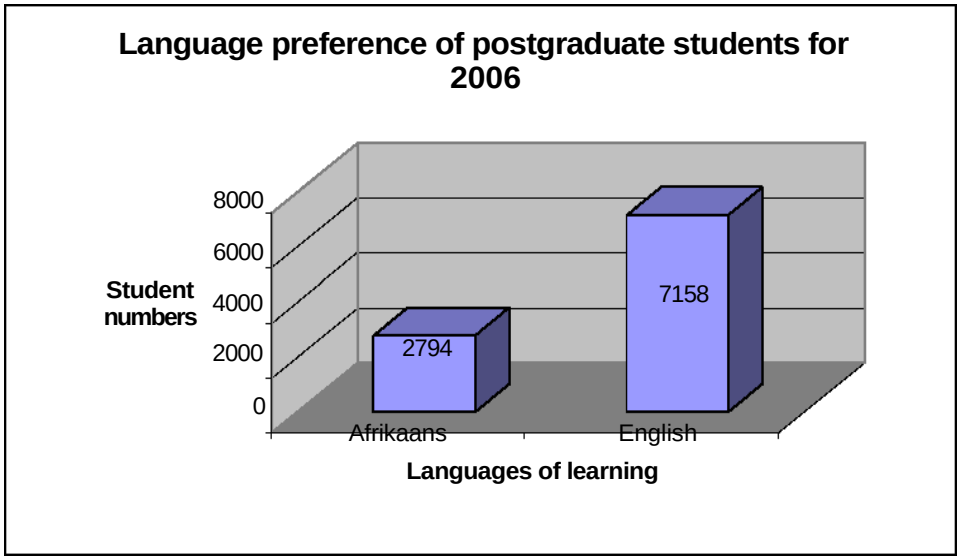
- a postgraduate test of academic literacy levels and
- an intervention designed, *inter alia*, on the basis of the test results, for the development of the academic literacy levels especially of master's and doctoral students.

The results will be disseminated both nationally and internationally, and the initiative will indirectly benefit also some of our European partners (at the Radboud University of Nijmegen) who are engaged in a similar project.

Motivation

Within the University of Pretoria, English is becoming ever more dominant as the language of learning and study at postgraduate level, especially at the master's and doctoral level. Current studies (Butler, forthcoming) about this that are being done within the Unit for Academic Literacy point out that some 72% of the almost 10000 postgraduate students registered at the University of Pretoria prefer English (Figure 1, below):

Figure 1 Language preference of postgraduate students at the UP for 2006



There is legitimate concern, however, about the academic literacy levels of the increasing numbers of students both from South Africa and from other parts of Africa and the world who are not first language users of English, and who are currently enrolling for master's and doctoral degrees. Only 2731 of the registered postgraduate students are first language users of English; a good 45% of post-graduate students are additional language users of English.

Both our experience with the currently available advanced writing skills course for postgraduate students and the outcomes of the study being referred to above indicate that this is a problem that is becoming ever more serious, and that a special initiative is needed to overcome it.

A top institutional priority

The quality of tuition and learning available to postgraduate students is currently a top institutional priority of the University of Pretoria, whose strategic intent is to be an internationally recognised and locally relevant institution of higher education.

Since the South African higher education context also presents a more affordable, and often more relevant, alternative to study in the developed world, being locally relevant also means being relevant to the continent of Africa. The increasing numbers of postgraduate students, especially from Africa, are evidence of the growing responsibilities of the University of Pretoria in this regard.

The increasing demand for our current course (EOT 300) is also a clear indication to us that this is a problem requiring urgent attention.

The project

The Postgraduate academic literacy initiative has two major components:

Test development

We intend to produce an assessment instrument to test the academic literacy levels of postgraduate students. The production of this test will entail:

- researching an appropriate construct, i.e. a definition of academic literacy that can be operationalised, before deciding on test task types and drawing up test item specifications that are aligned with this construct;
- the development of test items within the various task types;
- piloting these items on selected groups of postgraduate students;
- after item selection, preparing a final draft of a first version of such a test.

