The design, refinement and reception of a test of academic literacy for postgraduate students

Colleen Lynne du Plessis

The design, refinement and reception of a test of academic literacy for postgraduate students

Colleen Lynne du Plessis

A thesis submitted to meet the requirements for the degree Magister Artium (Language Studies) in the Faculty of the Humanities (Department of English) of the University of the Free State.

January 2012

Supervisor: Prof. A.J. Weideman

Acknowledgements

This study was greatly facilitated by the professional input and advice of Prof. A.J. Weideman whose knowledge of the field and enthusiasm were always a source of great inspiration. I also wish to thank both the Council of the University of the Free State and the Inter-institutional Centre for Language Development and Assessment (ICELDA) for the financial assistance granted me in the form of bursaries. My heartfelt gratitude is further due to the many family members and friends who encouraged me to pursue my studies.

"Act justly, love mercy, walk humbly"
Micah 6:8

Declaration

I herewith declare that this thesis, which is being submitted to meet the requirements for the qualification Magister Artium (Language Studies) in the Faculty of the Humanities (Department of English) of the University of the Free State, is my own independent work and that I have not previously submitted the same work for a qualification at another university. I agree to cede all rights of copy to the University of the Free State.

Table of contents

Chapter 3

The design of a test of academic literacy for postgraduate students

Chapter 4

The refinement phase of the test

Chapter 5

The reception of a test of academic literacy for postgraduate students

List of tables

List of figures

Chapter 1

The assessment of academic literacy at postgraduate level

1.1 The need for a test of academic literacy for postgraduate students

A number of studies have indicated that the academic literacy levels of students at tertiary institutions in South Africa are lower than required for academic success, largely as a result of the prevailing conditions and standards in South African schools (Van Dyk & Weideman 2004a; Van der Slik & Weideman 2007, Bhorat & Oosthuizen 2008). Whether the literacy levels are actually showing a decline, or whether they have been sub-standard for a long time, or may be expected to increase, is not the focus of this study. The immediate issue to be addressed is how to respond to the evidence of inadequate success as manifest in the form of through-put rates of university students (DIRAP 2011, Annexure A). The institution of language testing at South African universities can be seen as one plausible step towards identifying students at risk and assisting them to gain an awareness of their current academic literacy levels and English language proficiency. On the basis of such a system of measurement certain inferences can be made that may be helpful both to the individual student in terms of addressing a language-related ability, and to the tertiary institution in respect of resource allocation and enrolment planning.

Currently universities are using the National Benchmark Tests (NBTs) for access purposes and as a means of assessing the literacy levels of first-entry students. In addition hereto, some institutions are employing the Test of Academic Literacy Levels (TALL), initially developed at the University of Pretoria's Unit for Academic Literacy, but used by the four partnering institutions that are collaborating as the Inter-Institutional Centre for Language Development and Assessment (ICELDA), viz. the Universities of Pretoria, Free State and Stellenbosch and North-West University. This test, or close derivatives of it, is also being employed more widely now at institutions in Namibia, Singapore and Vietnam. Whereas some tertiary institutions use these kinds of literacy tests as a predictor of potential or as a gatekeeper in terms of which only certain students gain access to tertiary or postgraduate study ('high-stakes testing'), others rely on such assessment measures to determine which students are in need of supportive interventions to increase their chances of success, an approach referred to by some test developers (Bachman & Purpura 2010: 456-461) as the door-opener function. Either way, the institutionalized system of language and literacy assessment at tertiary institutions is already well entrenched and can be expected to continue for the foreseeable future in the light of the low academic literacy levels of students.

While it may be accepted that a number of first-year students will display inadequate academic literacy levels for the purposes of studying at an institution of higher education, it is a disconcerting prospect that students may be able to

graduate at a tertiary institution with low levels of academic literacy. Even more disturbing is the possibility that students may be admitted to postgraduate study without having attained an adequate level of academic literacy during their undergraduate course work. The proposed study identifies the need for postgraduate literacy assessment in terms of both the door-opener and gatekeeper functions and sees the design of an appropriate measurement instrument in the form of an academic literacy test as a useful tool for identifying students at risk, as well as channelling postgraduate student applications, provided the inferences drawn from those tests do not serve as the sole basis for access to postgraduate study.

Students need to have attained a certain minimum level of academic literacy in order to stand a chance of successfully completing a postgraduate field of study (Weideman 2003b). Students who fail to meet this minimum standard should most likely not be admitted to postgraduate study until they have undergone a number of necessary interventions to increase their literacy levels, following which they may be re-assessed. On the other hand, students who achieve the required minimum level of academic literacy, but do not receive a sufficiently high score, may possibly be admitted to postgraduate study on the condition that they register for specially designed support modules to strengthen their academic literacy on the basis of the general weaknesses identified through the test. This should play a part in reducing the number of postgraduate students who fail to complete their studies.

Any potentially helpful measures should be welcomed, considering that the termination of postgraduate study on the part of a student can create a predicament for the academic department concerned and have a negative impact on the institution's government subsidy, in addition to having an adverse effect on the student's self-actualization.

Apart from these more obvious intentions for the administration of a language test, there are a number of further reasons why tests such as the TALL are increasingly being employed. These tests can also be used to benefit the greater society. For example, national benchmark tests in South Africa are being used not only to determine literacy trends amongst new entrant students, but also to influence education policy, resource allocation and curriculation (see www.nbt.ac.za/cms/). Another use of language assessment is to certify that an individual has attained a proficient level of knowledge and skills for a specific purpose. A high test score reflects that both the language has been mastered and the ability to communicate with competence (Bachman & Purpura 2010: 459). Such information is of use to employment agencies where particular language skills are required for various professions.

In view of the somewhat unrealistic expectations that may exist about what may be accomplished through language testing, it should be stated unequivocally from the outset that academic literacy tests are not psychometric tests designed to test

potential, but rather instruments that reflect the current literacy level of a candidate. For this reason, literacy tests are never to be used in isolation, but preferably in conjunction with a number of other indicators of academic success, as advocated by authorities such as Bachman and Palmer (1996), Messick (1980) and Weideman (2009a). Nonetheless, users of language tests may still have unrealistic expectations as to what language tests are able to do. One of the mistaken ideas about language testing mentioned by Bachman and Palmer (1996: 3-4) is that there is a magical formula for designing a model language test. It is to be doubted that any single 'best' test exists for a specific language testing situation. Those who are of the opinion that any single model exists, fail to take into account that the processes involved in language learning and mastery of language are not the same for all learners, situations and purposes.

1.2 The evolution of language testing

Although language testing has been around for a long time, its history as a "theoretically founded and self-reflexive institutional practice" (McNamara & Roever 2006: 1) is of a brief nature. Language testing was only institutionalized in the second part of the twentieth century on an interdisciplinary basis incorporating the fields of psychometrics and applied linguistics. Some see this development simply as psychometrics prescribing the measurement rules, and the subsequent addition of language to existing forms of testing. However, as McNamara points

out, “a psychometrically good test is not necessarily a socially good test” (McNamara & Roever 2006: 2), and it would seem that since language is inextricably linked to a social context, the social dimension of testing may be expected to be more marked in language testing than it is within the ambit of general cognitive ability assessment. McNamara believes that the cognitive bias in psychometrics has actually hindered some aspects of language testing and that more research is needed on the social effects of testing.

As the communicative approach to language teaching gained ground in the nineteen eighties, largely in reaction to the prevailing modernist and technocratically inspired audio-lingual approach, the way language testing was conducted also came under review. Test developers realized the need to relate performance in a language test to the use of that language in a specific setting (Bachman & Palmer 1996). It then follows that if language performance is the outcome of language ability, this ability that is to be assessed needs to be defined before any test can be constructed. During the audio-lingual era of language teaching a restrictive view of language was adopted by most test developers in terms of which language was considered to be a combination of sound, form and meaning (phonology, morphology, syntax and semantics), necessitating the assessment of four separate language skills: listening, speaking, reading and writing. This restricted view of language came under criticism for failing to take into account the communicative and expressive role of language as a social

instrument used to mediate and negotiate interaction in a variety of specific contexts (Van Dyk & Weideman 2004a). Attempting to define language ability in terms of the four mentioned skills was subsequently considered inadequate and a negation of the role and use of language to carry out specific tasks (Bachman & Palmer 1996, Blanton 1994, Van Dyk & Weideman 2004a).

The above developments are relevant as they influenced the format that institutionalized literacy testing would adopt in South Africa. When low academic language proficiency levels were recognized as one of the main reasons behind the lack of academic success of many non-native English speakers at South African universities, the Unit for Language Skills Development (ULSD) was established at the University of Pretoria (UP) in 1999 and given the task of developing the academic language proficiency of students at risk. From 2000 onwards all students at the university were required to be certified language proficient prior to obtaining a degree (Van Dyk & Weideman 2004a). The English Literacy Skills Assessment for Tertiary Education (ELSA PLUS), which was designed by the Hough and Horne consultancy, was initially employed to assess students' literacy levels. This test version was an adaptation of an industrial and commercial test which was refined in collaboration with the ULSD for use at the university. It proved to be problematic precisely as a result of its emphasis on a restrictive, limited view of language, as sound, form and meaning, that would lead to the assessment of the separate 'skills' of listening, reading and writing. Not only this

limiting view of language, but also practical considerations further necessitated a switch to a different construct (Van Dyk & Weideman 2004a). When the test designed by Yeld and her associates at the University of Cape Town in 2000 as part of the Alternative Admissions Research Project (AARP) – the precursor of the National Benchmark Tests (NBTs) – also proved to be inappropriate and unavailable, mainly for practical reasons since it included writing tasks that required sufficient time for marking, the work of Blanton (1994), Bachman and Palmer (1996), along with that done by Yeld and her associates, was used by Weideman (Van Dyk & Weideman 2004a) to redefine a blueprint for assessing academic literacy. The result was the adoption of the innovative placement test of academic literacy at the UP referred to as the Test of Academic Literacy Levels (TALL). This test has since been written annually by tens of thousands of students at the four participating ICELDA institutions.

When the need became apparent for the assessment of academic literacy levels at postgraduate level, the existing format of the TALL was used as the basis for developing a test for use at this more advanced level. The latter test is referred to as the Test of Academic Literacy for Postgraduate Students (TALPS) and is already being employed under the auspices of ICELDA, but there is an urgent need to design more such tests and to research the effects of different task types. A single version of a test not only is a security risk, but also limits cross-comparisons that might be useful in refining the test design.

1.3 Research methodology

This study falls within the scope of the discipline of applied linguistics. Different paradigms of applied linguistics will briefly be examined by means of a literature study to show how these have been relevant to the field of language teaching and learning, from which literacy assessment derives. Six identifiable generations of applied linguistics will be discussed and a seventh introduced. Of central concern will be the move away from the positivist and prescriptive approaches to language teaching and assessment that relied on a form of presumed scientific proof. The inadequacies of such an approach will be discussed critically. At the same time it will also become apparent why applied linguistics cannot simply be regarded as an extension of linguistic theory or as a mediator between linguistics and other disciplines, but as a full-fledged discipline in its own right, in which the technical design element is at the forefront (Weideman 1987, 2007, 2009b).

The constitutive and regulative dimensions of applied linguistics as a discipline of design will be dealt with as the necessary foundation for any applied linguistic practice, including the design of an instrument to assess academic literacy. Although language tests have conventionally been required to show validity and reliability, in terms of more contemporary thinking such tests must also possess what is referred to as consequential validity, a notion that refers to the impact and

power dimension of tests (Shohamy 2006). Moreover, since language testing is not without controversy and has in the past fallen prey to abusive power relations, it is essential that any literacy test should have a high face validity[1] and that it should only be employed for the purpose for which it has been designed. Apart from designing an appropriate alternative test version, and presenting a theoretical justification for this design, this study will have the further objective of examining the reception of this kind of test amongst its test takers, in order to ensure that it is considered to be fair and credible. If literacy tests are to be consistent and theoretically justifiable, they should incorporate a multiplicity of evidence (Bachman & Palmer 1996, McNamara & Roever 2006, Van Dyk 2010, Weideman 2009a) to back up their validation. Each of the constitutive and regulative dimensions of language tests will be examined, including factors such as test acceptability, utility, accountability and transparency. Based on the above framework, it is evident that language tests as applied linguistic artifacts will have both a leading technical mode (including a set of regulative dimensions) and an analytical founding dimension (as well as a number of constitutive elements).

The central part of the study will involve the design of a test of academic literacy for postgraduate students based on the current versions of the TALL and in particular the TALPS, which already have a well-established test construct (Butler 2009). The various phases involved in the design of a test will be covered,

1. Face validity in brief refers to the perceived credibility of a test and will be discussed in more detail later.

including piloting and refinement. In order to proceed with the design of the test, a literature study will be carried out to give a theoretical articulation of academic literacy, since this constitutes a crucial aspect of construct validity. Various definitions will be considered, with particular attention being given to the definition of functional academic literacy provided by Weideman (2003a). The identified ability will be reflected in the blueprint for the test construct and will be further specified in the task types selected for inclusion in the test (such as vocabulary knowledge exercises, cloze procedure, text editing, interpretation of visual and graphic information and writing tasks). Task types will be closely aligned with the actual language tasks that postgraduate students are required to perform and will be evaluated in terms of their ability to be productive, based on a quantitative system of measurement and the application of appropriate statistical procedures (Bachman 2004, Paltridge & Phakiti 2010). Moreover, practical and logical constraints pertaining to the administration of the test and subsequent process of marking individual test papers will also be taken into consideration, since these may play a role in determining the format of the selected task types.

In addition, the research will include a reception study in the form of a survey conducted amongst a cohort of postgraduate students at the University of the Free State. Survey questionnaires were distributed to both prospective and current postgraduate students who wrote the TALPS, with the objective of determining their perceptions of the test. It is envisaged that the students' comments could

possibly be of assistance in future test administration. The main objective of the survey questionnaires is to assess the face validity of the test. Although the survey respondents are unlikely to be well versed in the different technical aspects of academic literacy, high face validity is considered essential for the future employment of a potentially high-stakes assessment instrument.

1.4 Value of the research

The investigation will conclude with a summary of the main findings of the research and the identification of necessary areas of further study, including the development of subject-specific literacy tests and the availability of academic literacy tests in languages other than English. The alignment of academic literacy course modules with literacy assessment constitutes a further challenging area of investigation that will finally be focused on.

The main value of the study is likely to be found in the demonstration of the viability of designing an assessment instrument that can serve as a useful tool for identifying students at risk of not completing their studies, and as an initial step towards addressing the problem of inadequate through-put rates at postgraduate level.

Chapter 2

Academic literacy assessment as a sub-discipline of applied linguistics

2.1 Applied linguistics as a discipline of design

This study falls within the domain of the discipline known as Applied Linguistics. Delineating the field of reference of this discipline, however, continues to be an elusive and contentious matter. At the one extreme scholars have argued the modernist case for a theoretical continuity in terms of which applied linguistics is regarded as a subdivision of linguistics. Towards the middle of the spectrum others have reconceptualized applied linguistics as a problem-solving enterprise and mediator between linguistics and other disciplines. The resultant contradiction that applied linguistics can both constitute an inherent part of linguistics, while at the same time fall on the continuum between linguistics and other disciplines, has yielded an alternative, postmodernist view, which lies towards the opposite end of the spectrum. It is a view that emancipates applied linguistics from the control of linguistic theory and acknowledges it as a discipline in its own right (Hall, Smith & Wicaksono 2011, Sealey & Carter 2004,[2] Weideman 2007). All of these views, however, have had a significant role to play in attempting to define applied

2. Sealey and Carter regard applied linguistics as a social science and see language use as a form of social practice, hence their view that social science disciplines are better able to describe certain aspects of linguistic behaviour than are those disciplines which are concerned primarily with language. They redefine applied linguistics as "problem-based researching into communication-mediated issues in social life" (2004: 17).

linguistics and in endeavouring to provide a theoretical foundation for language solutions to specific problems, particularly within the context of language acquisition and education.

Although applied linguistics may first have gained recognition in language teaching and learning, an understanding of its nature cannot be restricted to the teaching and learning of language. The discipline covers a much broader scope and multiplicity of fields of language practice, such as translation science, language planning and language management to mention but a few (see Crystal 1987: 412). Nonetheless, a number of scholars have adopted the more structuralist view that applied linguistics is a reflection or application of linguistic theory in language teaching, or linguistics applied. Corder (1973: 31) sums up this perspective with a statement that a comprehensive plan for a language-teaching operation "must be expressed in 'linguistic' linguistic terms – lists of grammatical structures and vocabulary …" and that the linguistic approach determines "how we describe what we are to teach". Crystal (1987: 412) himself describes applied linguistics as the "application of linguistic theories, methods, and findings to the elucidation of language problems that have arisen in other domains".

The tradition of applied linguistics as both an intra-disciplinary and inter-disciplinary field, still firmly entrenched in linguistics, can be seen in further definitions such as the following:

> …It would … make … sense to regard applied linguistics as just that part of linguistics which, in given situations, turns out to have applications in some other field (Kaplan 1980: 3).

> …To see linguistics as comprising no more than phonetics, phonology, morphology, syntax and semantics, plus a statement concerning their foundations and interrelationships, plus a statement concerning their relationship to the individual and to society: this will suffice to provide a perspective for the notion of 'application' (Crystal 1981: 2).

A more recent view is that of Hall, Smith and Wicaksono (2011: 15) who employ the wording "autonomous applied linguistics" to emphasize that applied linguistics is not limited to any application of the findings of general linguistics. They agree with scholars such as Brumfit and Weideman that the scope and methodology of the subject field differ and that the real issue is the investigation of solutions to real-world problems in which language features as a central issue. Hall *et al.* define autonomous applied linguistics as a "discipline concerned with the role language and languages play in perceived problems of communication, social identity, education, health, economics, politics and justice, *and* in the development of ways to remediate or resolve these problems" (2011: 15). As such applied linguistics draws on multiple theories and methodologies from other fields, rendering the notion of 'linguistics applied' redundant.

The attempt to define applied linguistics is relevant to the field of language assessment and testing, since the latter is a reflection of a theoretical belief as to how language is learned or acquired, and, in the case of the present research study,

more specifically how academic literacy is developed and can be assessed. A review of the different views on and traditions of applied linguistics and their relation to language teaching and learning is thus necessary to understand the impact of each on language assessment.

2.1.1 Traditions of applied linguistics

Weideman (2009b: 62) provides a concise summary of the successive traditions of applied linguistics, which is presented in table 2.1.

Paradigm/Tradition		**Characterized by**
(1)	Linguistic/behaviourist	"scientific" approach
(2)	Linguistic "extended paradigm model"	language is a social phenomenon
(3)	Multi-disciplinary model	attention not only to language, but also to learning theory and pedagogy
(4)	Second language acquisition research	experimental research into how languages are learned
(5)	Constructivism	knowledge of a new language is interactively constructed
(6)	Postmodernism	political relations in teaching; multiplicity of perspectives
(7)	A dynamic/complex systems approach	language emergence organic and non-linear, through dynamic adaptation

Table 2.1: Seven successive traditions within applied linguistics

From the above it is apparent that the first attempts to delimit applied linguistics were largely influenced by prevailing Western thinking that sound knowledge was to be found in science and that technology was a form of applied science. Such essentially 'technocratic' thinking led to arguments that technical-scientific methods should be used to analyse man and society, which obviously included the lingual reality (Weideman 1987). Not surprisingly the first tradition in applied linguistics relied heavily on some form of purported scientific proof. This can be seen in the application of behaviourist theory to language learning that was characteristic of the middle of the previous century. Weideman reinforces the main points of critique against the modernist perspective that science provides the only guarantee of an authoritative solution to a language problem by pointing out that scientific analysis itself is not neutral. Not surprisingly the supposed benefit of scientific analysis for applied linguistics has been rejected in postmodernism.

Weideman (1987) comments that evidence of a bi-directional and reciprocal feedback between linguistics and applied linguistics played a role in creating credibility for the intra-disciplinary view that applied linguistics was indeed part of linguistics. Accordingly applied linguistics was seen as the carrying over of linguistic knowledge into language teaching. This view was shattered with the arrival of the theory of transformational-generative grammar when no evidence could be found of a mentalist approach in the prevalent language pedagogy and teaching materials. Scholars faced the predicament of explaining why linguistic

theory was not being reflected in language teaching. Consequently, the notion of applied linguistics as a continuation of linguistic theory started to lose its firm footing. Interestingly, a trace of cognitivism can be seen in some later communicative techniques applied in second language acquisition studies requiring learners to discover grammatical organization for themselves.

The modernist view of applied linguistics forwarded by advocates of linguistic theory was also criticized for its positivist and prescriptive focus on a scientific foundation that emphasized analogy and linguistic distinctions rather than analysis (Weideman 1987). On the matter of the monotonously repetitive audiolingual method of teaching sound systems and sentence patterns, Weideman (2007: 591) states that rather than providing any demonstration of the application of linguistics to the design of a solution to a language problem, "the 'linguistic paradigm' of first generation applied linguistics…..has left us with a language teaching design devoid of proper theoretical justification". Furthermore, the 1970s transformational-generative grammar also failed to acknowledge the instrumental communicative function of language. As a result hereof linguistics started to lose its iron grip on applied linguistics, which came to be seen instead as a mediating discipline. Weideman (2007) believes that although the mediating perspective is problematic, developments such as the above have helped to emancipate applied linguistics from its direct dependency on linguistics as mother discipline. The proposition that applied linguistics fulfils a mediating role remains problematic

since, of necessity, in order for there to be a mediating role, the nature of the two things being mediated needs to be entirely different. Yet, if the one is considered to be part of the other, the implication is that the two are not inherently different. Rather, as Weideman (1987) shows, there is a difference in principle between the two, with applied linguistics operating in a much more specified and contextualized environment, a view shared by Sealey and Carter (2004). The study of language and linguistic concepts can therefore not be equated with the application of language plans as instruments of design to address an identified problem. The two aspects may be related, but applied linguistics cannot simply be seen as a continuation of linguistics, since the latter deals with an analysis of the learning and use of language and the structure of lingual objects, while the former attempts to address a language problem in a particular and complex context through the design of a solution.

The difference in emphasis and the distinguishable design element of applied linguistics can be discerned in the explanation of Widdowson (1984) that the term applied linguistics indicates the use of theoretical studies of language to generate solutions to problems arising in different domains, without assuming that a relevant model of language must of necessity derive from a formal model of linguistic description. The fact that theories started to be developed from work already done within applied linguistics is described by Weideman (1987) as the discipline's coming of age. The point to be noted in the many searches for a

theoretically justifiable basis for applied linguistics, however, is that "in designing solutions to language teaching problems, theory does not lead the way" (Weideman 2007: 594). Widdowson (1984: 8) goes so far as to state that the relevance of linguistics to language teaching cannot be taken for granted and that it is likely that "linguistics, as customarily conceived, may *not* be the most suitable source for a practical teaching model of language". This is obvious in the failure of both behaviourism and cognitivism to provide an enduring theoretical basis for language learning and teaching. The move towards communicative language teaching (CLT) in the nineteen eighties illustrates this point further. Only after the implementation of CLT did research on second language acquisition and constructivism come to provide a theoretical justification for the already designed and applied solution in the language classroom.

Weideman (1987) points out a further definitive distinction between linguistics and applied linguistics by referring to what he terms logico-analytical and technico-analytical analyses. In terms hereof linguistics may be conceived as the insights gained through a theoretical analysis of the lingual mode of experience, whereas applied linguistics should be viewed as those insights obtained through an analysis of a language problem with the purpose of mastering the latter in a technically designed solution. Linguistic knowledge may thus be subsequently used to identify a language problem and so justify a technical design which will provide the solution. The fact that the anticipation of a design is referred to

suggests the dilemma of attempting to provide applied linguistics with a scientific status in terms of which a particular method of language teaching or assessment may be deemed to be scientific and henceforth foolproof and acceptable or credible. By now it should be clear from the complex nature of the subject field that such a notion of scientific status is rather unrealistic. Since science is founded on theory and not absolute truth, as is evident in the evolutions of language teaching methodology and changing philosophical paradigms, the inference can be drawn that science and theory can never be neutral or fixed concepts. Anything being studied will inevitably come under the influence of political, cultural, social and other realities of a changing nature.

The reaction against the notions of absolute truth and scientific discovery, coupled with the increasing consciousness of political power relations, explains the emphasis placed in postmodernism on political and social accountability in relation to language solutions. Though this shift elucidates what alternative conceptualizations of applied linguistics might be entertained, it does not fundamentally alter the disciplinary character of the field. Weideman (2007) states that although postmodernist approaches signal a break with their modernistic predecessors, discontinuity is an impossibility, since the latter continue to define them, albeit negatively.

From the preceding overview it should be evident that each of the different traditions has played a part in helping to define what applied linguistics is or, considering the continuing differences of opinion, what it is not. Complex systems theory draws attention to the fact that aspects of previous schools of thought may re-emerge in later paradigms, whether in a similar or new format. It would thus be a mistake to attempt to base applied linguistics solely within any one particular tradition. Of more significance is the common thread that Weideman (2007) notes in all of the mentioned paradigms – the element of design found in the creative solutions to language-related problems. This seems best to define the nature of applied linguistics for the purposes of the current study, which will thus proceed on the basis of Weideman's (2007: 589) view. Although we may never arrive at a succinct definition of applied linguistics that satisfies all parties, at least the above understanding provides a functional framework within which to operate. Taking into account the diverse aspects raised in the preceding discussion, applied linguistics can then be referred to in very simple and brief terms as the design of theoretically justifiable solutions to complex language-related problems in very specific social and political contexts.

2.1.2 The social and critical turn of language testing

The same change in paradigmatic thinking that can be seen in language education may be evidenced in language assessment, with a definite move away from the

assessment of knowledge and rules of structural grammar (evidenced in the first three traditions of applied linguistics) towards a task-based form of testing in which the communicative function of language within specific contexts is given pre-eminence within an integrative approach (characteristic of the later traditions of applied linguistics; see Truscott 1996, Weideman 1987, 2002). Not surprisingly language testing has incorporated the use of authentic texts and engages learners and students in tasks such as extracting information and interpreting meaning beyond sentence level. The emphasis in academic literacy testing likewise can be seen to fall on critical reading, analytical thinking and persuasive writing, the kind of tasks typically required of postgraduate students.

McNamara (2005) makes it clear that the character of applied linguistics is receiving an increasingly critical focus which is also being reflected in language testing. What he refers to as the "social turn" (p. 775) of language testing is evidenced in new concerns that are being raised about values and consequences of a social nature, along with epistemological debates on the socially embedded nature of knowledge and language. In the same vein, the extent to which language proficiency should continue to be conceptualized with its current individualistic focus is also coming under scrutiny. As a result hereof, the social context of assessment is receiving emphasis with calls for ability to be interpreted in the light of the social values and practices engendered (McNamara 2005: 776), reminiscent

of Sealey and Carter's view that applied linguistics should be regarded as a discipline of the social sciences.

In line with more recent thinking, applied linguistics research methodology in language assessment is no longer being based only on traditional psychometrical procedures, but incorporates, for example, discourse analysis and qualitative research methods. Language performance is no longer viewed solely as projecting individual ability or competence, but rather as being of a collaborative nature. A further indicator of the changing nature of applied linguistics mentioned by McNamara (2005) is Shohamy's introduction of the notion of critical language testing which endeavours to take stock of political and social agendas behind language assessment practice.

2.2 Constitutive and regulative conditions for applied linguistic practice and literacy assessment

The term "assessment" may be employed in different ways across diverse fields of study, but within the ambit of language testing Bachman (2004: 7) describes assessment broadly as "the process of collecting information about a given object of interest according to procedures that are systematic and substantively grounded". 'Object of interest' in this context refers to a particular aspect of language ability, also termed a test construct. As Bachman points out, when an

assessment can be replicated on the basis of explicit and transparent procedures, it may be considered to be systematic in its design and implementation. What is meant by 'substantively grounded' is that firmly accepted language theory must underpin the assessment – an aspect that has a bearing on construct validity – and what is meant by 'systematic' we can return to later, when discussing test validation. Measurement in turn is defined as the process of "quantifying the characteristics of an object of interest according to explicit rules and procedures" (Bachman 2004: 8). The specification of rules and procedures is necessary to link the (unobservable) ability to be measured to the number allocated to the (observable) performance thereof.

The above descriptions of assessment and measurement form part of what is commonly referred to by language testing specialists as some of the components of the process of validation. Various theories have developed in the field, with the main emphasis falling on construct validity. The confidence that may be placed in any language test is considered to be directly proportional to the evidence collected in the process to support the evaluation instrument's validity (Davies *et al.* 1999: 220). The latter refers to the systematic presentation of this evidence as a unity within a multiplicity of arguments setting out the relationship of the test to the definition of the ability being tested (the construct). Three of the main interpretations of validity theory will briefly be discussed in the section that follows. In particular, attention will be devoted to a number of essential criteria

which shall be referred to as the constitutive and regulative conditions for language testing.[3]

2.2.1 The views of Messick and Angoff

Validity has adopted a central place in the work of scholars such as Messick and Angoff, presumably as a result of its fundamental importance to psychometrics. Traditionally in the 1940s and earlier, validity was narrowly regarded as a correlation of a test score with another form of objective measurement of what the test was supposed to measure, often expressed as the square root of test reliability (see Angoff 1988: 20). Only later was it understood that such (objective) validity was part of a process of (subjective) validation that depended on the "interpretations and inferences" (Angoff 1988: 24) drawn from the scores and the decisions resulting from these inferences. In terms hereof the designer of the test and the user thereof both share the responsibility for providing evidence that the testing is valid. Moreover, according to this view, validity extends from the very start of designing a test and continues beyond its administration, unlike traditional validation that was criterion related and product oriented (see Angoff 1988: 25).

3. Testing itself is a very general term used in all subject fields and is considered to refer to a particular procedure to establish the "quality, performance, or reliability of something" (*Concise Oxford English Dictionary*, 2006, 11th edition, p. 1489.

Whereas validity could generally be classified into four types in the 1950s, namely content, predictive, concurrent and construct validity, enabling a test to be shown to be valid on the basis of any of these, this was succeeded by a more unitary view that the first three types are to be found within construct validity (see Angoff 1988: 25). This view is largely propagated by Messick (1980: 1015) who avers that it is the "unifying concept of validity that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships". He defines validity itself as a comprehensive judgment that is evaluative by nature, founded on empirical evidence and theoretical rationales, and related to the adequacy and appropriateness of inferences and actions that are based on test scores. In brief, "validity is an inductive summary of both the adequacy of existing evidence for and the appropriateness of potential consequences of test interpretation and use" (Messick 1988: 33-34). Messick further states that all educational measurement should be construct-referenced, since it is the interpretation of the construct that constitutes the basis for all inferences based on scores and that even if construct-related evidence "may not be the whole of validity, there can be no validity without it" (*ibid.*: 35). Messick's (1988: 20) incorporation of the social dimension of assessment into validity theory is demonstrated in the table below.

	TEST INTERPRETATION	TEST USE
EVIDENTIAL BASIS	Construct validity	Construct validity + relevance/utility
CONSEQUENTIAL BASIS	Value implications	Social consequences

Table 2.2: Messick's facets of validity

McNamara and Roever (2006: 14) further elucidate the above model to illustrate how Messick brings the social context of testing to the fore in table 2.3.

	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?

Table 2.3: How to understand Messick's validity matrix

McNamara and Roever (2006) point out that the relationship between fairness and empirical evidence, as well as social dimension, has never been resolved within the field of language testing and that Messick's insistence to investigate the

overtly social dimension remains controversial. Nonetheless, if language testing research is to move beyond validity theory and truly contribute to a broader discussion of the functions of tests in society, then it must develop "an ongoing critique of itself as a site for the articulation and perpetuation of social relations" (McNamara & Roever 2006: 40).

2.2.2 Bachman and Palmer's notion of test usefulness

Bachman and Palmer (1996) suggest an alternative and more manageable notion to dealing with the essential criteria of language tests, which they term *usefulness.* McNamara (2003) considers their notion of test usefulness to be their most helpful contribution to language testing theory and a replacement of Messick's construct validity. Bachman and Palmer (1996: 9) consider two principles to be of fundamental importance for language test development: performance in a test must correspond to actual language usage in a non-test situation, and the usefulness of a test should be measured in terms of quality control variables such as "reliability, construct validity, authenticity, interactiveness, impact, and practicality" (1996: 9) as shown in the table that follows.

Test usefulness					
Reliability	Construct validity	Authenticity	Interactiveness	Impact	Practicality

Table 2.4: Bachman and Palmer's exposition of test variables

In order to meet the first objective, a conceptual framework needs to be in place which describes the salient features of the test performance and non-test language use. This will enable the identification of suitable texts and task types. Attention also needs to be devoted to the characteristics of the test takers, including their topical knowledge, language ability and "affective schemata" (Bachman & Palmer 1996: 12), as these affect the way the test takers interact with the test tasks. A task that requires test takers to relate topical content to their own topical knowledge can be expected to be more interactive. According to Bachman and Palmer (1996: 26) interactiveness is an essential quality of language test tasks, because it provides the necessary link with construct validity.

The second objective of test usefulness requires further elucidation. Bachman and Palmer (1996) point out that although there may be a measure of tension among the mentioned variables, test usefulness should be seen as a function of the respective attributes since they are interrelated. The overall usefulness of the test should be emphasized rather than the individual qualities that exert an influence on

usefulness. As such the variables should be assessed in terms of the combined effect that they have on a test's usefulness. Furthermore, test usefulness needs to be determined for each particular testing situation. Bachman and Palmer point out that judging the usefulness of a test remains subjective to a large extent, depending on the aspects which the test developer wishes to emphasize. Of the six test qualities identified by Bachman and Palmer (1996: 19), reliability and validity are considered to be the two most essential variables when it comes to justifying using test scores for the purpose of making inferences.

Reliability is referred to by Bachman and Palmer (1996: 19) as "consistency of measurement". This implies that test scores may be deemed to be reliable if they remain consistent from one set of tests and tasks to another. Reliability is thus a function of score consistency between different administrations of tests and tasks. A test taker should thus obtain the same if the same test is administered to the same group of test takers on two separate occasions and settings. Reliability is essential if a test score is to provide any information about the test taker's language ability. Note should nonetheless be taken of the fact that it is impossible to eliminate inconsistencies completely. It should thus be endeavoured to use the test design to minimize the effects of the sources of inconsistency. Bachman and Palmer (1996: 135) consider the purpose for which the test is intended as probably the most important aspect when determining a minimum acceptable level of reliability. For a high-stakes test the minimum acceptable level of reliability

should be set as high as possible (a Cronbach's *alpha* of 0.7 is considered to be suitable for basic testing and research purposes; see Hogan 2007: 149-150). Reliability is harder to achieve when the construct is complex and covers a range of language ability components and topical knowledge.

Reliability is a prerequisite for *construct validity.* In brief this form of validity refers to the extent to which the test "adequately captures the concept in question" (Paltridge & Phakiti 2010), or, stated differently, the extent to which a given score can be interpreted as "an indicator of the ability(ies) or construct(s)" to be measured (Bachman & Palmer 1996: 21). Van Dyk and Weideman (2004b: 17) offer a third understanding of construct validity as the alignment of the definition of the construct (ability) with what the testing instrument actually measures. Construct validity provides the necessary justification for the interpretation and generalization of test scores. Since academic literacy is the construct under consideration, this needs to be assessed with an enriched, open view of language and what is meant by academic language ability, rather than in terms of a mere four skills-based (reading, listening, writing, speaking) restrictive approach (Van Dyk & Weideman 2004a).

Since test designers and users need to justify the validity of the interpretations they make, evidence should be produced that the test scores do reflect a particular area of language ability that is being measured. Whereas the term construct refers to the

definition of an ability that is to be tested, construct validity pertains to the degree to which a given test score can be interpreted as a valid indication of ability with reference to the definition of that ability. Of consideration here is the necessity to ensure that the tasks which are to be performed in the test correspond with the actual tasks that will be performed outside the test context in the target language usage (TLU) domain. Bachman and Palmer use the term *authenticity* to refer to this correspondence. They assert that authenticity can assist test takers to perform at their best levels and that it facilitates a positive affective response towards the test tasks. As such it is an important control variable for test usefulness (1996: 39).

It should be noted that test validation is a continuous process and that no interpretation of a test score can be considered as absolutely valid. Bachman and Palmer (1996: 22) agree with Messick and others that the process of justifying any interpretations of test scores "starts with test design and continues with the gathering of evidence". Interpretations, however, remain questionable. In terms of this view, construct validity cannot be stipulated in statistical form and needs to be indicated on the basis of the kind of evidence that is needed to support an interpretation of a given score. More evidence is needed for high-stakes testing purposes. Both quantitative and qualitative evidence may be required.

Bachman and Palmer also consider *interactiveness* to be an essential test quality as it concerns the extent to which the constructs that are being assessed constitute an

integral part of the test task. Furthermore, interactiveness is extremely relevant from the point of view of current language teaching and learning principles. Bachman and Palmer (1996: 39) describe interactiveness as a "function of the extent and type of involvement of the test taker's language ability (language knowledge plus metacognitive strategies), topical knowledge, and affective schemata in accomplishing a test task".

The last control variables that provide for the usefulness of a test mentioned by Bachman and Palmer are test *impact* and *practicality*. Tests scores obviously have consequences and as a result the social dimension of language testing is receiving much emphasis. So as to facilitate a positive impact in a testing situation, Bachman and Palmer advise involving test takers by providing them with as much information about the test procedure as possible. This, they claim, will enhance authenticity and interactiveness, while contributing towards a positive perception about the test and a higher motivation to participate (1996: 32).

Finally, practicality refers to the relationship between the available and required resources necessary for the design, development, implementation and use of the test, and also includes logistical constraints.

2.2.3 Weideman's constitutive and regulative conditions

One of the main problems of validation as a process that continues beyond test administration according to the Messick view, is being able to arrive at a point where it can be decided that the evidence obtained is now sufficient. To complicate matters further, in keeping with postmodernist thinking, the results of one validation process may not even be suitable for another testing context. Weideman elucidates this further by explaining that although evidence may never be found to be *sufficient*, evidence is still *necessary* (Weideman 2009a: 236). Drawing on his paradigm of applied linguistics as a discipline of design, he provides a framework for language testing based on two main tenets, which he terms the constitutive and regulative conditions for language assessment. Rather than attempting to subsume any of these conditions under a unitary notion, he discusses each as being of relevance and interrelated, cautioning that the "requirement of conceptual acuity for the sake of an improved designed instrument is not served if concepts are conflated" (Weideman 2009a: 241).