Intervention design

We will use both our experience with the current course (EOT 300) for post-graduate students, as well as the diagnostic results of the test of academic literacy levels that we will produce, to design an appropriate intervention for students who are identified as being most at risk, in order to enable such students to develop their academic literacy to a desirable level in the shortest possible time. The design of this intervention will comprise:

- the identification of a set of appropriate design principles for an academic literacy development course;
- the investigation of the most appropriate task types to be included in such a course;
- a first draft of materials that would form the basis of such a course.

It is our intention to produce both the test and the intervention on the basis of current research internationally, and within our department. After setting out the milestones and deliverables of the initiative, we return below to the departmental expertise in this field.

Objectives and project schedule

The following objectives have been set: to

- complete the final draft of a first version of the test of academic literacy levels for postgraduate students before September 2007;

- finalise the development of materials for a generic postgraduate academic literacy development course before December 2007.

The eventual administration of the test, as well as the implementation of the materials designed for the intervention, falls outside of this proposal, but these are envisaged for January 2008.

The following schedule sets out the steps to be taken to achieve this:
The two main deliverables of the project will therefore be the test and the course materials.

Target date	Action
30 November 2006	1. Set up project team. 2. Recruit master's students for researching aspects of the test design and development. 3. Finalize construct.
31 January 2007	4. Decide on task types. 5. Draw up item specifications.
31 March 2007	6. Complete development of test items.
30 April 2007	7. Complete first round of trials (pilot test).
31 May 2007	8. Complete second round of piloting. 9. Decide on course design principles.
30 June 2007	10. Finalise draft of first version of test. 11. Decide on task types for intervention.
31 October 2007	12. Write up results of project so far; make substantial advances with completion of master's dissertations involved.
30 November 2007	13. Complete first draft of materials for course.
End 2007	14. Produce project report.

Though the following will fall outside of the time frame of the above schedule, the project will also have as its outcomes:

- two master's dissertations within the field of applied linguistics, focussing specifically on test design and construction;
- four academic articles in accredited journals that are directly related to these studies;

- one local conference presentation (to either SAALA or SAALT), and two international conference presentations at the 2008 AILA conference in Essen, Germany;

- at least one presentation on the results of the project will be made to our partners at the Radboud University of Nijmegen in the Netherlands.

The synergy and networking effects will probably be wider, so we are confident that the above will turn out to be a conservative estimate of the outcomes of this initiative.

Accountability and expertise

The project leader will be Prof. Albert Weideman, director of the Unit for Academic Literacy of the University of Pretoria, who will also take final accountability for the completion of the project. The project official will be Mr Gustav Butler, currently a lecturer within the same unit, and a doctoral student of Prof. Weideman.

The plan provides for the recruitment of two master's students, who will be involved in investigating and writing up the development of the envisaged test.

The initiative will also directly involve one of our research associates, Prof. Frans van der Slik of the Radboud University of Nijmegen, and indirectly a project team at that university who are also currently working on the development of a test of academic literacy for postgraduate students.

Prof. Weideman is an NRF rated researcher, who is a member of the team currently drafting the National Benchmark Test of Academic Literacy under the auspices of HESA. He leads the development of the undergraduate *Test of Academic Literacy Levels* (TALL; TAG in Afrikaans), which is done jointly with the Universities of Northwest and Stellenbosch, and is administered annually to more than 20 000 first year students at various South African universities. He is also involved with the team that, under the Alternative Admissions Research Project (AARP) at the University of Cape Town, develops the *Placement Test in English for Educational Purposes* (PTEEP). He has published widely on this part of his experience, as well as on course design. He is the author of several textbooks, along with a number of coursebooks. Here is a selection of some of his relevant publications:

1985a. *Making certain: a course for advanced learners of English*. Bloemfontein: Patmos.

b.- with G.J. van Jaarsveld. *Doelgerigte Afrikaans; St. 8 & 9: kommunikatiewe oefeninge, aktiwiteite, rolspeletjies en strategieë*. Bloemfontein: Patmos.