Weideman agrees that scores are meaningless objects without human interpretation, but emphasizes that objective measurements are used to make subjective interpretations. There is thus a distinction to be made between the "subjective process of validation and the objective validity of a test" (2009a: 243). As such, validity can be seen as the achievement of validation. The conditions that

should form part of the validation process are set out in figure 2.1 (based on Weideman 2009a: 248).

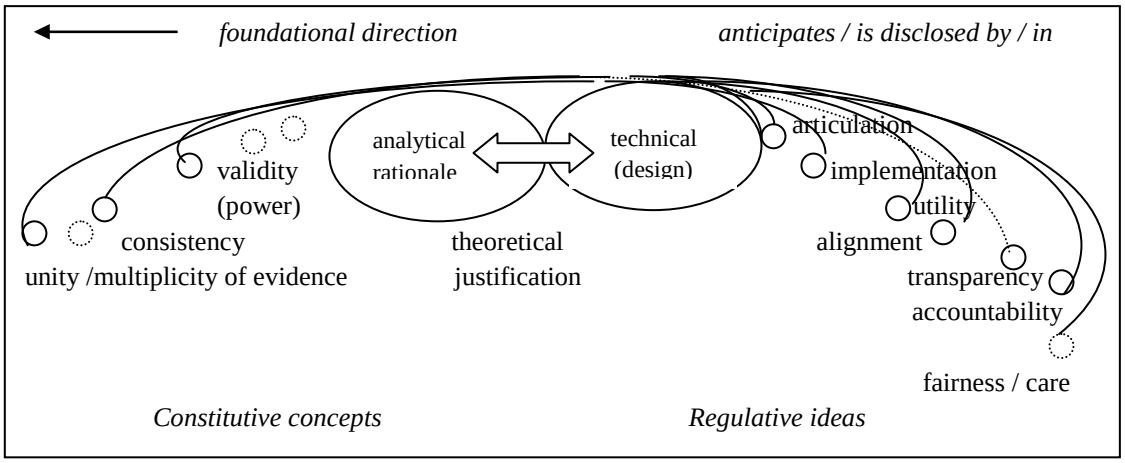


Figure 2.1: Constitutive and regulative conditions for the validation of language tests

In terms of the above representation, the theoretical justification for a language test is to be found in the reciprocal relationship between the analytical and technical modes. The portrayed dimensions cannot be considered absolute and are mutually related. In language testing the technical (design) mode leads and qualifies the design of a solution to a language related problem, while the analytical dimension provides the foundational basis for the intervention. Validity, reliability (consistency) and a unity within a multiplicity of sources of evidence are amongst the constitutive and necessary foundational conditions, while notions such as Bachman's practicality and impact feature under the regulative side of language testing. Weideman (2009a: 246) explains that "in the theoretical

justification that is sought for the design of a test, the original understanding of test validity is mediated through the analytical mode".

The administration of the actual test links the assessment instrument to the social context that features prominently in the regulative conditions. The role of the latter can be clarified by examining the table below (Weideman 2007: 602):

Applied linguistic design	Aspect/function/ dimension/mode of experience	Kind of function	Retrocipatory/anticipatory moment
is founded upon	kinematic	constitutive	internal consistency (technical reliability)
	physical		internal effect/power (validity)
	analytical	foundational	design rationale
is qualified by	technical	qualifying/leading function (of the design)	
	lingual		articulation of design in a blueprint/plan
	social		implementation/administration
is disclosed by	economic		technical utility, frugality
	aesthetic	regulative	harmonisation of conflicts, resolving misalignment
	juridical		transparency, defensibility, fairness, legitimacy
	ethical		accountability, care, service

Table 2.5: Constitutive and regulative moments in applied linguistic designs

The technical function of a test is analogically connected to the economic dimension in the aspect of utility or frugality, which pertains to resource and financial considerations. As part of the validation process, certain trade-offs have to be made between reliability and utility. For example, while longer tests tend to produce a higher reliability coefficient (Geldenhuys 2007), this may prove to be too demanding on the available resources, or the time needed to administer them. Similarly, the technical function of language assessment is also related to the lingual, social, aesthetic, juridical and ethical dimensions. The articulation of a blueprint for the test is indicative of the analogical relation between the technical and the lingual mode of experience and directly related to the actual language tasks required of the test taker. Weideman points out that in attempting to align all of the mentioned factors, the test designer "brings them into harmony within the design", which in turn refers to the aesthetic dimension (2007: 601). The juridical dimension has a bearing on the public defensibility and fairness of a test, whereas the ethical domain elicits considerations of beneficence, such as how employing the test can contribute towards the care or development of test takers so that they are not disadvantaged in any way. Some scholars would be in agreement with this framework, and similarly consider this element to be part of the social or political dimensions of language test designs.

From Weideman's exposition of the many dimensions that contribute towards the validation of a test, it becomes evident that conflating everything under a unitary

concept of validity is not only unnecessary, but undesirable. Notions such as fairness and accountability are themselves subjective concepts and it can be very problematic deciding when a test is fair enough to be considered valid on the basis of Messick's model. For example, in South Africa where little opportunity exists for speakers of African languages to obtain an education in their mother tongue at high-school and tertiary level where the languages of instruction are English and/or Afrikaans, how fair is it to expect these students to write a test of academic literacy and study at tertiary level in the English language medium? However, the reality is that often no other options are available. Academic literacy testing of these students will never be as fair as that of students who could choose to study in their mother tongue and write a literacy test in the language of their choice. The important point to be made is that this situation does not detract from the fact that a test of academic literacy can still be found to be valid and measure what it is supposed to, namely the current level of academic literacy of a test taker, without prejudicing the candidate. From the three different approaches towards the validation of a language test discussed above, it is clear that the main concerns of responsible language testers can all be addressed under the tenets of constitutive and regulative conditions on the basis of the comprehensive framework provided by Weideman.

Chapter 3

The design of a test of academic literacy for postgraduate students

3.1 Defining academic literacy for the purposes of quantification

In language testing the ability to be measured is referred to as the test construct. One definition of the concept of a test construct is that it is an articulation of "…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 *et al.* 1999: 31). Since the particular construct to be measured in the present study is that of academic literacy, this needs to be articulated clearly for the sake of construct validity. Moreover, the test developer must be able to give the assurance to test users and other stakeholders that the "interpretations of ability to be inferred from the assessments generalize to the target language use (TLU) domain, which is the domain that defines the context in which the decisions will be made" (Bachman & Purpura 2010: 463-464). In the instance of a test of academic literacy for postgraduate students, the TLU domain refers to academic tasks that the test taker is likely to encounter within the academic discourse community (see Blanton 1994, Van Dyk & Weideman 2004a).

Although the selection of appropriate test tasks constitutes a major part of the design phase, the construct needs to be substantiated by a theory of language. Academic literacy testing has been preceded by numerous forms of language testing that were initially based on an understanding of language ability as a combination of four finite components – grammar, vocabulary, pronunciation and spelling. These were considered to relate to the four skills of listening, speaking, reading and writing (Bachman & Palmer 1996), sometimes with the further distinction of channel (audio/visual) and mode (productive/receptive). Language testing was thus considered to be a simple matter of testing these four separate skills and nothing more, regardless of the context and without any consideration for specific testing needs. When questions started to be asked about the literacy levels of students at tertiary institutions and pedagogical methods started to move away from a restricted view of language, the dilemma of defining academic literacy became apparent. Rather than attempting to define what academic discourse was, Blanton (1994: 6) proposed asking "*what academic readers and writers do*". Since language is a social activity and medium through which human beings conduct their lives and business, the emphasis should fall on actions and behaviours, rather than on knowledge and linguistic forms. Thus, in order for students to acquire academic reading and writing behaviour, they need to become academic readers and writers. The answer to the problems of language teaching and assessment is then to be found in engaging students in interactions with texts.

Blanton (1994: 8) lists the following 'literate behaviours' that academic readers and writers should be capable of performing:

1. Interpret texts in light of their own experience and their own experience in light of texts;
2. Agree or disagree with texts in light of that experience;
3. Link texts to each other;
4. Synthesize texts, and use their synthesis to build new assertions;
5. Extrapolate from texts;
6. Create their own texts, doing any or all of the above;
7. Talk and write about doing any or all of the above;
8. Do numbers 6 and 7 in such a way as to meet the expectations of their audience.

Blanton also distinguishes between the above literate behaviours and literacy skills, with the latter pertaining more to abilities related to mechanical and formal features such as being able to spell, know vocabulary, grammar, etc. In her opinion it is the behaviours and not the skills that make for the successful employment of academic language. When students decode words of texts without interpreting the texts in terms of their own thoughts and experiences, they have reached a level of functional literacy, but not academic literacy. Blanton states that "individuals whom we consider academically proficient speak and write with something we call *authority*" (1994: 12). Powerlessness in speech and writing is the result of a lack of opportunity to develop an opinion of one's own, in other words the absence of critical reflection. Blanton's definition of academic literacy exposes the fallacy of considering academic language competence to be a mere matter of learning vocabulary and grammar. However, although her construct is well suited to implementation in a language teaching class situation and succeeds

in emphasizing how to give the students a voice while engaging with text, her definition of the ability to use academic language competently is difficult to operationalize in light of logistical constraints.

Taking Blanton's view a step further, Bachman and Palmer (1996: 75-76) do not consider language skills "to be part of language ability at all, but to be the contextualized realization of the ability to use language in the performance of specific language use tasks". Their view of language ability and language skills has hardly been contested in the scholarly community. Current socially enhanced views of language acknowledge the communicative function of language and the employment of listening, speaking, reading and writing for the purposes of negotiating human interaction (see Weideman 2003b). The very act of listening simultaneously incorporates other cognitive processes, including the formulation of thoughts and ideas through language, which itself can be considered an act of producing language, albeit silently within the confines of the human mind. Listening may therefore be conceived of as not merely receptive, but productive, since it involves cognitive processing that is not directly observable. It should be obvious that the ability to hear speech can also not be defined as a language ability *per se*, just as the ability to write letters of an alphabet or even form words or sentences cannot summarily be equated with a language ability, if language is to be understood as the negotiation of meaning, of possessing an interactive nature.

Weideman (2003b: 4) provides the following succinct representation of two contrasting 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

Table 3.1: Two different perspectives on language

Weideman considers a restrictive view of language to be a limited one. What is more, it promotes a deficit view of language ability. Although persons may vary in their levels of proficiency when using language, deficiencies cannot be remedied simply by providing a language user with the needed 'skill', since this is not in keeping with how learning takes place: "acquiring a language, or a new type of discourse in which you are not yet proficient, does not happen by 'receiving' something from an authority" (Weideman, 2003b: 5). Rather, it is through doing tasks or activities in which language is used in a purposeful way that language acquisition takes place. Instead of referring to speaking as a language skill, Bachman and Palmer (1996) propose identifying a language use task involving a speaking activity. Thus, what has traditionally been referred to as a skill can rather be seen as a combination of language ability and task characteristics.

Bachman and Palmer (1996: 61) define language use as "the creation or interpretation of intended meanings in discourse by an individual" or the "dynamic and interactive negotiation of intended meanings between two or more individuals in a particular situation". Meaning is not restricted to utterances or texts, but is connected to how these relate to the language situation and its characteristics. In view of the complex nature of multiple interactions, language ability should be considered within an interactional framework similar to that proposed by Blanton above. Interaction takes place among areas of language ability, topical knowledge, affective schemata and the dynamics of the language use situation. Topical knowledge generally refers to knowledge stored in long-term memory. Affective schemata denote emotional correlates of topical knowledge that influence the response to a task. It is important to note that performance on a task can be influenced just as much by affective schemata as by language ability, especially when emotionally charged topics are used in language tests (hence the avoidance of certain topics). Bachman and Palmer base their definition of language ability on the basis of the components identified by Bachman (1990): language competence (or language knowledge) and strategic competence (metacognitive strategies). And this derives, Van Dyk (2010: 65) has discovered, from the components of competence identified earlier by Skehan, namely linguistic, sociolinguistic, discourse and strategic competence. Together these components provide language users with the "capacity for creating and interpreting discourse" (Bachman &

Palmer 1996: 95). The Bachman model has, however, proven to be problematic and generally unhelpful to test developers. In a book review of Bachman's *Fundamental considerations in language testing*, McNamara (2003) questions the necessity of attempting to characterize the construct of communicative ability in a single model of proficiency. The model's double level of abstraction is also pointed out, along with the problematic *a priori* nature of the abilities it attempts to cover. The model may be fundamentally flawed in that it is "essentially psychological, seeing communicative language ability as a mental ability" (McNamara 2003: 468), whereas the context of language use is understood theoretically to fall within the social realm.

Van Dyk and Weideman (2004a: 8) provide the following representation of Bachman and Palmer's definition of language ability:

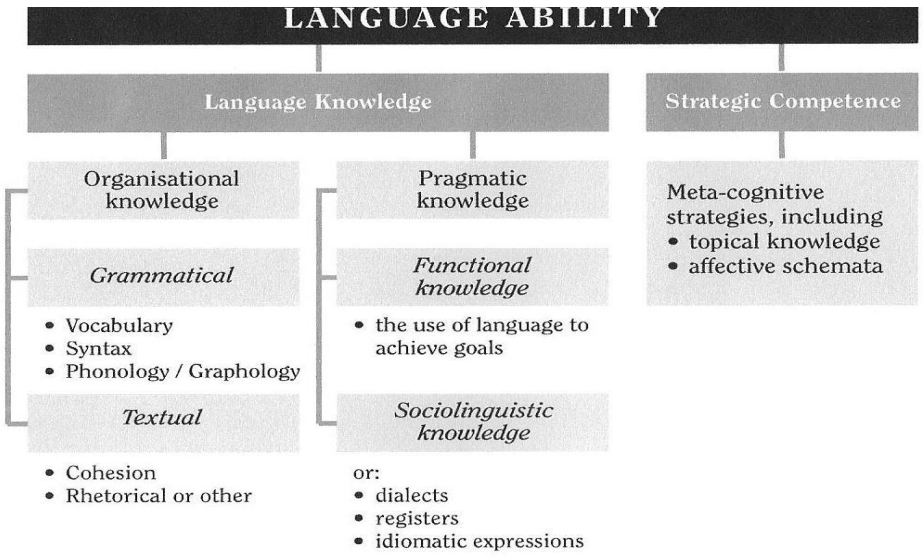


Figure 3.1: The Bachman and Palmer construct of communicative competence

The question may also be asked whether it is possible to separate textual knowledge (knowledge of rhetorical or conversational organization), which resorts under the category of organizational knowledge, from pragmatic knowledge, since it is also related to communicative goals and functional and sociolinguistic knowledge. There may well be seepage amongst a number of these categories.

The above construct of Bachman and Palmer was contextualized as part of the work of the AARP at the University of Cape Town and translated into a number of typical academic literacy tasks and language functions, culminating in the development of the Placement Test in English for Educational Purposes (PTEEP). That test assesses "students' potential to process academic reading and writing at the level required of an entry-level student" (Cliff & Hanslo 2009: 268) and has become a further contributor towards the prediction of academic success. However, the test design used for PTEEP was found to be difficult to administer in the case of large groups of students who literally needed to be tested within a day at tertiary institutions such as the University of Pretoria. In addition to the practical constraints, the AARP test served different purposes, namely the measurement of potential for academic success, i.e. cognitive ability, and the granting of access to tertiary study in a number of cases (e.g. Faculties of Medicine). The desire for an alternative test at the University of Pretoria was premised on the necessity of measuring students' current levels of academic literacy for the purposes of placing

them in the most suitable academic programmes and not, in the first instance, for access reasons. Following the refinement of the definition as part of the AARP, Weideman (2003a: xi) summarized the following idea of academic literacy, in which he incorporated elements from Blanton, Bachman and Palmer, and Yeld and her associates. In terms of this construct, students should be able to:

- understand a range of academic vocabulary in context;
- interpret and use 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 the 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, categorize 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).

For a fairly similar notion to the above, consult Ratangee (2007, chapter 3). Having arrived at a workable framework for and construct of academic literacy, the following section sets out the specification of the subcomponents that are to be assessed and their operationalization into actual task types.

3.2 Specification of test components and task types

Aligning the test construct and its specifications with suitable task types constitutes the next important step in the design process. The following task types were selected initially for inclusion in pilot versions of the TALL (Van Dyk & Weideman 2004b: 18):

TASK TYPE	EXPLANATION
Academic writing tasks	Particularly ones requiring a measure of classification, and that comparisons be made
Cloze procedure	The restoration of a text that has been systematically mutilated
C-procedure	A derivative of cloze, in which half, or just more than half, of every second word is deleted
Dictionary definitions	A derivative of the vocabulary knowledge type, but based on real dictionary definitions of terms taken from an academic word list such as that of Coxhead (2000)
Error identification	The identification of especially South Africanisms
Interpreting and understanding visual and graphic information	Questions on graphs; simple numerical computations
Longer reading passages	Testing a wide variety of academic language abilities
Register and text type	A task that requires the candidate to match the sentences in one list with those in another; usually from different registers
Scrambled text	A paragraph whose sentences have been jumbled and have to be re-arranged
Vocabulary knowledge	Within the context of a short sentence

Table 3.2: Explanation of task types for the pilot version of the TALL

The above tasks were re-organized by Van Dyk and Weideman (2004b: 18-19) into the following specifications and task types for use in further versions of the TALL and also serve as the basis for the current test of academic literacy at postgraduate level which is being administered under the auspices of ICELDA:

SPECIFICATIONS	TASK TYPES
Vocabulary comprehension	Vocabulary knowledge Dictionary definitions Cloze C-procedure
Understanding metaphor and idiom	Longer reading passage
Textuality (cohesion and grammar)	Scrambled text Cloze C-procedure (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 Cloze procedure Longer reading passages Academic writing tasks (possibly also) C-procedure
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 (Interpreting and understanding visual and graphic information)
Communicative function	Longer reading passages (possibly also:) Cloze, scrambled text
Making meaning beyond the sentence	Longer reading passages Register and text type Scrambled text Interpreting and understanding visual and graphic information

Table 3.3: Test construct alignment with specifications and task types

The specifications and task types for inclusion in the test being developed for the purposes of this study will be discussed in more detail in the section that follows.

3.3 The blueprint of the test of academic literacy for postgraduate students

The test that has been designed as part of the current study consists of eight sections (see Annexure D), which in turn serve as subtests of certain aspects of literacy and make an initial contribution towards providing the multiple sources of evidence needed for the attainment of high validity and reliability. The respective test sections and primary elements that they measure are indicated in table 3.4.