1991 - with M. Rousseau. *Starting English: a first course for children*. Johannesburg: Heinemann-Centaur in association with English Language Methods and Programmes (ELMAP).

2002 *Designing language teaching: on becoming a reflective professional*. Pretoria: BE at UP. ISBN 1-86854-436-2. 109 p.

2003 a. *Academic literacy: prepare to learn*. Van Schaik, Pretoria. ISBN 0 627 02541 2.

b. Justifying course and task design in language teaching. *Acta academica* 35(3): 26-48.

c. Assessing and developing academic literacy. *Per linguam* 19 (1 & 2): 55-65.

2004a.- with Tobie van Dyk. Switching constructs: on the selection of an appropriate blueprint for academic literacy assessment. SAALT *Journal for language teaching* 38 (1): 1-13.

b.- with Tobie van Dyk. Finding the right measure: from blueprint to specification to item type. SAALT *Journal for language teaching*. 38 (1): 15-24.

2005 - with Frans van der Slik. The refinement of a test of academic literacy. *Per linguam* 21 (1):23-35.

2006 (ed.) with B. Smeija. *Empowerment through language and education: Cases and case studies from North America, Europe, Africa and Japan*. Volume 65 of the series *Duisburg papers on research in language and culture*. Frankfurt am Main: Peter Lang. ISBN 3-631-55088-X.

2006a. Transparency and accountability in applied linguistics. *Southern African linguistics and applied language studies* 24(1): 71-86.

b. Assessing academic literacy in a task-based approach. Forthcoming in *Language matters* 37.

c. with F. van der Slik. The stability of test design: Measuring difference in performance across several administrations of a test of academic literacy. To be submitted to *Acta academica*.

d. with F. van der Slik. Measures of improvement in academic literacy. To be submitted to SAALT *Journal for language teaching*.

Mr Gustav Butler is currently in the process of completing his doctoral thesis, entitled *A framework for course design in academic writing for tertiary education*. This is directly related to the work envisaged on this project. He has published the following articles in accredited journals that are relevant to this project:

1996. Alternative proposals to second language syllabus design. *Tutor* 4 (2): 94-116.

1999. Collaborative language teaching: an alternative strategy for implementing ESP courses in technikon education. *Tutor* 7 (1): 1-19.

2000 - with A.J. van Rooy. Functional grammar and the teaching of grammar in outcomes-based education. SAALT *Journal for language teaching* 34 (2): 194-209.

2004 - with T. van Dyk. An academic English language intervention for first year engineering students. *Southern African linguistics and applied language studies* 22 (1 & 2): 1-8.

Monitoring, evaluation and dissemination of results

The project will be evaluated in terms of the outcomes envisaged above, and by the ability of the project team to deliver these on time.

The dissemination of results will take place as envisaged above under "Objectives and project schedule".

Budget summary

A detailed cash flow budget, together with budget working papers setting out the rationale for each of the lines of expenditure envisaged, is available on request. The following is a summary of this budget:

Account no		Rand
	Co-ordination and administration	
4476	*Project committee meetings*	1640
4511	*Consultation services and co-ordination*	3430
4652	*Computer programmes*	800
4322	*Photocopying*	1800
4384	*Postage & courier services*	900
4420	*Stationery*	1200
4426	*Telephone & fax*	900
	Test development	
4476	*Recruitment of master's students*	1250
4686	*Bursaries for 2 master's students for 2 years of study*	28000
4322	*Printing of first pilot test*	750
4322	*Printing of second pilot test*	750
4322	*Printing of final draft plus front page*	1050
	Intervention design	
5579	*Printing draft materials*	11625
	Dissemination	
4408	*Conference fees*	1750
4408	*Travel*	6810
5583	*Accommodation and subsistence*	6400
	TOTAL	**69055**