TEST SECTION	ASPECT OF LITERACY MEASURED
Section 1: Scrambled text A number of sentences that need to be re-organized into a coherent passage.	Textuality (knowledge of cohesion, grammar) Understanding and responding to the communicative function of the text
Section 2: Interpreting graphs and visual information A short text passage and accompanying graph requiring numerical calculations and visual inferences	Understanding genres Visual literacy Interpreting of information Extrapolation and application of information
Section 3: Academic vocabulary This section includes vocabulary items based on Coxhead's (2000) word list, mainly from the selection of less frequently used words	Advanced vocabulary knowledge Understanding and responding to the communicative function of the text
Section 4: Text types A selection of phrases and sentences representing different genres which have to be matched with a second group of phrases and sentences	Understanding genres Identifying registers Making meaning beyond sentence level
Section 5: Understanding texts A lengthy reading passage and series of questions to be answered	Critical thinking Understanding and responding to the communicative function of the text Deriving meaning beyond sentence level Extrapolating and applying information Distinguishing essential/non-essential information Drawing conclusions and making inferences
Section 6: Grammar and text relations A variation of cloze procedure in which certain words are deleted from a text	Meaning making Understanding and responding to the communicative function of the text Knowledge of cohesion
Section 7: Text editing A passage in which a number of grammatical errors have been made requiring correction	Knowledge of syntax Knowledge of morphology Knowledge of semantics
Section 8: Academic writing A short structured essay assignment based on information provided in the test	Ability to synthesize texts Making meaning beyond the level of the sentence Interpreting information Understanding and responding to the communicative function of the text Extrapolation and application of facts Knowledge of genres and registers Applying coherence Referencing

Table 3.4: The blueprint for the test under development

What is evident from the above table is that academic literacy skills cannot be compartmentalized and measured in isolation. More than one skill is needed simultaneously to access a text at different levels. Even when measuring knowledge of grammar, for example, the ability to understand the message being communicated comes under scrutiny. Of relevance during the refinement stage of test design is the test construct validity reflected through the correlations between the subtests.

3.3.1 Multiple-choice item format

The multiple-choice item format is widely used to assess ability, especially where there are time constraints and large numbers of test takers. However, test takers may score higher in multiple-choice items than they would in an open-construct item format in certain cases. In a study on the effect of the multiple-choice item format on the measurement of knowledge of language structure, which was conducted by Currie and Chiramanee (2010) amongst a cohort of English Foreign Language (EFL) university undergraduates in Thailand, it was found that test participants scored significantly higher marks in the multiple-choice (M/C) test than in the constructed-response (C/R) test equivalent. Their conclusion was that the two formats could not be considered to be measuring exactly the same construct. Currie and Chiramanee (2010: 485) suggest that the two formats are both valid ways of measuring language-related constructs, but that the multiple-

choice format “had the effect of distorting the measurement of the language based abilities” used by the test takers. While 91% of the responses in the C/R were shown to be based on knowledge or the lack thereof, 64% of the corresponding responses in the M/C test items were based on “cued recall, a test taking strategy or guessing” (Currie & Chiramanee, 2010: 485). At the same time it should be mentioned that the guessing of answers does not preclude the application of knowledge.

Since the results of the above study in Thailand can only be generalized to the testing of knowledge of language structure, they were not considered adequate reason to move away from the multiple-choice format for academic literacy testing purposes, especially where critical thinking and reading are involved, as in the current study. A section of the postgraduate test does, however, include the assessment of grammar and text relations and the effect of multiple-choice format testing on these subcomponents will require further attention and research before suggesting any alternative formats. All the same, note should be taken of the reservations expressed by some scholars about using the multiple-choice item format for the testing of knowledge of subject content in particular, as well as concerns about the test takers’ familiarity and experience with this type of questioning. In the case of assessing academic literacy levels there are indications that the multiple-choice format does not pose a serious problem. Test takers have the opportunity to consult sample academic literacy tests before attempting to

write the TALPS and gain an understanding of how multiple choice test items work. Weir (2005: 126) also points out the problematic nature of using an open construct rather than multiple choice format when assessing reading, and requiring test takers to "employ writing to record answers", since this may interfere with the measurement of the envisaged construct. The multiple-choice format remains a practical, efficient and affordable method of administering literacy tests to vast numbers of students simultaneously.

With regard to the effect of the number of distracters used in M/C format questions, Currie and Chiramanee's study (2010) showed very little effect was detected in the use of the 4- and 5-option items, while the 3-option format was found to make the items slightly easier. The test of academic literacy developed for application at postgraduate level only employs 4- and 5-option items.

3.3.2 Vocabulary tasks

A golden rule to be kept in mind when deciding on task types is that the greater the variety of sub-tests used, the more likely the test instrument is to be reliable (Van Dyk & Weideman 2004b). The assessment of vocabulary is one of the main specifications of a test of academic literacy. Although the current trend in language testing is to embed vocabulary within a full reading text so that the meaning of words is context dependent, there still appears to be a need to include

vocabulary tasks not related only to one particular passage. The rationale for this is that by including other task types, provision can be made for testing academic vocabulary across a wider front on the basis of Coxhead's (2000) list of frequently occurring words identified in academic texts. Note should nonetheless be taken of the fact that the *alpha* reliability coefficient of the separate vocabulary section of the current version of the TALPS as calculated by using the Iteman 4.2 programme and indicated in table 3.5, is lower than that of the remaining task types, which may suggest a need for the design of different task types to test vocabulary. The results are provisional and only include the test scores of a cohort of 326 postgraduate students at the University of the Free State (UFS) who wrote the TALPS in the course of 2011.

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	0.924	3.711	0.859	0.655	0.888	0.924	0.792	0.940
Scrambled Text	0.808	0.807	0.619	0.655	0.661	0.765	0.791	0.796
Graph	0.790	1.279	0.646	0.611	0.679	0.785	0.759	0.809
Vocab	0.487	1.358	0.370	0.358	0.369	0.540	0.527	0.539
Text types	0.634	0.925	0.542	0.599	0.270	0.703	0.749	0.426
Under= standing	0.759	1.846	0.446	0.607	0.659	0.617	0.756	0.795
Grammar	0.860	1.590	0.782	0.644	0.805	0.877	0.784	0.892
Text editing	0.860	1.079	0.751	0.746	0.762	0.858	0.854	0.865

Table 3.5: Reliability in an administration of the TALPS at the UFS

The combined *alpha* of 0.924 is very high, attesting to the high quality of the test. Notwithstanding the lower *alpha* for the vocabulary tasks, this section still produced a highly satisfactory facility value (mean P) and discrimination index (Rpbis), as can be seen in table 3.6. Good items are able to differentiate between test takers of high and low ability. A point-biserial (pbis) of 0.0 means that there is no differentiation between test takers. The discrimination index is rarely above 0.50, though (Guyer & Thompson 2011).

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
All items	76	48.512	13.460	16	76	0.638	0.357
Scored Items	76	48.512	13.460	16	76	0.638	0.357
Scrambled	5	2.209	1.841	0	5	0.442	0.272
Graph	10	6.325	2.788	0	10	0.633	0.438
Vocabulary	10	6.908	1.895	1	10	0.691	0.230
Text types	5	2.433	1.529	0	5	0.487	0.219
Understanding	21	14.684	3.757	4	21	0.699	0.299
Grammar	15	8.684	4.243	0	15	0.579	0.468
Text editing	10	7.270	2.884	0	10	0.727	0.470

Table 3.6: Summary statistics of the 2011 administration of the TALPS at the UFS

On the basis of the above statistics there is sufficient evidence to conclude that the blueprint used for the current test version is usable and effective.

3.3.3 Choosing reading texts

Prominence is given throughout the test to the element of reading, since the way a reader engages and interacts with texts is considered by a number of scholars to be the best articulation of academic literacy (Moore, Morton & Price 2007). Identifying appropriate reading texts is thus one of the most important aspects in designing a test. In the case of the TALPS there is usually a central theme running through the test. Selecting texts that complement each other is thus necessary to provide sufficient factual information and pre-writing activities for the undertaking of the writing task at the end of the test. The theme of renewable energy was elected for the purpose of the current study on the basis that this is a topic of global importance, which all postgraduate students should be familiar with to some or other extent.[4]

There is much discussion currently on the desirability of subject-specific tests since these tend to facilitate a positive attitude towards test taking – an affective aspect which is important from the perspective of test transparency. However, the point of view adopted in the present study is that a test of academic literacy should not evolve into a test of knowledge of subject content. Rather, candidates should in fact be able to negotiate unfamiliar content using the language proficiency at their disposal. It may be hypothesized that electing to use subject content of which

4. General knowledge itself may be considered to be a form of academic literacy.

test takers have a general, but not in-depth knowledge, will have the effect of compelling test takers to draw on language-related abilities, which may provide a better indication of their literacy levels than would be the case when employing subject material from the test takers' own or chosen fields of study. Nonetheless, the subject reading matter included in the test should not be so discipline-specific that test takers are unable to relate to it.

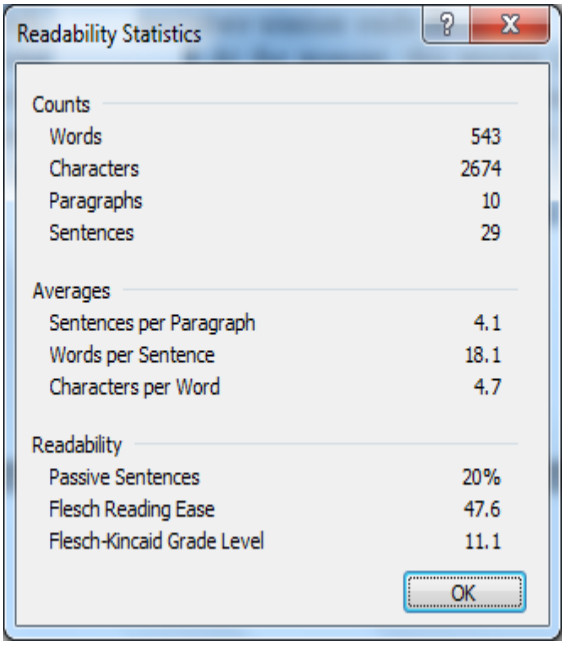
Controversial topics pertaining to religion, race, crime and personal health, for example, are generally to be avoided. These can elicit negative emotions and connotations that could become a hindrance to test performance and diminish or weaken the juridical mode in terms of which every care must be taken to ensure fairness of test items. Amongst the topics covered in current test versions available from ICELDA are subjects such as global warming and the desirability of genetically modified foods. Selecting texts of this nature, which deal with issues of actuality that are widely debated amongst both members of the public and academe, reinforces further the alignment of the test construct with authentic discourse in keeping with the communicative approach and an enriched view of language.

The reading passages envisaged for use in a test are scrutinized beforehand on the basis of a matrix to determine their suitability for inclusion. Van Dyk (2010: 208) provides the following matrix:

Comments on Text 1			
Reading ease index	**Grade level**	**Text length**	**Comments for improvement (if relevant)**
E.g. 55%	E.g. Grade 11	E.g. 500 words	E.g. Too few discourse markers

Table 3.7: Matrix for the comments of subject specialists on the text passage

The reading texts included in the test designed as part of the current study had an initial Flesch-Kincaid reading ease index of 47.6 and a Grade level of 11.1, which were considered adequate for the purposes of starting to design the test and in line with the grade levels used in versions of the TALL. However, since the test under construction is aimed at postgraduate students, the reading text needed to be made more challenging. The original text and amended version are provided below. The shaded sections in the second text illustrate how the reading grade level has been increased through the addition of a number of words and phrases to the original text, including verbs, linking adverbials and modifiers. Through these changes the reading ease has been lowered by nearly 10% and the grade level increased to 13.0, which is in line with that of the existing version of the TALPS.



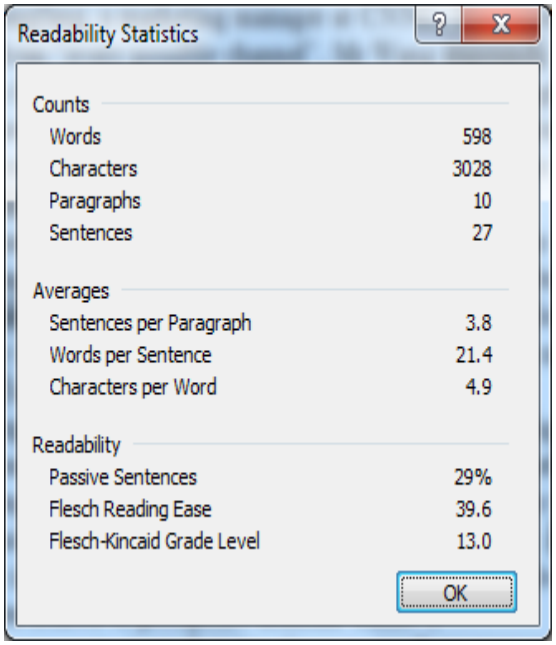


Table 3.8: Increasing the difficulty of the reading comprehension text

Original text: Novel sources of uranium

Rising from the ashes

Coal ash, fertiliser and even seawater may provide nuclear fuel

ONE of the factoids trotted out from time to time by proponents of nuclear power is that conventional coal-burning power stations release more radioactivity into the environment than nuclear stations do. The reason is that the ash left over when coal is burned contains radioactive elements, notably uranium and thorium.

2 Turn that logic on its head and it suggests that such ash is worth investigating as a source of nuclear fuel. And that is exactly what Sparton Resources, a firm based in Toronto, is doing. It has signed a deal with the China National Nuclear Corporation (CNNC), the authority that runs the country's nuclear-power stations, to recover uranium from coal ash at a site in Lincang, in Yunnan province.

3 Uranium is usually extracted from ore that contains 1,000 or more parts per million (ppm) of the element. The Lincang coal ash holds much less, about 300ppm. That said, it does not need to be mined—which brings costs down. Sparton says it can extract a kilogram of uranium for $77 or less. Uranium's spot price is now near $90 a kilo. That is not a huge margin, but it is a profit nonetheless.

4 To extract the uranium, Sparton adds sulphuric and hydrochloric acids to the ash, along with water, to make a slurry. With some sorts of ash, nitric acid is also used. The acids dissolve the uranium, and various other things, leaching them from the ash. The trick is to get the dissolved uranium out of the resulting solution.

5 China is developing ash-mining for reasons of energy security more than economics, according to Wang Hongfang, a marketing manager at CNNC. The country wants to get uranium from "every possible channel", Mr Wang says. This includes stripping it out of the tailings from gold and copper mines, and also from phosphoric acid produced during the manufacture of fertiliser. Nor is CNNC alone in this aspiration. NUKEM, a German-American company that enriches and sells nuclear fuel, hopes soon to begin "mining" fertiliser in Florida.

6 Some people are even turning to seawater as a source of uranium, in an eerie recapitulation of Fritz Haber's attempt to pay off Germany's first-world-war debts by extracting gold from the ocean. Though seawater contains only three parts per billion of uranium, mostly in the form of uranyl tricarbonate, the element can be sucked out of it by ion exchange.

7 Several organisations, including Japan's Atomic Energy Agency and the Bhabha Atomic Research Centre in India, are attempting to do so. Their methods include the use of strips of ion-exchanging plastic, braided with polystyrene to toughen them up. These are placed in wire cages and anchored in a current of seawater. After a month or two, the plastic is removed and soaked in acid to dissolve the uranyl tricarbonate. The solution is then treated to precipitate uranium oxide.

8 At the moment, this process costs more than ten times as much as conventional mining, but some countries might regard that as a small price to pay for security of supply. Perish the thought that the supply is for anything other than providing fuel for civilian nuclear-power stations. ■

[Adapted from *The Economist*, 10 April 2010, pp. 74-75.]

Amended text: Novel sources of uranium

Rising from the ashes

Coal ash, fertiliser and even seawater may provide nuclear fuel

ONE of the factoids trotted out from time to time by proponents of nuclear power is that conventional coal-burning power stations release more radioactivity into the environment than nuclear stations do. The reason furnished for this is that the ash left over after coal has been burned contains certain radioactive elements, notably uranium and thorium.

2 So, if we turn that logic on its head and it suggests that such ash would be worth investigating as a potential source of nuclear fuel. And that is precisely what Sparton Resources, a firm based in Toronto, is attempting to do. It has concluded a contract with the China National Nuclear Corporation (CNNC), the authority that is responsible for running the country's nuclear-power stations, to recover uranium from coal ash at a site in Lincang, which is located in Yunnan province.

3 Uranium is usually extracted from ore that contains 1,000 or more parts per million (ppm) of the element. The Lincang coal ash, however, holds much less, about 300ppm. Be that as it may, it does not need to be mined—which brings costs down considerably. According to calculations done by Sparton, a kilogram of uranium can be extracted for $77 or less. Considering that uranium's spot price is now approaching $90 a kilo, this does not represent a huge margin, but it is a profit nonetheless.

4 In order to extract the uranium, sulphuric and hydrochloric acids are added to the ash, along with water, to make a slurry. With some sorts of ash, nitric acid is also used. The acids dissolve the uranium, and various other things, leaching them from the ash. The trick is to get the dissolved uranium out of the resulting solution.

5 China is engaging in ash-mining for reasons of energy security more than economic considerations, according to Wang Hongfang, a marketing manager at CNNC. The country is pursuing a search to obtain uranium from "every possible channel", Mr Wang expounds. This includes stripping it out of the tailings from gold and copper mines, and also from phosphoric acid produced during the manufacture of fertiliser. Nor is CNNC alone in this aspiration, since NUKEM, a German-American company that enriches and sells nuclear fuel, hopes soon to begin "mining" fertiliser in the vicinity of Florida.

6 Some entrepreneurs are even turning to seawater as a source of uranium, in an eerie recapitulation of Fritz Haber's attempt to pay off Germany's first-world-war debts by extracting gold from the ocean. Although seawater contains only three parts per billion of uranium, which is to be found mostly in the form of uranyl tricarbonate, the element can be sucked out of it through the process of ion exchange.

7 Several of the leading international organisations, including Japan's Atomic Energy Agency and the Bhabha Atomic Research Centre in India, are attempting to do so. Their methods include the use of strips of ion-exchanging plastic, braided with polystyrene to toughen them up. These are inserted in wire cages and subsequently anchored in a current of seawater. After a month or two has lapsed, the plastic is removed carefully and soaked in acid to dissolve the uranyl tricarbonate. The solution is then treated to precipitate uranium oxide.

8 At the present moment, this process costs more than ten times as much as conventional mining activities, but some countries might regard that as a small price to pay for security of supply. Perish the thought that the supply is intended for anything other than providing fuel for civilian nuclear-power stations. ■

[Adapted from *The Economist*, 10 April 2010, pp. 74-75.]

3.3.4 Writing tasks

There are certain practical considerations that should be taken into account when deciding on task types. For example, longer reading passages are more time consuming than shorter extracts of texts, while writing tasks are liable to subjective marking. This raises questions as to the necessity of including writing tasks, especially in the light of studies that show a positive correlation between reading and writing ability (Flower 1990, Hirvela 2004). The writing section was in fact removed from the academic reading test of the International English Language Testing System (IELTS) – one of the biggest international language testing organisations – when studies showed that test takers differed considerably in the way they employed the reading material in the writing section of the test and that this could have implications for test fairness and construct validity. A second noteworthy reason was the possible confusion of assessing writing ability and reading ability (see the report by Moore, Morton & Price 2007: 6). The test which is being designed as part of the current study still includes a writing section at the end of the test for the purposes of face validity, but the dispute as to whether this task type assesses reading or writing ability certainly requires further attention.