APPENDIX B TALPS MARKING RUBRIC (SECTION 8)

Content and organisation		**Poor**	**Average**	**Good**
Introduction (5)	Statement of issue – angle to be argued	No clear statement of issue; no point of view to be argued; abrupt or no introduction	States issue and point of view weakly; not clear what relevance is	Clearly states issue and point of view, explains relevance and importance
	Framing of reader expectations	No or little interest in explaining clearly what will follow, or in guiding reader	Attempts unsuccessfully to frame reader's expectations of what will follow	Clearly sets out what is to follow, providing a frame for what reader can expect
Body (argument) (5)	Nature of problem/issue	No or little discussion of the nature of problem/issue, or why it is necessary to deal with it	Unsuccessfully attempts to discuss nature of problem/issue and its importance in South Africa	Clear discussion of nature of problem/issue, and necessity of addressing it in South Africa
	Discussion of pros and cons	Gives no or little indication that there is more than one side to an argument	Attempts to provide both pros and cons, but does so unconvincingly	Provides a comprehensive discussion of possible pros and cons
	Argue convincingly for specific point of view	Argumentation is weak, one-sided, unconvincing	Argument deals with some of the important issues, but not in any convincing way	Strong, balanced argumentation that leaves the reader convinced of point of view
Conclusion (5)	Emphasising again the point of view advanced – link with introduction	No connection between the issue/thesis introduced in the introduction and what is said in conclusion	Attempts to restate the issue/thesis, but does so unconvincingly	Clearly emphasises the thesis again without making it a word by word repetition of the introduction
	Clearly states again the most important issues	No attempt to highlight again the most important issues in the text	Attempts to again include the most important issues, but does so in an unconvincing and incomplete manner	Clearly emphasises the main issues again in a structured and non-repetitive manner (exact repetition of the sentences used in body)
Language and style				
Technical aspects and language (5)				
	Academic style and referencing	No or little acknowledgement of authorities, weak structure, interrupted flow of argument	Argument patchy in its logic and structure; some acknowledgement of authority, but inadequate	Authorities used appropriately acknowledged, well-structured argument, logical flow
	Grammar and spelling	The number of grammatical and spelling errors seriously interferes with the meaning	Contains some typical errors that could easily have been eliminated	Primarily error-free and fluent

[**20**]

	Poor	**Average**	**Good**
5	0 – 1	2 – 3	4 – 5
10	0 – 4	5 – 6	7 – 10

APPENDIX C THE TALPS HOME PAGE

ICELDA
Inter-institutional Centre for Language Development and Assessment

TALPS HOME PAGE

Home | Feedback | Search | Contents | Links | FAQ

START TEST

Click here to download the TALPS brochure

About ICELDA

What is TALPS ?

Student Infomation

Supervisor and Test Designer Information

Interpreting the Results of the Test

The Intervention Program

Sample Test

Research

Meet the Test Designers

Welcome to the TALPS webpage. The webpage has been designed to give you as much information as is possible about the test, whether you are a postgraduate student, a supervisor of postgraduate students, a test designer, a researcher or someone simply interested in language testing.

If you are a **student** then you will find that the website has been designed to be as interactive as possible. We have tried to provide you with as much help as possible. We have also tried to anticipate the questions you may have but do not hesitate to **contact us** should there be other questions, suggestions or comments.

A Sample Test has been provided to give you an idea of what to expect in the test. We have also provided you with links to the following:

Click here to test your reading speed

Click here for the academic vocabulary exercises

Click here to access Coxhead's Academic Wordlist

Click here to access the ICELDA website

Click here to access a sample of the academic literacy test

Click here to access sample questions on each section of the test.

If you are a supervisor of postgraduate students, a test designer, a researcher or someone simply interested in language testing, we have tried to provide you with the kind of information you will be interested in: The need for the test, the subtests in TALPS, as well as a validation of the test. The **Research link** provides you with a detailed reading list of articles published on TALL and TALPS should you be interested.

We are interested in your feedback. If you have any comments, questions or suggestions we would like to hear from you. Click here to access our **Comment Box** available on the **Feedback Page**.

APPENDIX D The TALPS Brochure

Attachment Appendix D.doc

APPENDIX E STANDARD PROCEDURES FOR THE ADMINISTRATION OF TALPS

---

1. Introduction[1]

TALPS is a postgraduate test of academic literacy. It is used to identify at the earliest possible stage students whose academic language proficiency (also called academic literacy) may be an impediment in successfully completing their studies. The test results are generally used to make recommendations regarding the student's placement in suitable support courses. No student can be disadvantaged by completing this test.

TALPS is a reliable, valid and standardised measuring instrument:

A **reliable** test is one that can be taken at any other location, at any other stage, in more or less similar conditions with more or less the same test population, and should deliver more or less similar results. This indicates the internal consistency of the test.

If a test is **valid** it measures what it is supposed to measure (in this case the level of academic literacy). Validity is established on a qualitative and quantitative level by means of empirical research.

A **standardised** test presupposes certain standard criteria that are maintained at a constant level from one test to the next. The criteria dictate standard procedures for compiling and developing the test contents, conducting the test and awarding marks. There is a link between the reliability and the validity of a test and its standardisation. If the standard procedures are not complied with, the reliability and

1. This set of standard procedures derives from the procedures employed to administer the Test of Academic Literacy Levels (TALL/TAG) used by the Universities of Pretoria and Stellenbosch, and Northwest University. Its original author is Tobie van Dyk, head of the Unit for Afrikaans in the University of Stellenbosch..

validity of the test is influenced, resulting in possible discrimination against certain students.

2. Administering the test

Since TALPS is a standardised test, it is necessary that you should thoroughly prepare yourself beforehand. Here follows an explanation of the test procedure and the tasks to be executed.

2.1 Before the test

- Familiarise yourself with the test, any other test material, and the procedure for conducting the test.
- Effect the necessary rearrangement of the test location: rearranging the furniture, adjusting the air-conditioning and lighting, facilities for disabled students, etc.
- Ensure that the following are available in the test location:
    - Sufficient test/examination registration forms
    - Sufficient test books and answer sheets
    - Sufficient pencils, sharpeners and erasers
    - An overhead projector in sound working order

2.2 During the test

Phase 1

- At each entrance to the test venue two invigilators should provide test/examination registration forms and pencils to students entering the test room or hall.

- Students take their seats as with any test/examination: one vacant seat between candidates; all candidates in rows behind one another. If the venue proves to be too small for such a seating arrangement, all seats should be filled – start at the front of the venue to fill all seats. Silence should be maintained.
- The rest of the invigilators should move through the venue and ensure that students have nothing but the pencil and the test/examination registration forms in their possession and that they do not complete the test/examination registration forms.

Phase 2

- Discuss the aim of the test (p.1 of the test book) with the students.
- Explain that the questions in the test are multiple-choice and that these should be answered on the loose answer sheet – only the last section should be answered in the test book.
- Explain that participation in the test cannot disadvantage anyone.
- Inform those taking the test how much time is allowed for completing it.
- Explain where and when the test results are to be announced.
- Announce that no scribbling paper is permitted since there is sufficient space in the test books on which to scribble or make calculations, and that calculators / cell phones are also not allowed and should be put away.