3.4. Conclusion

The design of a test of academic literacy is first and foremost dependent on a clear understanding and articulation of the construct to be measured. Failure to

define the construct in specific terms before proceeding with the design of the test components, may have serious repercussions in terms of construct validity and could even render any interpretations of ability based on the test scores to be null and void. Careful consideration thus needs to be given to what a test of academic literacy is supposed to assess and which theoretical framework will guide the process of task selection and development. Additionally, any testing event will inevitably create an artificial testing context which can only be counteracted through the meticulous alignment of test tasks with authentic moments of language interaction and production. Not only should the content material selected for inclusion in the test resemble the discourse encountered within a tertiary environment, but the very test items themselves should reflect those tasks typically demanded of students at higher education institutions. In this respect the most contentious aspect of designing a test of academic literacy at postgraduate level would seem to be whether to include a writing section or not. Is an essay type of question necessary to determine the ability to argue a case, or can an indication of writing ability be obtained through alternative test items that assess the ability to identify evidence that supports a point of view? If communicative competence is to be regarded as an combination or amalgamation of linguistic and strategic competence, of language knowledge and cognitive ability, then one would expect there to be a strong correlation between critical reading and thinking ability, and writing ability. Testing critical reading and thinking may then suffice to gain an indication of writing potential, which would make literacy assessment logistically less complicated.

In addition hereto, rater or marker bias constitutes a real threat to any endeavour to ensure parity and fairness of assessment and removing the need for subjective marking would also serve to bolster consistency of measurement.

Chapter 4
The refinement phase of the test

4.1 The need for piloting and revision

Academic literacy tests require considerable trialing and revision as part of the development process, in order to meet the required constitutive and regulative conditions for test design discussed in Chapter 2. In addition to the design stage, in which the target language use domain is described in full and test specifications are set, as expounded on in the preceding section, there are a further two primary stages in the development of an academic literacy test, which Bachman and Palmer (1996: 90-91) refer to as operationalization and administration. While the former concerns the writing of the actual tasks, providing test instructions and stipulating scoring procedures, the administration phase includes the piloting of the test on a trial basis, after which information is collected and analyzed for any necessary amendments before continuing with the full administration of the test. Re-piloting may even be necessary.

The above procedures are essential to gather the evidence that will substantiate the usefulness of the inferences based on the test scores and to prevent any prejudicial treatment of test takers. The employment of recognized statistical procedures for the analyses of the quantitative data obtained through the piloting of the test is just as crucial a part of the validation process as the

definition of the construct and determination of test specifications. Statistical analyses provide a means of evaluating and improving the quality of tests and of ensuring to some degree that tests are used in a fair manner (Bachman 2004: 135).

Note should be taken of the fact that the measurement process has its limitations. Bachman (2004: 27) mentions six restrictions: "underspecification, indirectness, incompleteness, imprecision, subjectivity and relativeness". There are a number of factors that cannot be accounted for in the language testing process, such as topical knowledge and anxiety, for example. These aspects are thus not specified and are ignored when defining the construct and as a result any interpretations based on test performance will not be fully determinate. Indirectness refers to the fact that abilities are not observable directly. Moreover, any observations of performance will not be complete, since they cannot reflect the complete ability of an individual to perform. The challenge is to be able to consider test performance as being representative of an individual's complete performance. Imprecision relates to measurement procedures and rating scales that can impact on the accuracy of assessment. The fifth limitation, subjectivity, is not necessarily something negative, but Bachman cautions that the usefulness of measurements may be limited by decision errors on the part of test designers, test takers and test users. Design decisions can therefore positively or negatively influence the usefulness of the scores. Relativeness is reflected in the fact that the abilities being measured can

only be interpreted relative to a standard that has been set outside of the measure itself.

All professionally developed tests should meet the accepted standards of testing practice with regard to statistical reliability and validity to avoid allegations of unfair item bias. Hunter and Schmidt (2000) point out that findings of biased test items may be attributable to statistical or mathematical errors resulting from statistical procedures that have not been founded on substantive theory (Hunter & Schmidt 2000). Kline (2004: 559) cautions that measurement bias should be probed at the test level, because this is the level at which decisions about individual test takers are made. She advises against the summary removal or revision of items that show bias, since these may in fact be indicative of a difference in ability and thus directly related to the construct being measured. This is a common mistake, in fact, in the advice of some testing professionals who assume that every item that shows differential item functioning (DIF) must necessarily be removed.

4.2 Piloting the alternative test

Owing to time constraints and logistical problems, the test as a whole, which takes at least 2 hours to complete, could not be piloted. However, sections 1, 4 and 5 of the 'renewable energy' alternative version of the TALPS were piloted as a class test on a cohort of 150 students at the UFS, as well as parts of sections 3 and 7. The decision to include these sections in the class test was

based on the course design of the academic literacy module for which the above-mentioned students had registered. Section 1 included a scrambled text in which the sequence of the five sentences had been altered. Students had to place the sentences in the correct order, drawing on their knowledge of cohesive ties and other textual clues. Only 4 of the 10 vocabulary words from section 3 could be included in the class test owing to time constraints. Section 4 required of students to match text extracts representing different genres and registers. In section 5 students had to answer a number of questions based on the reading passage, "Novel sources of uranium", to test their general understanding of the text, as well as their ability to interpret the information provided. Students were expected to make inferences, comparisons and associations, arrive at conclusions, distinguish between fact and opinion and know what counts as evidence to support an argument. In addition hereto, they were also required to do basic numerical calculations involving percentages and fractions based on numerical information provided in the text. Their language competence was also assessed through questions relating to the use of anaphoric references, metaphors, idioms and synonyms. The last part of the class test, which forms part of section 7 of the alternative TALPS, required students to edit a passage of text in which a number of grammatical errors had been made. Aspects such as knowledge of subject-verb agreement, verb tenses, prepositional phrases, pronouns and spelling were assessed. This section of the class test had to be expanded considerably for the 2-hour test version which also includes a section on grammar and text relations (section 6) in which

students have to identify the missing words in a text. Section 8 could not be piloted, as the writing section is dependent on the completion of sections 1 to 7. In the full-length test students have to use information provided in the test passages to write a short structured essay in which they argue which sources of energy South Africa should be investigating, so as to be able to meet the growing need of the country for electricity. Sufficient information is provided in the selection of text passages to enable the test takers to be aware of the different sources of energy that are currently available and the advantages they offer. The objective of this task item is to see whether students are able to formulate their own opinions, base their arguments on factual evidence gleaned from the reading texts, and communicate this coherently according to accepted conventions for producing academic texts.

The students who participated in the pilot were representative of different academic departments and faculties. None of these students were postgraduate enrolments, but were in their second to fourth year of study. Since ENG 104 is an academic literacy module, the students were expected to perform at a lower level than might be the case with an actual cohort of postgraduate students. It was not practically possible to find a large enough cohort of voluntary postgraduate test takers, which necessitated piloting the test as a class test in an undergraduate course module. The mean for the test was 45%, which suggests that the test was too challenging for the group of undergraduate students, but should be appropriate for graduates. Results from different administrations of

the TALL indicate that the mean for comparative groups of students is usually considerably higher (see Loan Le 2011, Van der Slik & Weideman 2009).

4.3 The productivity of test items

As part of the refinement process, the task types contained in the test needed to be tested for productivity, which means that the test items should neither be too difficult, nor too easy and should also discriminate well between candidates of different ability. This is usually done on the basis of a quantitative measure (see chapter 4 of Bachman 2004).

Not all task types are equally productive and can be considered to fall into one of four categories:

> *acceptable* (a high degree of alignment with the test construct, but apparently not productive), *unacceptable* (low productivity coupled with small degree of alignment with blueprint), *desirable* (high alignment with construct, as well as productive), or *not ideal* (potentially productive, but not quite aligned with framework (Van Dyk & Weideman 2004b: 17-18).

Mention should also be made of the possibility of specifying different weightings to task types in accordance with a particular ability that is given prominence in a test so as to achieve the desired productivity. It is advisable to assign weightings to the test questions after all the test sections have been piloted and refined.

4.4 Results of the statistical analyses of the pilot test

For the purposes of this study both descriptive and inferential statistical analyses were used to determine the productivity of the test items. Whereas the first-mentioned helps to describe score characteristics of a group or sample of individuals, inferential statistical analyses assist with the making of inferences or generalizations about the performance of a larger group or test population. A full report of the analyses of the test items, which was generated through the Iteman 4.2 programme, is attached as Annexure B.

4.4.1 Distribution of test scores

Figure 4.1 displays the distribution of the raw scores for the dichotomously scored items across all domains.

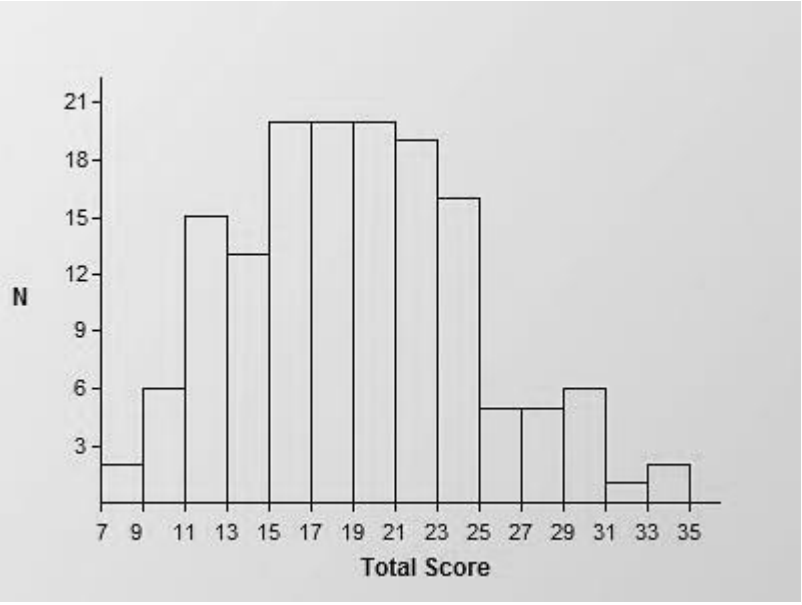


Figure 4.1: Distribution of raw scores for the ENG 104 class test

Most of the students scored in the 15 to 22 band. The highest score was 35 and the lowest 7. The same results can be presented in the form of a table using Iteman 3.6.

Number Correct	Freq-uency	Cum Freq	PR	PCT	
. . . No examinees below this score . . .					
6	0	0	1	0	|
7	1	1	1	1	|#
8	1	2	1	1	|#
9	4	6	4	3	|###
10	2	8	5	1	+#
11	8	16	11	5	|#####
12	7	23	15	5	|#####
13	7	30	20	5	|#####
14	6	36	24	4	|####
15	13	49	33	9	+#########
16	7	56	37	5	|#####
17	8	64	43	5	|#####
18	12	76	51	8	|########
19	13	89	59	9	|#########
20	7	96	64	5	+#####
21	7	103	69	5	|#####
22	12	115	77	8	|########
23	8	123	82	5	|#####
24	8	131	87	5	|#####
25	3	134	89	2	+##
26	2	136	91	1	|#
27	2	138	92	1	|#
28	3	141	94	2	|##
29	3	144	96	2	|##
30	3	147	98	2	+##
31	0	147	98	0	|
32	1	148	99	1	|#
33	1	149	99	1	|#
34	0	149	99	0	|
35	1	150	99	1	+#
36	0	150	99	0	|
37	0	150	99	0	|
. . No examinees above this score . . .					|
					|----+----+----+----+---+
					5 10 15 20 25
					Percentage of Examinees

Table 4.1: Score distribution of the ENG 104 pilot using Iteman 3.6

The distribution is negatively skewed, which is to be expected in a classroom test with relatively small numbers of scores. According to Bachman (2004: 74), a kurtosis of between -2 and +2 is indicative of a reasonably normal distribution. The kurtosis of the class test was -0.184, which is well within these parameters.

Iteman 3.6 generated the following scale statistics for the class test:

Scale:	1
N of Items	41
N of Examinees	150
Mean	18.620
Variance	31.396
Std. Dev.	5.603
Skew	0.359
Kurtosis	-0.184
Minimum	7.000
Maximum	35.000
Median	18.000
Alpha	0.739
SEM	2.861
Mean Pcnt Corr	45
Mean Item-Tot.	0.219
Mean Biserial	0.288
Max Score (Low)	15
N (Low Group)	49
Min Score (High)	22
N (High Group)	47

Table 4.2: Scale statistics generated by Iteman 3.6 for the class test

The following provides a summary of the main statistical findings of the test as a whole, as generated by Iteman 4.2. These will be explained in more detail in the sections to follow. Note should be taken of the fact that the *alpha* value is slightly higher than that provided by Iteman 3.6. Iteman 4.2 is considered to be

more authoritative since it is a more refined version (Guyer & Thompson 2011: 1).

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis	
Scored Items	41	18.620	5.622	7	35	0.454	0.207	
Score	**Alpha**	**SEM**	**Split-Half (Random)**	**Split-Half (First-Last)**	**Split-Half (Odd-Even)**	**S-B Random**	**S-B First-Last**	**S-B Odd-Even**
Scored items	0.748	2.821	0.534	0.443	0.641	0.696	0.614	0.781

Table 4.3: Summary statistics of the ENG 104 class test using Iteman 4.2

The above results can be indicated per test domain (section area) to show how each section compares to the rest of the test.

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
All items	41	18.620	5.622	7	35	0.454	0.207
Scored Items	41	18.620	5.622	7	35	0.454	0.207
Scrambled text	5	1.900	1.587	0	5	0.380	0.283
Understanding	22	9.067	3.065	2	17	0.412	0.166
Text types	5	3.367	1.397	0	5	0.673	0.287
Vocabulary	4	1.993	1.245	0	4	0.498	0.292
Text editing	5	2.293	1.240	0	5	0.459	0.165

Table 4.4: Summary statistics of the ENG 104 class test according to content domain

4.4.2 Consistency or reliability

Cronbach's *alpha* is a reliability statistic that shows the "degree to which the observed scores represent the 'true' scores", without measurement error (Van

der Slik & Weideman 2005: 26). The Iteman 4.2 programme calculates the contribution of each item to the test as a whole and helps to indicate items that do not perform well (Van der Slik & Weideman 2005: 24). Not only is the *alpha* KR-20 reliability coefficient calculated with each item deleted, but a number of split-half reliability coefficients are also generated, both with and without Spearman-Brown correction. The split-half method provides the most common way to estimate the reliability of a test. In terms hereof the test is split into two halves and the correlation coefficient between these is calculated (see Hardyck & Petrinovich 1975: 155 for a more detailed discussion). Iteman 4.2 provides three configurations of split-half reliability, first as uncorrected correlations, and thereafter as Spearman-Brown (S-B) corrected correlations. The reason for this is that reliability is underestimated when an uncorrected split-half correlation is referenced to a 'test' that only contains half as many items as the full test (Guyer & Thompson 2011: 23). Iteman 4.2 calculated an *alpha* of 0.748 for the ENG 104 test, which is acceptable for a class test. The goal, however, remains to achieve an *alpha* reliability in excess of 0.9 in keeping with the high standard of the existing versions of the TALL and TALPS. Reliability should also be viewed in the light of the purpose of a test. In the case of high-stakes tests which are used for access purposes, the *alpha* value needs to be in excess of 0.9. On the other hand, where a test is used for placement or other purposes and constitutes one of several types of information on the basis of which certain inferences are made, an *alpha* value of 0.8 is desirable (see Hogan 2007: 149-150 for a more detailed explanation).

Both facility and discrimination values play a role in determining the reliability of a test and will be discussed next. Often the reason for the weakness of a test item can be found in either or both of these values. Through further rounds of refinement and piloting on a significantly larger cohort of test takers, it is possible to increase the reliability index. The *alpha* values for the ENG 104 test as a whole and per domain are indicated in table 4.5.

Score	Alpha	SEM	Split-Half Random	Split-Half First-Last	Split-Half Odd-Even	S-B Random	S-B First-Last	S-B Odd-Even
Scored items	0.748	2.821	0.644	0.443	0.641	0.784	0.614	0.781
Scrambled text	0.749	0.795	0.722	0.551	0.722	0.838	0.711	0.838
Understanding	0.524	2.115	0.464	0.297	0.420	0.634	0.458	0.592
Text types	0.720	0.739	0.659	0.578	0.437	0.794	0.732	0.608
Vocabulary	0.533	0.851	0.296	0.368	0.296	0.457	0.538	0.457
Text editing	0.379	0.977	0.286	0.229	0.220	0.445	0.373	0.360

Table 4.5: Summary of a*lpha* values for the test

As can be seen in the above table, the text editing section had a low reliability index and needed to be revised substantially.

4.4.3 Facility values

Facility values reflect the percentage of correct answers for the whole of the test population. From the facility values (Mean P) in table 4.4, it is evident that the test was difficult. A more senior cohort of students with higher academic literacy levels could thus be expected to perform better and produce higher

facility values. Items should vary in difficulty and should reflect p-values between .2 and .8 (Bachman 2004: 138). There should also be a normal distribution of scores indicating that the number of items that are very difficult or very easy is not disproportionate to the facility values of the remaining test items. The ENG 104 class test showed a good distribution of easy and difficult items, although some further refinement would be necessary.

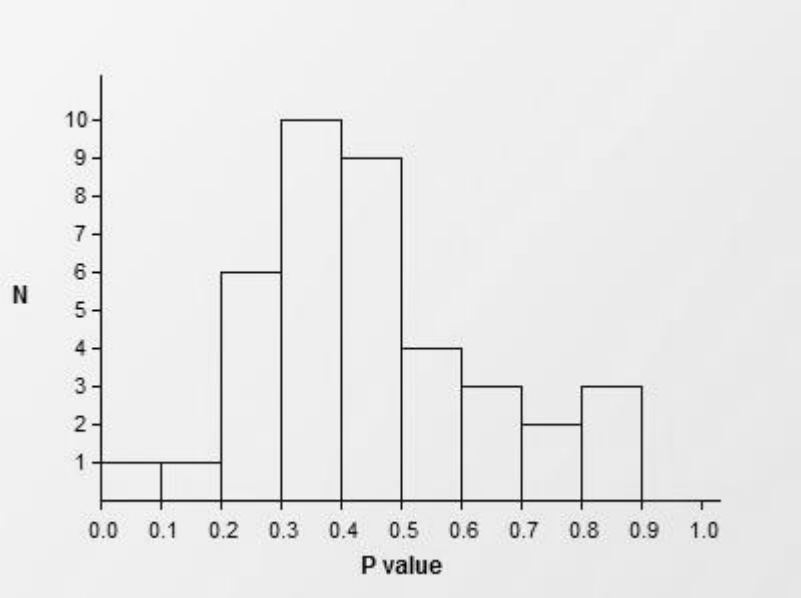


Figure 4.2: Distribution of p-values for dichotomously scored class test items

Most of the items fall within the parameters of .2 and .8, but there is one item that is much too difficult and three that may be considered too easy.

With the exception of Text Types, the mean p-values of the individual sections closely resemble the mean P of the test as a whole (see table 4.4).

4.4.4 Discrimination indexes

How well a test item can distinguish between strong and weak test takers is referred to as the discrimination index. Items with high discrimination indexes make the test more reliable (Loan Le 2011: 22). Accordingly, individuals who answer a test item with a high discrimination index correctly should obtain a high score in the test. Similarly, weaker candidates who answer the item incorrectly could be expected to obtain a low score in the test. Thus, there should be a high positive correlation between the respective item scores and total test scores.