Phase 3

- Complete the test/examination registration forms together with the students by showing an example on the overhead projector or similar device. Invigilators should circulate through the venue to ensure that students complete this correctly.
- Upon completing the test/examination registration forms, students should pass them on to the sides of the row (to the left or the right, depending on the venue's layout).
- Invigilators move from one row to the next and count whether the number of test/examination registration forms corresponds with the number of students sitting in the relevant row. As soon as the test/examination registration forms have been collected and counted, one or two invigilators should count them again to ensure that the total number of persons in the venue coincides with the total number of test/examination registration forms.
- Now arrange the test/examination registration forms in alphabetical order and complete the relevant control form – see Annexure A.
- Announce that the test books will now be handed out, that absolute silence should be maintained and that the test books may not yet be opened. Invigilators now hand out the test books with the inserted loose answer sheets row by row (only the number required for each row).
- Please ensure that not too many test books per row have been handed out accidentally and that no one has been omitted.
- The students may now open their test books on the first page. Remind them that you have already discussed *General information* (p.1 of the test book) with them and that they should now sign their names in the designated space in the test book (also p.1 of the test book).

- Once the signing has been completed, page through the test book and have the students also page through their books to ensure that all pages have been printed and that a loose answer sheet has been inserted in the book. Request the students to remove the loose answer sheet from the test book.
- Start to complete the cover of the test book with the students by showing an example on the overhead projector. Invigilators circulate throughout the venue to ensure that students complete this correctly.
- Using the overhead projector to illustrate an example, the biographical details on the loose answer sheet should then be filled in together with the students. No scribbling is permitted on this loose answer sheet. Invigilators circulate through the venue and ensure that students complete this correctly.

  NB: Please remind students that they should not place the loose answer sheets on top of the coloured cover of the test books and then exert pressure on them, because the colouring of the cover may soil the answer sheet. Invigilators should ensure that students code their dates of birth and student numbers correctly on the answer sheets.

Phase 4

- Remind students that they have to hand in both the test books and the answer sheets after the test and that no calculators/cell phones or scribbling paper are permitted.
- Announce that students have exactly one hour to complete the test. Should they experience any problem in any section/question, they should immediately proceed to the next section/question. They should work fast and accurately. No one may enter or leave the test venue during the test period.

- The test now commences. The time allowed for completing the test thus commences as from this point.
- Invigilators should circulate quietly through the venue and ensure that no irregularities occur. Should there be any irregularities, the tester/chief invigilator must be informed immediately.
- Announce after half an hour that half the allotted time has passed. Announce after an hour that the time has elapsed and that the pencils should now be laid down.

2.3 After conclusion of the test

- After the conclusion of the test period, students should place their answer sheets between the cover and page one of the test books (approximately 2 cm. should protrude from the top of the test book). The test books are then passed to the sides of the row (either to the right or left, depending on the layout of the venue).
- Invigilators move from row to row and count whether the number of test books and the enclosed answer sheets coincide with the number of persons sitting in that row. In the meantime all pencils are also passed to the sides. As soon as the test books and enclosed answer sheets have been collected and counted, the students may leave the venue.
- Invigilators recount the total number of test books with the enclosed answer sheets and check that they agree with the total number of persons in the venue. This number should also coincide with the total number of test/examination registration forms. Place the test books with the enclosed answer sheets in alphabetical order. Only now may the answer sheets be

removed. The alphabetical order of both the test books and the answer sheets should be maintained.

- Invigilators should now place the answer sheets in the relevant container.
- Invigilators should group the test books in parcels of 50 and put elastic bands around each parcel. Ensure that each parcel of test books is clearly marked (1-50, 51-100, 101-150, etc.) and clearly indicate the venue, test time and tester/chief invigilator's name. The tester/chief invigilator signs the relevant form when he/she is satisfied that everything was done in terms of the standard procedure.
- Deal with answer sheets, test books and test/examination registration forms as agreed beforehand with the test coordinator(s).
- Invigilators sharpen the pencils and prepare the venue for the next test session.