A correlation coefficient is calculated to show the "strength and direction of the relationship" between two variables (Bachman 2004: 84). The Pearson point-biserial correlation (*r-pbis*) is normally used in language testing to measure the differentiating strength of a test item. The correlation can range from -.0 to 1.0, but does not usually exceed 0.50 (Guyer & Thompson 2011: 30). If a negative point-biserial is obtained, this shows that the item is a poor one and that strong candidates are answering it incorrectly, while weaker test takers are providing correct answers. If the point-biserial is 0.0, this means that the item does not provide any differentiation between low and high scoring examinees and that this item also needs to be refined or rejected. In addition to the Pearson point-biserial (*r-pbis*), Iteman 4.2 also provides an item biserial (*r-bis*) correlation. This is an estimate of Pearson's *r* and is larger in magnitude than the point-biserial. Guyer and Thompson (2011: 30) explain the reason for this, stating

that the correlation is "computed between the item and total score as if the item was a continuous measure of the trait". Preferably, items should yield *r-bis* values of at least 0.25 and higher (Weideman 2012: 9).

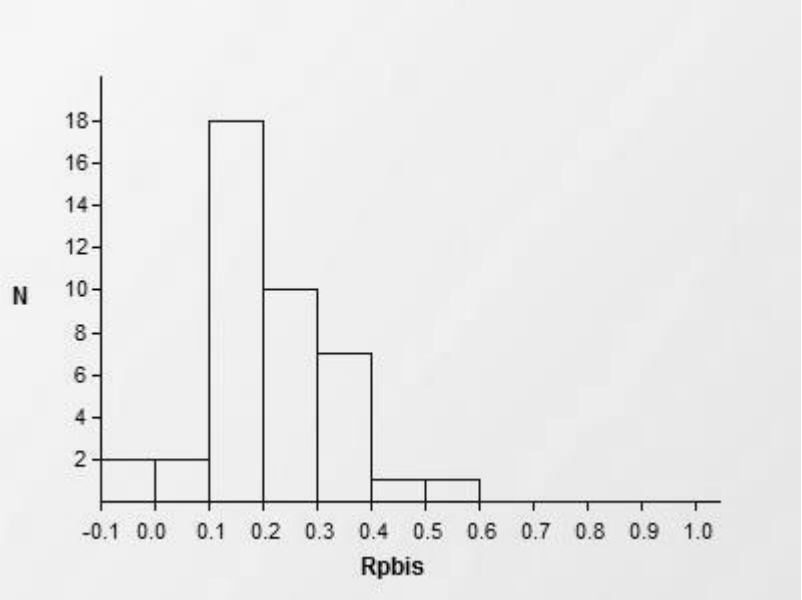


Figure 4.3: Pearson point-biserial (*r-pbis*) for the class test

Only two items had negative discrimination indexes, but several fell within the 0.0 to 0.2 range. At the same time, it should be kept in mind that the *r-bis* values will be higher than the *r-pbis* readings provided in figure 4.3. Many of the items with low discrimination indexes came from the section Understanding Texts and this suggests that a number of items in this section are in need of refinement. That would also help to explain why the *alpha* for Understanding Texts was only 0.524.

4.4.5 Dimensionality

A scatterplot is used to illustrate to what extent the test is one-dimensional or multi-dimensional, i.e. whether one or more abilities are being tested at the same time. This is depicted by the way in which the items of test sections cluster in more or less the same range. A rich construct such as academic literacy is likely to indicate a certain measure of heterogeneity and that more than one ability is being assessed simultaneously, as is evident in the scatterplot provided in figure 4.4 which illustrates dimensionality in a TALL test administered at the University of Pretoria in 2008 (Weideman 2012: 9).

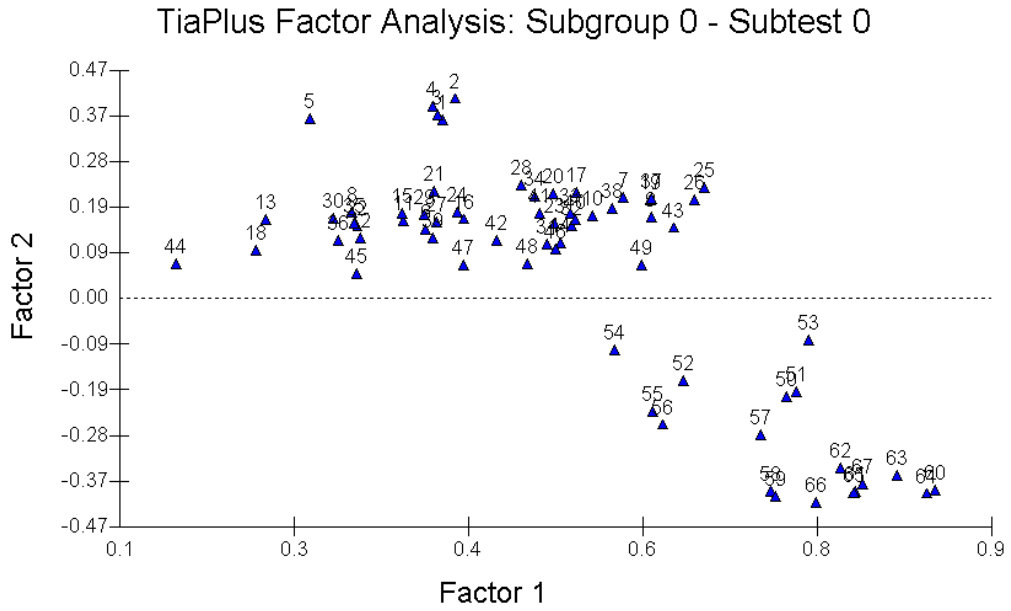


Figure 4.4: An indication of dimensionality in a TALL test

In figure 4.4 we see that there are two subtests (items 1-5 and 51-67) further away from the zero line that is associated with a homogeneous construct (the measurement of a single factor). Provided that the outlying test items display a

reasonable measure of association with one another, a heterogeneous test construct is acceptable. In order to generate a scatterplot such as the above, larger cohorts of test takers are required, something that was not within the ambit of the present study.

4.5 The refinement of unproductive test items

The test items in table 4.6 were flagged by Iteman 4.2, which indicates that they were not productive and needed to undergo refinement. A word of caution is due here. Just as the assessment of literacy is not a simple matter of testing listening, speaking, reading and writing skills, the refinement of task types does not merely entail the elimination of less productive tasks and the retention of the most productive ones. It should be kept in mind that multiple abilities are involved in any one particular task and that possible causes for the unproductivity of a test item need to be investigated in full before deciding to eliminate an item. Furthermore, a much larger cohort of test takers is needed in order to obtain a more accurate picture.

Item ID	P / Item Mean	R	Flag(s)
6	0.067	-0.006	K, LR
8	0.333	0.142	K
16	0.260	0.015	K
18	0.600	-0.005	K, LR
26	0.287	0.114	K

Table 4.6: Summary statistics for the flagged items
Key: K = Key error[5]; LR = Low *r-pbis*

5. The *r-pbis* for a distractor is higher than that for the answer key.

An analysis of the flagged items follows.

4.5.1 The refinement of item 6

Iteman 4.2 provides the following data on item 6:

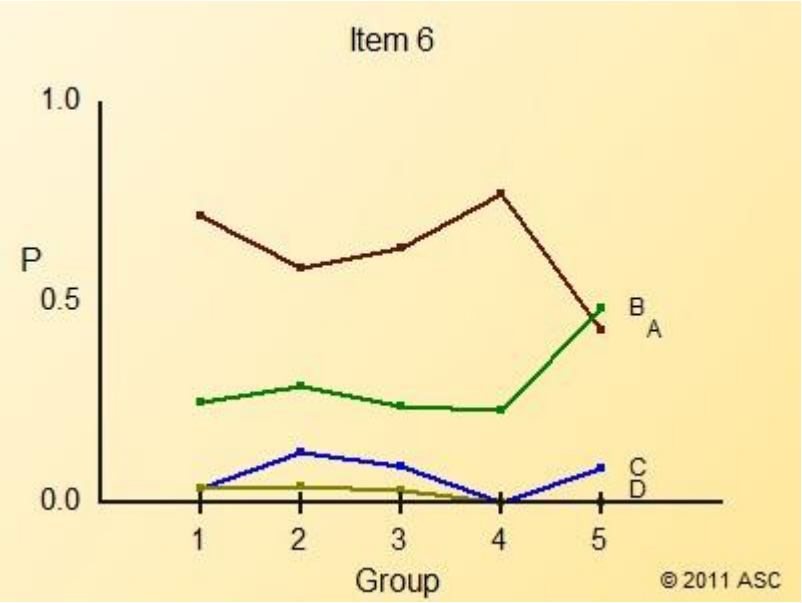


Figure 4.5: Poor discrimination indexes in item 6

N	P	Total Rpbis	Total Rbis	Alpha w/o
150	0.067	-0.006	-0.012	0.750

Table 4.7: Discrimination statistics for item 6

Option	N	Prop[6]	Rpbis	Rbis	Mean	SD	Color	
A	90	0.600	-0.167	-0.212	17.878	5.173	Maroon	
B	45	0.300	0.221	0.292	20.489	6.014	Green	
C	10	0.067	-0.006	-0.012	19.500	6.258	Blue	**KEY**
D	3	0.020	-0.129	-0.373	13.667	4.163	Olive	
Omit	2	0.013	-0.068	-0.227	13.000	2.828		
Not Admin	2				13.000	2.828		

Table 4.8: Distractor statistics for item 6

The coefficient *alpha* of the test item would increase slightly if item 6 were to be removed from the test (see table 4.7, which indicates that the resultant coefficient *alpha* would be 0.750). However, it was decided to refine the item rather than remove it at this stage. Apart from its negative discrimination value, the facility value (P) of the item is very low and only 10 students selected the correct answer. The item was judged to be too difficult and amended as follows:

Original test item:

From the phrase "factoids" trotted out in the first paragraph we can infer that

A. coal-burning does increase the amount of radioactivity in an area.
B. nuclear power is an environmentally safer option to use than coal.
C. coal-power stations should not necessarily be replaced by nuclear ones.
D. uranium and thorium are radioactive elements of no real worth.

6. The proportion of examinees who selected the option.

Refined test item:

From the phrase “factoids” trotted out and the information given in the first two paragraphs, we can infer that the writer of the article

A. agrees that coal-burning is dangerous and should be stopped.
B. thinks nuclear power is a much safer option than coal mining.
C. believes coal-power can generate energy in more than one way.
D. considers uranium and thorium to be elements of notable worth.

4.5.2 The refinement of item 8

Iteman 4.2 provides the following data on item 8:

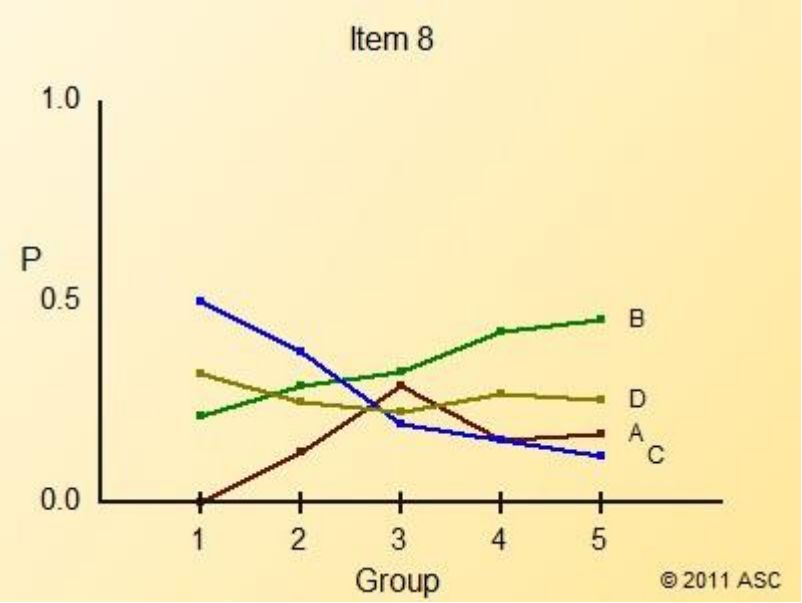


Figure 4.6: Poor discrimination indexes in item 8

N	P	Total Rpbis	Total Rbis	Alpha w/o
150	0.333	0.142	0.184	0.747

Table 4.9: Discrimination statistics for item 8

Option	N	Prop	Rpbis	Rbis	Mean	SD	Color	
A	22	0.147	0.146	0.225	20.273	4.377	Maroon	
B	50	0.333	0.142	0.184	20.460	5.779	Green	**KEY**
C	37	0.247	-0.288	-0.393	15.676	4.750	Blue	
D	38	0.253	0.013	0.018	18.500	5.811	Olive	
Omit	3	0.020	-0.074	-0.215	13.667	4.163		
Not Admin	3				13.667	4.163		

Table 4.10: Distractor statistics for item 8

This item is also too difficult and option A has a higher discrimination value than the answer key. Item 8 was refined as follows:

Original test item:

The title "Rising from the ashes" is associated with

A. cremation and burial.
B. destruction and renewal.
C. nuclear fuel combustion.
D. burnt waste accumulation.

Refined test item:

The title "Rising from the ashes" in the context of the reading passage can best be associated with

A. incineration.
B. combustion.
C. destruction.
D. regeneration.

4.5.3 The refinement of item 16

Iteman 4.2 provides the following data on item 16:

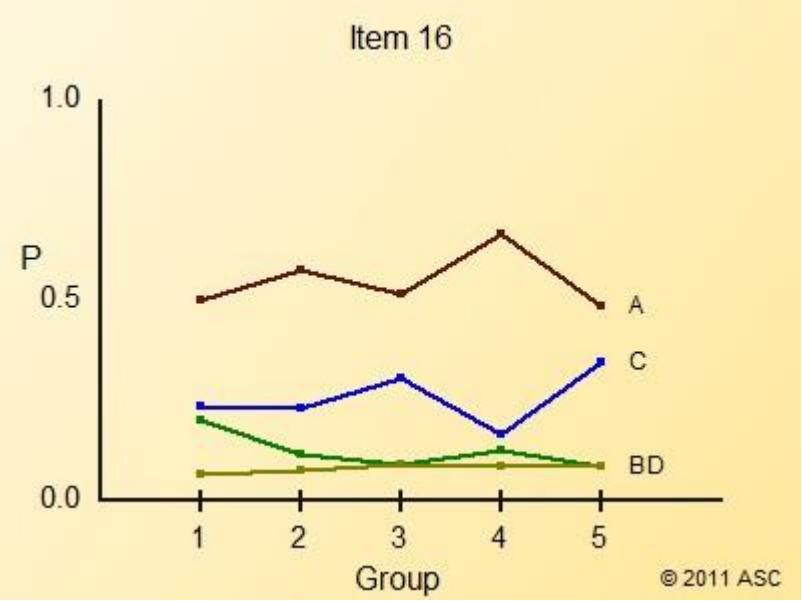


Figure 4.7: Poor discrimination indexes in item 16

N	P	Total Rpbis	Total Rbis	Alpha w/o
150	0.260	0.015	0.020	0.753

Table 4.11: Discrimination statistics for item 16

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
A	80	0.533	0.027	0.034	18.488	5.564	Maroon	
B	18	0.120	-0.094	-0.152	16.944	5.230	Green	
C	39	0.260	0.015	0.020	19.487	6.175	Blue	**KEY**
D	12	0.080	0.039	0.071	19.083	5.248	Olive	
Omit	1	0.007	0.017	0.073	20.000	0.000		
Not Admin	1				20.000	0.000		

Table 4.12: Distractor statistics for item 16

The problem with item 16 is that more than half of the test takers selected A as the correct answer key. This suggests that they understood the meaning of the word profit and how to calculate it, but did not know how to convert profit to

percentage. A postgraduate group of students, however, should be able to do this calculation. The item was thus retained for further piloting, but the word percentage was emphasized in the test question:

> If a kilogram of uranium can be extracted for $54 and sold for $90, what **percentage** of the selling price is profit?

4.5.4 The refinement of item 18

Iteman 4.2 provides the following data on item 18:

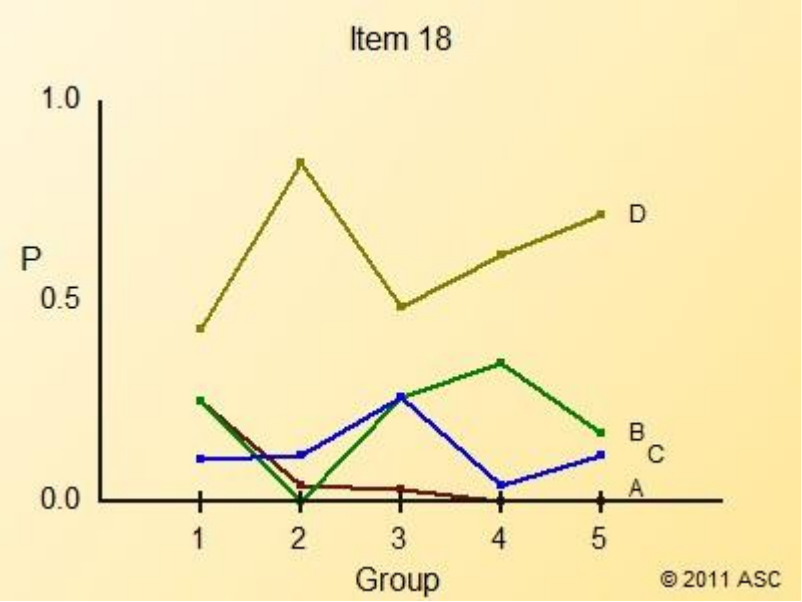


Figure 4.8 Poor discrimination indexes in item 18

N	P	Total Rpbis	Total Rbis	Alpha w/o
150	0.600	-0.005	-0.007	0.754

Table 4.13: Discrimination statistics for item 18

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
A	9	0.060	-0.233	-0.465	13.000	2.236	Maroon	
B	30	0.200	0.124	0.178	19.433	5.575	Green	
C	19	0.127	0.025	0.039	18.421	4.550	Blue	
D	90	0.600	-0.005	-0.007	19.044	5.907	Olive	**KEY**
Omit	2	0.013	-0.048	-0.161	14.500	4.950		
Not Admin	2				14.500	4.950		

Table 4.14: Distractor statistics for item 18

Most of the students selected the correct answer, but because the item produced negative discrimination values, it was decided to omit this question and increase the *alpha* value accordingly. A sequence question was inserted instead, as this was found to be lacking in the text comprehension.

Original test item:

Which is the odd one out in paragraph four?

A. sulphuric acid
B. water
C. nitric acid
D. solution

Refined test item:

Explain the process to obtain uranium in paragraph four by selecting the correct sequence:

A. Add acids, obtain ash, add water, leach uranium.
B. Extract uranium, add ash, make a slurry, dissolve uranium.
C. Add sulphuric acid, mix with water, add nitric acid, produce uranium.
D. Collect ash, add acids, combine with water, extract uranium.

4.5.5 The refinement of item 26

Iteman 4.2 provides the following data on item 26:

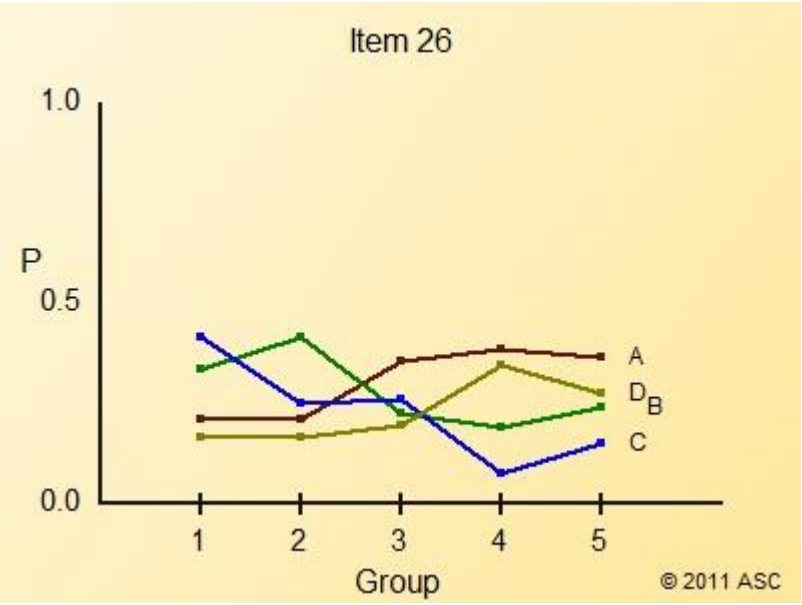


Figure 4.9: Poor discrimination indexes in item 26

N	P	Total Rpbis	Total Rbis	Alpha w/o
150	0.287	0.114	0.151	0.748

Table 4.15: Discrimination statistics for item 26

Option	N	Prop.	Rpbis	Rbis	Mean	SD	Color	
A	43	0.287	0.114	0.151	20.419	6.046	Maroon	**KEY**
B	38	0.253	-0.055	-0.075	17.974	5.485	Green	
C	31	0.207	-0.199	-0.282	16.419	4.952	Blue	
D	32	0.213	0.129	0.182	19.781	5.046	Olive	
Omit	6	0.040	-0.083	-0.190	15.000	5.514		
Not Admin	6				15.000	5.514		

Table 4.16: Distractor statistics for item 26

The reason that this item was flagged is most likely to be the fact that the test takers found all four distractors to be plausible and that this made the item too difficult. The remedy for this was to make the answer key more decisive by changing its wording and altering the order of the remaining distractors.

Original test item:

Based on the remaining information in the text, which of the following is not true?

A. There is a sufficient supply of fuel for future nuclear power stations.
B. Conventional mining is ten times cheaper than its seawater equivalent.
C. Countries might find the extraction of uranium from the ocean affordable.
D. Uranium may be mined for purposes other than of a civilian nature.

Refined test item:

Based on the remaining information in the text, which of the following is not true?

A. Conventional mining is ten times cheaper than its seawater equivalent is.
B. Countries might find the extraction of uranium from the ocean affordable.
C. Uranium can be used for purposes other than of a purely civilian nature.
D. Uranium is easily obtainable as a source for generating nuclear power.

4.5.6 Other refinements

The order of the test items was changed to ensure that test sections start with easy introductory items, so as not to discourage test takers or unnerve them. Moreover, a number of additions had to be made to lengthen the test. Extra test items were added to the sections on understanding texts and academic vocabulary and a section on grammar and text relations was introduced.

Although the items in the section on text types discriminated well in the original test version, some of them had facility values in excess of .8. These were substituted with extracts from more advanced texts.

Test items covering the biographical information of the test takers were included in the final test version for the sake of further research and so as to enable testing for gender bias through differential item functioning (DIF).

4.6 Conclusion

At this point the process of refinement can in no terms be considered to have been completed. The refined test needs to be repiloted amongst a larger cohort of test takers. This is the only way to ascertain whether the refined test items provide an adequate response to the initial statistical analyses. Hereafter the process of undertaking further statistical analyses will recommence and if necessary more refinements will be made. Once the desired *alpha* of at least 0.9 has been obtained and the discrimination and facility indexes reflect well, the test may be considered ready for administration as part of the process to facilitate the selection and placement of postgraduate students. Combinations of the current version of the TALPS and the alternative version also need to be piloted. This is usually the best way to introduce a new test version and ensure a measure of test equivalence.

Chapter 5

The reception of a test of academic literacy for postgraduate students

5.1 Assessing the face validity of an academic literacy test

Assessing the reception of any instrument aimed at measuring a particular ability is essential for a number of reasons. The extent to which a literacy test is received favourably can have a positive influence on the continued use of the test and can provide an indication of the amount of time, effort and resources that need to be invested in further test development. This kind of study will also make a contribution towards the existing data supporting research in the field, especially if one considers that postgraduate academic literacy testing does not have a lengthy history within the tertiary sector of South Africa. A further reason for conducting a reception study is to be found in the fact that research has shown that the attitude of a test taker towards a test can impact negatively on test performance (Bachman & Palmer 1996). It is thus extremely important to be able to create a testing environment that is conducive to optimal performance on the part of the test population, both from the point of view of fairness towards the test takers and validity of the inferences to be made on the basis of the test scores obtained.

A survey questionnaire was designed as a means of testing the face validity of the current version of the TALPS that is being administered at the University of

the Free State. This part of the study is aimed at assessing to what extent the existing version of the TALPS is well received by test takers, especially its ratings in terms of accuracy and fairness. The survey tests the face validity of the TALPS on the basis of the following definition provided by Davies *et al.* (1999: 59).

> The degree to which a test appears to measure the knowledge or abilities it claims to measure, as judged by an untrained observer (such as the candidate taking the test or the institution which plans to administer it).

Another explanation of face validity is the degree to which a test meets the expectations of its users and the test takers and how acceptable the test is deemed to be by its stakeholders (McNamara 2000: 133).

Every attempt must be made to increase face validity. As Butler (2009) points out, supervisors of postgraduate students are unlikely to use an academic literacy test if they perceive it to be irrelevant to the needs of their students and disciplines, and consequently not helpful. As part of the validation process of the test it is thus of importance that the expectations and perceptions of both the test users and test takers be determined. For the current study only data on the perceptions of the test takers were gleaned, since most of the students come from the same academic departments. When more departments make use of the postgraduate test, a reception survey should be conducted amongst the respective supervisors of the postgraduate students. In the interim it may be assumed that the test users perceive the test to be sufficiently credible, or they

would not be subjecting their students to this kind of assessment in the first instance.

A reception survey can also serve as a useful tool to raise awareness of academic literacy difficulties amongst students. The survey questionnaire that has been designed for the current study includes a section which is aimed at making the test takers aware of the complex nature of academic literacy and gives respondents a chance to reflect on their own literacy abilities and how these can impact on their academic progress. If students perceive the test to be a reliable indicator of their current literacy levels, they may be more likely to welcome intervention measures of a supportive nature.

The results of the reception study can be incorporated into future test versions with the purpose of increasing face validity. However, it should also be borne in mind that the test takers are not likely to be authorities in the field of academic literacy, nor mother-tongue speakers of the language of the test – as evident in the case of the survey population – hence the use of the word 'face' value. Such persons may thus not be in an ideal position to assess literacy levels accurately, which may influence the credibility of the responses given in a reception survey. The ideal situation would be for a high correlation to exist between the perceptions of the test takers and the actual test scores obtained. This would certainly contribute towards increased face validity.

5.2 Hypothesis of survey

It was postulated that the test takers who participated in the survey would not be convinced of the need for a test of academic literacy at postgraduate level for a number of reasons. Chan *et al.* (1997) show in a study on reactions to cognitive ability tests that the examinees' performance in an ability test does have an influence on their responses to test reaction items. Thus, when examinees perform poorly in a test, or experience difficulty in completing the test, they may tend to attribute their performance to low face validity of the test:

> Poor performance on a test for which the content is perceived as unrelated to the content of the job is more self-serving (i.e. less ego-threatening) than when test content is perceived as related to the content of the job. Hence test performance should positively affect face validity perceptions" (Chan *et al.* 1997: 302).

The same authors also found that low face validity impacted negatively on test-taking motivation. Should the same test or examination candidates be required to write a similar ability test at a later stage, they may suffer from a lack of motivation to do so and under-perform.

A further reason for the hypothesis that the student population would object to having to undergo a literacy assessment relates to the emotional and psychological stress generally associated with any testing or examination environment. Postgraduate students should therefore not be expected to be too enthusiastic towards the TALPS. Since the test is not content based and

students cannot prepare or be coached for it, they may view this test with suspicion and trepidation. Some may fear that the test scores will be used to deny them access to postgraduate study. Others may fear that the test will not be a true reflection of their potential. Taking into account the stressful nature of the test, its degree of difficulty and time constraints, one can expect the test to have a low face validity amongst the survey respondents.

5.3 Methodology used

Fowler cautions that "the quality of data from a survey is no better than the worst aspect of the methodology" (Fowler 1995: 150). Careful consideration was thus given to which questions should be included in the questionnaire to elicit the desired information. To ensure that the survey would be able to deliver useful information, the questionnaire was first piloted on a small group of students during the second quarter of the academic year, after which certain minor amendments were made to ensure clarity of questions and the final version was prepared for administration during the ensuing quarter.

5.3.1 Survey sample

In order to have as large a sample as possible, survey questionnaires were made available to postgraduate students at the University of the Free State who wrote the TALPS during the course of the third term. Of the 246 students who completed the test during this quarter, 139 filled in survey questionnaires. This

represents 57% of the test population and the survey responses may thus be considered sufficiently representative of the cohort of postgraduate students who wrote the test, as well as representative of postgraduate students from the Faculty of Economic and Management Science in particular.

5.3.2 Questionnaire design

Both qualitative and quantitative data were elicited through the questionnaire, although the emphasis was on the latter form of data. The open-ended question option provided an opportunity for respondents to display their knowledge of academic tasks at postgraduate level and the role played by academic literacy. The answers provided to open questions may in fact give a better indication of how the test takers really felt about the TALPS. Such narrative answers, however, require separate coding and are subject to a measure of unreliability owing to differences in language skills and styles on the part of the respondents. Moreover, diversity of answers also increases the heterogeneity of responses and consequently their analytic value is reduced (Fowler 1995: 179). Careful classification is thus required if such responses are to be of any assistance.

Likert scales were used to enable comparisons to be made across respondents and because these have been shown to be suitable for measuring attitudes and opinions, as well as for evaluating the extent to which users are satisfied with products (Sharp *et al.* 2007: 314). Rating scales also enable better measurement

in terms of simplicity of response and amount of information obtained per question item (Fowler 1995). The ideal is to be able to obtain maximum information from a small set of questions carefully selected for a specific reason, rather than relying on lengthy questionnaires that are too broad in their focus. Questionnaires should further be kept concise to limit respondent burden (see Fowler 1995: 152) in terms of which respondents may lose interest in completing the questionnaire as a result of its length and the many demands made on the survey participants.

5.3.3 Variables excluded

Questions related to socio-economic information were avoided in the survey questionnaire, since these were not considered to be relevant to the survey topic. Although generally speaking students from affluent communities may be expected to have access to more resources than their peers from poorer communities – an aspect which could place them at an educational advantage – the test of academic literacy for postgraduate students is aimed at students who would have been exposed to the same kind of teaching and learning over at least a three-year period of university instruction. This should help to bridge any divide or discrepancy between the respective socio-economic groups, as students would also have had similar access to library and research resources. Naturally some student populations may still be considered to be at a greater disadvantage than others, as a result of differing personal circumstances.

Respondents were also not asked to indicate their race. In the light of South Africa's apartheid past and the sensitivities surrounding academic performance between persons of different racial groups, it was decided to avoid creating any possible link between race and academic literacy. It is the opinion of the researcher that race has nothing to do with literacy and that data pertaining to home language and language of instruction could offer more meaningful information in this respect.

A deliberate decision was also taken not to include questions in the survey on whether the test should be administered in languages other than English. One reason for this was to ascertain to what extent students were conscious of issues of language rights in a multicultural environment without arousing such awareness by means of survey questions. Thus, students of minority or other groups whose mother-tongue languages were not English, but who were aware of the power and political dimensions of language tests, could be expected to describe the TALPS as unfair, since the test only recognizes knowledge of the English language as a literacy tool and does not take into account levels of academic literacy attained in other languages. However, should none of the respondents raise the issue of language, it could be accepted that the students may have undergone a process of assimilation at university in terms of which they had come to accept English as the necessary and dominant language of status and power within academe.

The complete survey questionnaire is included as Annexure C.

5.3.4 Ensuring accuracy of reporting

One way that psychologists have found to improve the accuracy of reporting is to help respondents to recreate the test experience in their minds (see Fowler 1995: 24). This involves incorporating sections from the test in the survey to jog the memories of the respondents. Question 27 on the actual test items serves this purpose in the survey questionnaire. Moreover, accuracy of reporting is also influenced by the span of time that lapses between the administration of the test and the completion of the survey questionnaire: "The more recent the event, the more likely it is to be recalled" (Fowler 1995: 22). For this reason respondents were given the opportunity to complete the survey questionnaire directly after the test. The majority of the respondents were willing to do so.

5.4 Results of the survey

5.4.1 Biographical information

5.4.1.1 Field of study

The overwhelming majority of the respondents, 128 (92.1 %) of the 139, represented the Faculty of Economic and Management Sciences. There were 10 students from the Faculty of the Humanities (7.2%) and one from the Faculty

of Education (0.7%). The explanation for this is to be found in the fact that the TALPS is a new initiative at the University of the Free State and that the Faculty of Economic and Management Sciences is currently the main user of the test. The number of students coming from other faculties is, however, expected to increase considerably and more representative data should be available within the next few years. Most of the respondents were completing their honours (125 or 89.9%). Four (2.9%) were already studying at master's level, eight (5.8%) were completing a postgraduate diploma or certificate course and five (1.4%) were not studying at the time of the survey, but had already completed an undergraduate qualification.

5.4.1.2 Age

In terms of age representation, 87 (68%) of the respondents were between 21 and 23 years of age and 30 (23.3%) were between 24 and 28 years of age. There were five students (4%) in their thirties, four (3.2%) in their forties, one student (0.8%) who was twenty years old and one who was 54 years of age. In eleven cases the respondents did not indicate their ages. The distribution of ages is provided in figure 5.1.

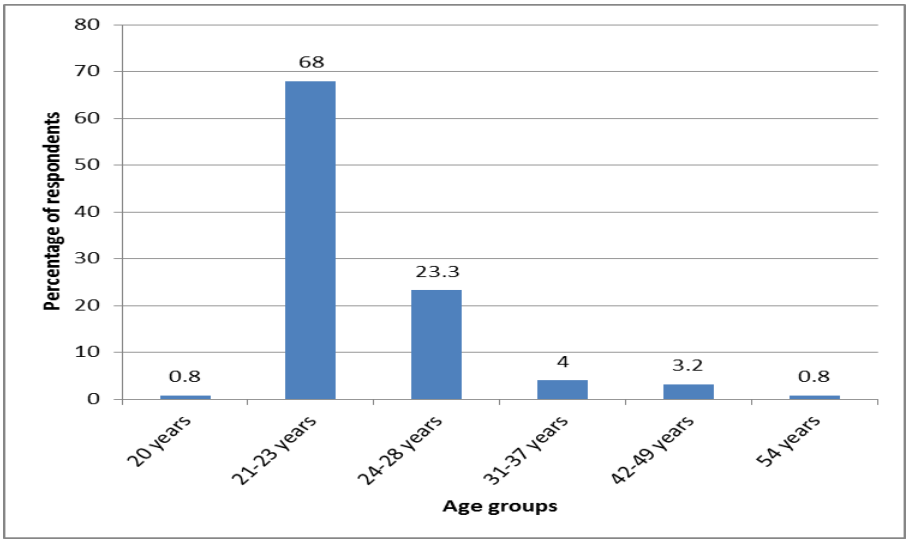


Figure 5.1: Age distribution of respondents

In terms of gender representation, 65.2% of the respondents were female and 34.8% were male.

5.4.1.3 Language diversity

Afrikaans was the home language of most of the participants in the survey (34.8%). This was followed by Sotho (15.9%) and Tswana (10.9%). Only 9.4% of the respondents indicated that English was their home language. Figure 5.2 shows that the cohort of postgraduate students is representative of many diverse language groups.

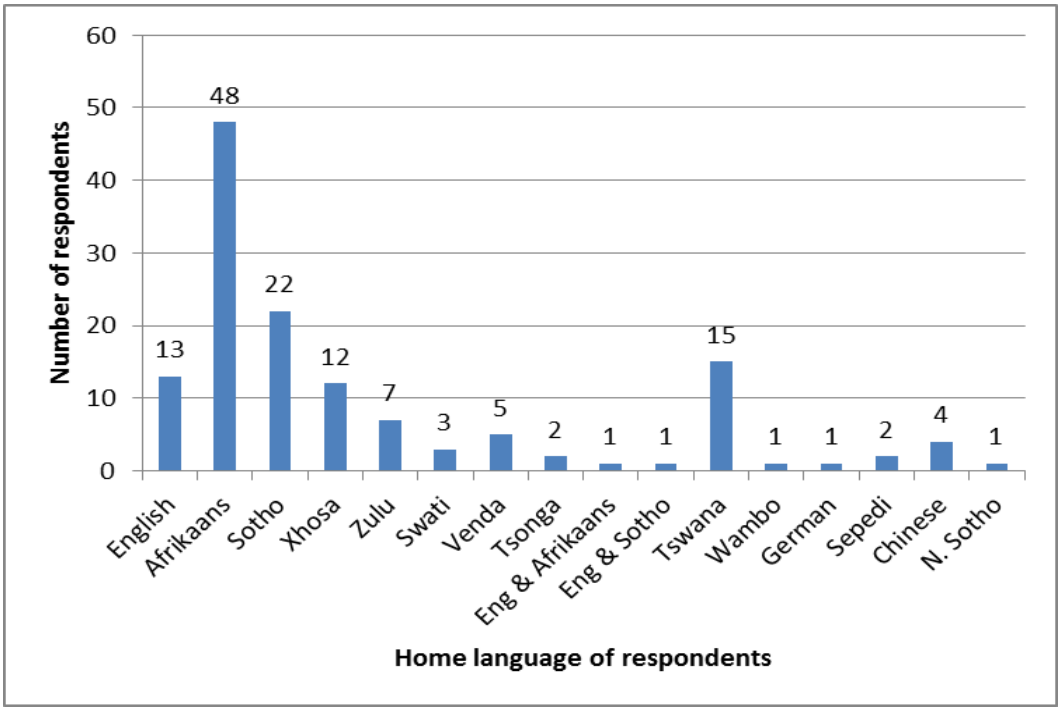


Figure 5.2: Representation in terms of home language

Respondents were also asked to indicate in which languages they had studied at primary and secondary school, and also what their language of instruction for their undergraduate course work at university had been. The information provided can be summarized briefly as follows:

	Primary school	**Secondary school**	**University**
Afrikaans	34.1%	33.3%	21.7%
English	50.7%	61.6%	72.5%
Other	15.2%	5.1%	5.8%

Table 5.1: Language of instruction

From the above it can be seen how English as the language of instruction increases incrementally by around 11% in each of the education phases. If this trend continues, Afrikaans may show a further decline as a language of

instruction at the University of the Free State. Although four of the Afrikaans students objected to the fact that the test advantaged English first-language speakers, only one of the respondents appealed for an Afrikaans version of the test to be developed, a possible indication of assimilation into the hegemony of English.