APPENDIX F THE COVER PAGE OF TALPS

TALPS

TEST OF ACADEMIC LITERACY FOR POSTGRADUATE STUDENTS

Time: 120 minutes **Marks: 100**

General information

1. This is an academic literacy test for postgraduate students. No part of it may be copied, electronically or otherwise, or distributed or used, without the written consent of the copyright holder.
2. You cannot be disadvantaged in any way by completing this test. It is used to determine your level of academic literacy, so that recommendations can be made about appropriate courses which you should/may enrol for. The results of the test are also utilized for research that is aimed at the progressive improvement of the test.
3. By placing your signature below, you declare that you are fully informed about the above, and give your permission that the information may be used for research without identifying you as an individual.

....

Signature Date

Test instructions

1. Write your name and student number on this test, as well as on the loose answer sheet. You have to hand in both of these documents. Answer all multiple choice questions on the loose answer sheet.
2. Use dark ink or, preferably, a soft pencil, which will allow you to erase and correct errors.
3. More than one answer for the same question, or answers that have been scratched out, are not accepted.
4. Use BLOCK LETTERS to complete the text fields, and use only one letter per block, without touching the sides of the block. Begin on the left. Do not place commas or full stops between letters or words. Leave a block open to indicate a space.
5. Enter your answers in the spaces (circles or blocks) as required – see the specific instructions.
6. Sections and questions are arranged vertically (from top to bottom). Please answer each question in the correct space. If you do not know the answer, leave the space open.
7. Do not fold the answer sheet, crumple it, scratch on it or damage it in any way.
8. No pocket calculators or own paper is allowed.

APPENDIX G THE TALPS QUESTIONNAIRE

Unit for Academic Literacy
University of Pretoria
Researcher: A. Rambiritch
Tel: (012) 420 4834
Cell: 0837819028
E-mail: avasha.rambiritch@up.ac.za

INFORMED CONSENT TO PARTICIPATE IN RESEARCH

Title of research: Transparency, accessibility and accountability as regulative conditions for a postgraduate test of academic literacy

Dear Student

The questionnaire based on the Test of Academic Literacy for Postgraduate Students (TALPS) forms part of a doctoral study based on the development of the test. The study aims to answer questions about the accessibility, transparency and accountability of the test developers and the test. The aim of the questionnaire therefore is to elicit information, comments, questions and reactions about the test from the people affected by its implementation, viz. the test takers.

Participation in this study is completely voluntary and you cannot in any way be disadvantaged by filling out this questionnaire. You are also not required to give us your name or any other personal details. Your personal contribution, however, is crucial to this study. Your responses will help answer questions related to key issues in this study as well as to ensure that future versions of the test will ensure greater transparency and accessibility. Your responses will help to open doors to issues largely ignored in the field of language testing. Your confidentiality regarding the information you have provided is guaranteed. You are also free to withdraw from participation in the study at any time. Any data collected from you will be destroyed should you withdraw. The statistical analyses will be done on the group as a whole and not on individual respondents. On completion of this study the data gathered will be incorporated into the UAL database which consists of ongoing research on academic literacy and language related matters. The data gathered will be stored for a period of 10 years as required by university policy.

Ethical clearance for the study has been obtained from the Research Proposal and Ethics Committee of the Faculty of Humanities at the University of the Free State.

Signature of Participant ---------- Date and place: ----------

Signature of Researcher: ---------- Date and place: ----------

UNIVERSITY OF PRETORIA
Unit for Academic Literacy
Questionnaire for students writing TALPS

Purpose of the questionnaire

The purpose of this questionnaire is to determine if this test is a fair and reliable measure of academic literacy. Being academically literate ensures that you are able to "communicate productively and perceptively through the language that you are required to use for academic purposes" (Weideman, 2007: x). In order to do this you should be able to do the following:

- Understand a range of academic vocabulary in context;
- Interpret the use of metaphor and idiom in academic usage, and perceive connotation, word play and ambiguity;
- 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 different parts of a text hang together;
- Interpret different kinds of text type (genre), and have a sensitivity for the meaning they convey, as well as the audience they are aimed at;
- Interpret, use and produce information presented in graphic or visual format;
- Distinguish between essential and non-essential information, fact and opinion, propositions and arguments, cause and effect, and classify, catergorise and handle data that make comparisons;
- See sequence and order, and 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;
- 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;
- 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.

(Weideman, 2007 *Academic Literacy: Prepare to Learn*)

Aim of this research

This questionnaire forms part of a doctoral thesis based on the development of the Test of Academic Literacy for Postgraduate Students (TALPS). The study aims to answer questions about the accessibility, transparency and accountability of the test developers and the test. The aim of the questionnaire therefore is to elicit information, comments, questions and reactions from testees about the test. You cannot in any way be disadvantaged by filling out this questionnaire. You are not required to give us your name or any other personal information.

How to fill out this questionnaire

Place a cross in the column containing the number that reflects your opinion most accurately.

QUESTIONS

	1 COMPLETELY DISAGREE	**2** DISAGREE	**3** NEUTRAL	**4** AGREE	**5** COMPLETELY AGREE
1. I think that it is important to know whether my academic literacy skills will affect my academic performance.					
2. I think that a programme designed to improve my academic literacy skills would be useful.					
3. I think it would be unfair to use the results of this test to deny students access into a desired programme.					
4. I think 'academic literacy' is more than or different from general language ability.					
5. I think that 'academic literacy' is more or less the same as general language ability.					
6. If one is good at languages, one should have no problems coping with academic language.					
7. Being good at languages is no guarantee of being successful in using academic language.					
8. I was well prepared for the test.					
9. I am aware of the purpose of the test.					
10. I understand what is meant by the score I receive for the test.					

	1 COMPLETELY DISAGREE	**2** DISAGREE	**3** NEUTRAL	**4** AGREE	**5** COMPLETELY AGREE
11. I understood all of the instructions.					
12. I was given enough time to complete the test.					
13. I understood all the questions.					
14. I was able to finish the test in time.					
15. I have a positive attitude to tests.					
16. I always understand what the results of a test mean will mean for me.					
17. Tests are sometimes unfair.					
18. Test-takers have little or no rights.					
19. If this test shows me to be at risk I think that people will see me as a failure.					
20. I do not believe that a test is a good measure of my performance.					
21. A poor performance on a test can lead to detrimental consequences.					
22. I do not think that I can in any way be disadvantaged by taking this test.					
23. I think that one needs to prepare specifically for all tests one has to write.					
24. I think that it is not always possible or desirable to prepare for all tests beforehand.					
25. Now that I have written the test, I think that the best preparation may have been attention to the way I have been using language in my own field all along.					

	1 COMPLETELY DISAGREE	**2** DISAGREE	**3** NEUTRAL	**4** AGREE	**5** COMPLETELY AGREE
26. I think that if the theme of the test was related to my field of study, I would have done much better.					
27. I would have liked the test to have a theme from my own field of study.					
28. I do not think that it is important when measuring academic literacy, to work with a theme from my own field.					
29. I think that if the test shows me to have risk associated with language, I would feel very disappointed.					
30. If I am shown to be at risk, I would gladly do a course to improve my language. I cannot risk extending my studies.					
31. I am familiar with the content of the course I may have to take if the result of the test shows me to be at risk.					

THE FOLLOWING QUESTIONS ALLOW YOU TO EXPRESS YOUR OPINION AND GIVE BRIEF REASONS

1. What is your understanding of the concept 'Academic literacy'?

2. How do you think you could have best prepared for this test?

3. Do you think a test of this nature is at all necessary? Explain.

4. Would you feel stigmatised in any way if the results of the test show you to be at risk? Explain briefly.

5. If you were shown to be at risk, how will you feel about taking a compulsory academic writing course?

THANK YOU FOR YOUR TIME