5.4.1.4 Language development

On the issue of language development, 61.8% of the respondents indicated that they had completed development modules to strengthen their language skills while studying at university. Just under a third of the respondents (28.8%) had taken English as a mainstream subject. The average score obtained for sections 1-7 was 69%, which indicates that the cohort of students showed reasonably high academic literacy levels as far as the ability to engage in critical reading and thinking was concerned. The language modules may have played a role in improving some of the students' academic literacy levels, but at this point it would be unwise to generalize or arrive at any definite conclusion. However, the picture changes when the essay writing section is added (section 8) and at least 35% of the students who participated in the survey study are at risk of not achieving success at postgraduate level in terms of the current risk bands identified by the test developers (ICELDA 2011) as indicated in table 5.2.

Risk associated with level of academic literacy as measured by the Test of Academic Literacy for Postgraduate Students (TALPS)		
Mark	**Code**	**Interpretation**
0-33	1	High risk
34-55	2	Clear risk
56-59	3	Risk
60-74	4	Less risk
75 +	5	Little to no risk

Table 5.2: Risk bands used for the TALPS

The above bands have been based on years of research undertaken by the test developers of the TALL and TALPS and the examination of test scores obtained at different levels of study. Further research is needed to investigate whether the scores obtained in tests such as the TALPS have any predictive ability. At the moment the tests serve to indicate current literacy levels and provide an indication of how well positioned a particular test taker is to negotiate advanced academic material through the medium of English.

5.4.2 Dimensions of face validity

5.4.2.1 Reaction to the TALPS prior to taking the test

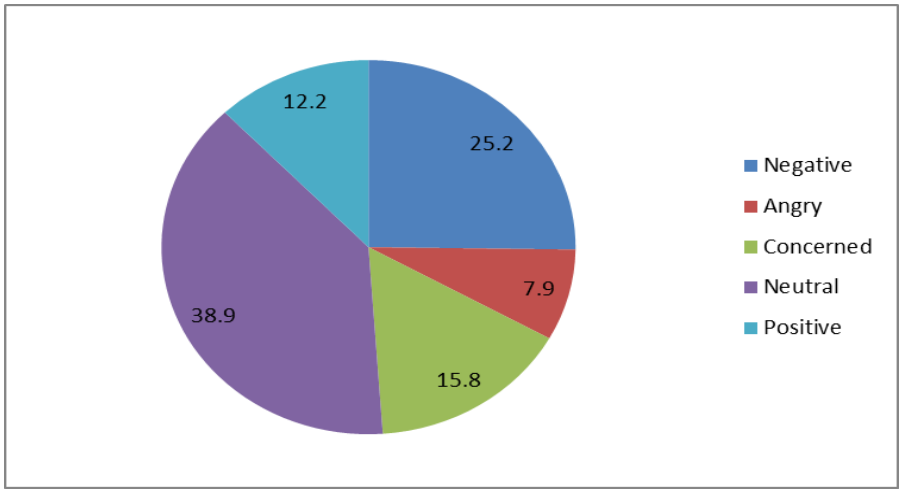


Figure 5.3: Attitude towards writing the test

The half of the pie-chart to the right may be considered indicative of negative sentiments, which means that the response of just less than half of the respondents (48.9%) towards being asked to take the test can be described as being not positive. This may be attributed to the fact that not enough information had been disseminated to the students on the nature of the test prior to its administration. In 87.8% of the cases, the test format had not been discussed with the students at all and only 6.5% of the respondents indicated that they had had access to an example test. This situation can easily be countered by providing sufficient information on the nature and purpose of the test to the students beforehand, and is a lesson that has been learned before: an external evaluation of the Unit for Academic Literacy of the University of Pretoria, for example, occasioned the wide-ranging dissemination of information on the undergraduate version of the Test of Academic Literacy Levels (TALL) some time ago.[7] Moreover, as the number of students required to write the test increases, the status and acceptance of the postgraduate literacy test may also be expected to increase simultaneously.

A total of 71% of the respondents stated that they would prefer to do a computerized version of the test rather than the paper format. An online version of the TALPS was launched in the course of 2011 at the University of the Free State. Preliminary indications are that the test works well for younger and computer literate test takers, such as those who participated in the reception

7. See the report, *Self-evaluation: Unit for Academic Literacy,* issued by the University of Pretoria in 2007.

study, and provided that the test takers have access to stable and sufficient bandwidth.

5.4.2.2 Anxiety experienced during the test

Any test situation generates a certain amount of anxiety, especially when the construct is unrelated to the testing of subject knowledge and students cannot prepare for the assessment. Nearly all of the respondents (92.8%) agreed that anxiety could impact negatively on test performance and a third of the students (31.9%) reported that they had experienced considerable anxiety while taking the test. Of these only 4.3% had described their test anxiety as severe, which may be considered non-representative of the cohort of test takers and negligible.

5.4.2.3 Difficulty of the test

More than half of the respondents described the test as being difficult. However, only a small percentage (6%) stated that the test was very hard. The majority of respondents selected a scale of 3 or 4 to describe the test difficulty (see figure 5.4).

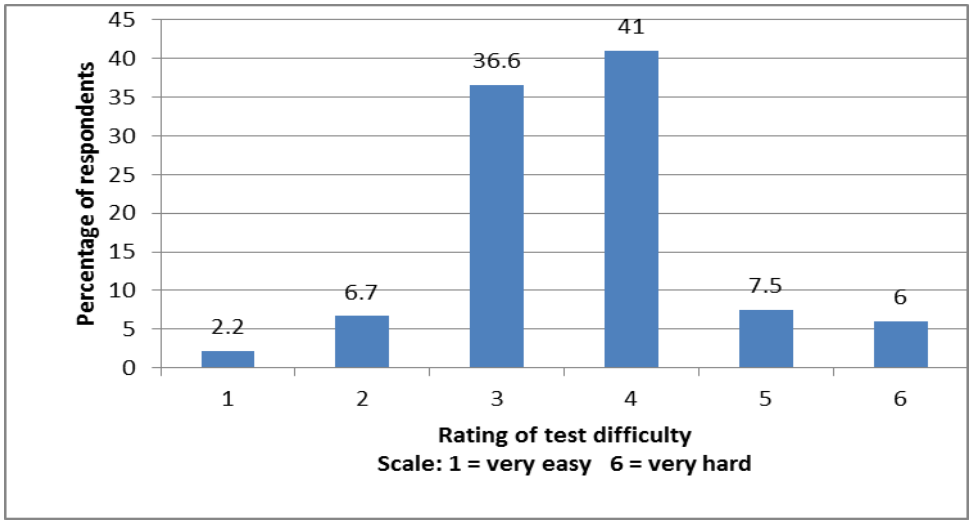


Figure 5.4: Perception of difficulty of test

The above graph resembles the bell curve of a normal distribution of test scores. The fact that students found the test moderately difficult, suggests that the cohort of test takers was strong academically and that the test is pitched at an appropriate level.

5.4.2.4 Time to complete the test

Most of the respondents (59.8%) felt that more time should be allowed for completing the test. This is to be expected, considering the pressurized nature of the test. However, as long as the TALPS continues to maintain its high reliability values and discriminates well between stronger and weaker candidates, there is no need to adjust the time allocated to complete the test. There are no indications from the analyses of test scores available up to now that the amount of time allowed for the test is unfair towards the test takers.

5.4.2.5 Accuracy of the test

The answers provided by respondents in the open sections of the test indicate that students do not necessarily understand the nature of academic literacy and thus tend to confuse the test construct with the four language composites of listening, speaking, reading and writing. Not surprisingly about half (49.7%) of the test takers were skeptical of the test's ability to measure their academic literacy levels (figure 5.5).

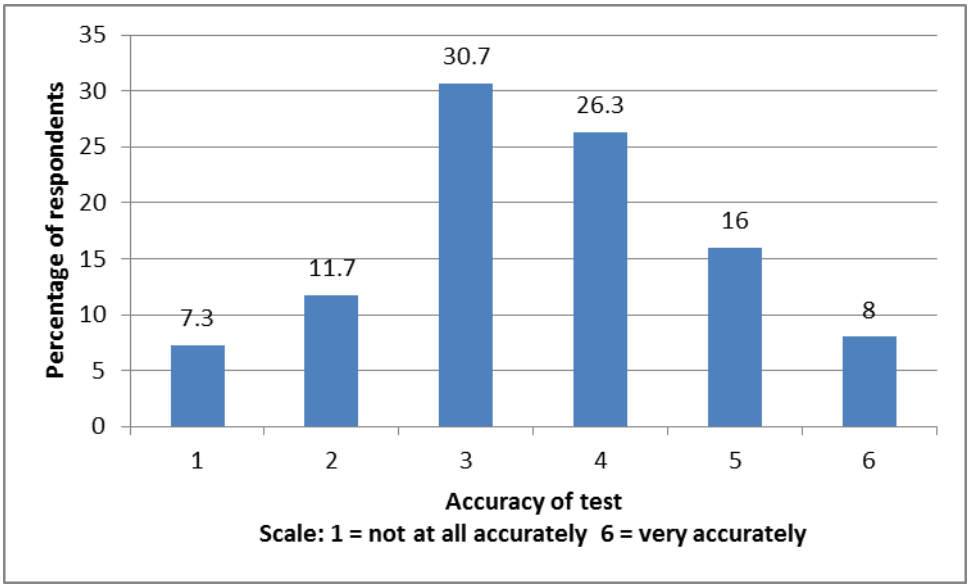


Figure 5.5: Perception of accuracy of test

The fact that about half of the respondents considered the test to be accurate may be viewed positively in the light of the hypothesis that respondents would be negative towards the test and the findings of Chan *et al.* (1997) that poor test performance leads to low face validity. If one takes into account how little information on the test was disseminated to the students before they wrote the test and the negative attitude of nearly half of the respondents towards taking

the test, the responses to the question on the test's accuracy are better than expected. No positive relationship was found between test performance and perception of accuracy. A Pearson correlation revealed a negative relationship of -.056, which in this case runs contrary to the findings of Chan *et al.* (1997: 308) and suggests that the respondents were either more realistic about their literacy levels than was to be expected, or perhaps hopeful that their results would be favourable.

The answers to the question on whether all undergraduate students who wished to register for postgraduate study should be required to write the test, delivered very similar responses to the question on accuracy: 50.4% were in favour hereof and 49.6% against this.

5.4.2.6 Fairness of the test

In a reception study the perception of the fairness of a test is one of the most important aspects. Encouragingly, 69.6% of the respondents considered the test to be fair. The main reason given by those who disagreed that the test was fair, was the fact that students could not prepare for the test and had not been told the content or test topic beforehand (11 comments). This supports the hypothesis that students misunderstand what academic literacy entails and confuse language proficiency with achievement in a content-related subject field. Other responses which may be attributed to a lack of comprehension of

the nature of academic literacy include the following statements gleaned from the open questions in the survey:

- The test should only be written by students who did not study English at school (two comments).
- The test should be administered during Grade 12 (one comment).
- Students doing honours already have an adequate knowledge of English (one comment).
- The test is not relevant to the field of study (one comment).
- The test was a waste of time (one comment).

Another reason related to the above as to why the test could not be considered fair was that it did not measure literacy sufficiently; the respondent failed to clarify what should be measured more adequately.

Six students objected to the fact that English was their second language and that the test was easier for mother-tongue speakers. Of these four were Afrikaans students and two Chinese speaking students. In the case of the Afrikaans students, providing them with an opportunity to do parts of the test or the whole test in Afrikaans would help to establish equity between the English and Afrikaans speaking students. Unfortunately no readily conceivable alternatives exist for the Chinese students, or mother-tongue speakers from other language groups, other than to ensure that they receive sufficient English language support during their undergraduate studies and, preferably, before being admitted to mainstream university courses. One respondent in the survey stated that students came from different backgrounds and as a result the test could not be considered fair. Again, providing academic literacy support at undergraduate level provides a means of addressing some of the imbalances,

but no immediate redress is available in terms of dealing with the socio-economic discrepancies that exist between the different student population groups. Another respondent commented that the test was unfair, because it penalized students for incorrect language usage. This may be considered a valid objection in the instance of students who undertake their postgraduate studies in the medium of Afrikaans and who would be expected to display correct grammatical usage when writing in Afrikaans. Here too the option of doing parts of the test in Afrikaans could resolve this issue.

Further reasons cited by respondents as to why in their opinion the test was not fair, included the following comments:

1. Too little time was allowed for completing the test (three comments).
2. The test caused too much anxiety (two comments).
3. The test was written at the wrong time of the year (one comment).
4. There was no essay writing involved in Accounting (one comment).
5. Not only the Harvard method of referencing should be allowed (one comment).
6. The reading material was not informative enough (one comment).
7. Some test sections were confusing (one comment).
8. The test advantages literacy students (one comment).
9. It was unfair that students should have to pay for the test (one comment).

The first two comments above have already been addressed. On the matter of the most suitable time of the year for writing the test, the specific respondent stated that the test should be written at the beginning of the academic year when students were under less pressure. However, it will not be possible to address this complaint, as most students only apply for postgraduate study in the second half of the year and the academic literacy test is usually administered after application has been made. The fourth objection listed can

be resolved by separating the essay section from the rest of the test and allowing academic departments to prescribe whether their students should do the essay or not. Allowance can also be made in the essay section for different referencing conventions (see fifth comment).

With regard to the sixth and seventh comments, the test developers must ensure that sufficient factual material is provided in the reading texts and graphs to facilitate the writing of an academic essay and that test tasks are set out as clearly as possible. Much refinement has already been undertaken on the test sections dealing with grammar and text relations and the current format used is much clearer than that found in previous versions of the TALL, the predecessor of the TALPS. Only one respondent commented that questions were not clear, which shows that this is not in fact a problem.

The respondent who provided the eighth comment did not elaborate on what was meant by ‘literacy students’. Presumably this would refer to those students who had taken English as a course module. This can hardly be considered a valid objection, since all students have the choice to do language modules as part of their degree courses.

The comment that it was not fair to expect students to pay for the test does not reflect negatively on the test itself, but the manner in which it is administered. All new entrants to the university are compelled to write the NBTs, but the

costs hereof are included in their fees. The situation with the TALPS is similar. However, in the case of the survey respondent who objected to having to pay for the test, the correct application procedure was not followed and this necessitated the payment of a separate fee. The financing of the test is a matter that can be negotiated with the university authorities, especially since more departments are expected to incorporate the TALPS as part of their selection procedures in future.

The following general comments were provided under the open section at the end of the test and not in direct response to the question on the fairness of the test:

1. The test should be free as students already pay tuition fees (three comments).
2. The test content did not cover other subject areas or relate to the respondents' fields of study (three comments).
3. The test should be done in the first year of study, as literacy does not change much and cannot be addressed later at university (two comments).
4. The test was a waste of time and money as the test takers already had undergraduate degrees or High School English (two comments).
5. The test should be written early in the day (two comments).
6. The communication about the test was poor (two comments).
7. Some questions, such as the graphs and text editing, were not clear (two comments).
8. The time allowed for the test was too short (two comments).
9. The test was good (two comments).
10. The test should be done before honours (one comment).
11. The test should be voluntary (one comment).
12. Test takers should be able to use different styles of referencing for the writing section (one comment).
13. The essay section should be removed (one comment).
14. The test should be shortened to one hour (one comment).
15. The reading comprehension passage was too long (one comment).
16. There should be an Afrikaans test (one comment).

All of the above comments have already been discussed in the preceding section. Comments 2, 3, 4, 10 and 11 once again relate to a lack of understanding of academic literacy.

5.4.2.7 Raising awareness of academic literacy

The next part of the survey questionnaire was aimed at increasing awareness of academic literacy by asking respondents to rate their abilities to read with understanding, apply critical thinking and produce written academic texts. This section of the survey was included on the basis of the belief that a better understanding of academic literacy can foster a more positive attitude towards literacy tests. The results can be summarized as follows:

	1 Very poor	**2**	**3**	**4**	**5**	**6 Very good**
Reading with understanding	0.0%	2.2%	9.5%	43.8%	31.4%	13.1%
Critical thinking	0.0%	4.4%	16.2%	43.4%	25.7%	10.3%
Writing academic texts	0.7%	7.2%	35.3%	37.4%	18.0%	1.4%

Table 5.3: Exploring student perceptions of their own academic literacy levels

A clear pattern is observable. Most of the respondents perceived their academic literacy levels to be above average (scale 4) to very good (scale 6) as far as their ability to read with understanding and critical thinking were concerned. They found critical thinking to be more challenging than reading with understanding, but still rated themselves above average. This situation changes when one examines the respondents' perception of their ability to write

academic texts. Here there is a definite shift towards the left side of the scale with almost half of the respondents (43.2%) rating their ability to produce written texts as below average. In actual fact the average score obtained for the writing section was 30%, which shows that the students' perceptions of their ability to produce written academic texts were inflated considerably. Of the 139 students, only 15 (11%) managed to obtain a score of 50% or higher in the essay writing section. The low mark for writing may suggest that a different construct is in fact being assessed in this section of the test (see Moore, Morton and Price 2007).

Students were also asked to rate the level of difficulty of the respective test sections. A summary of their ratings is provided in the graphs that follow.

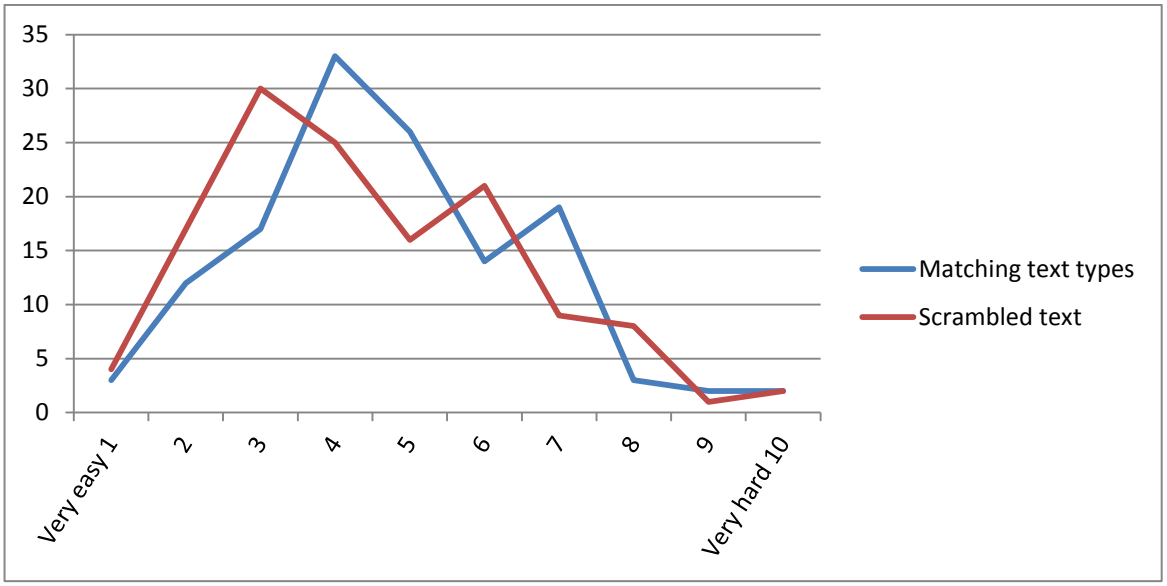


Figure 5.6: Student ratings of sections one and four of the TALPS

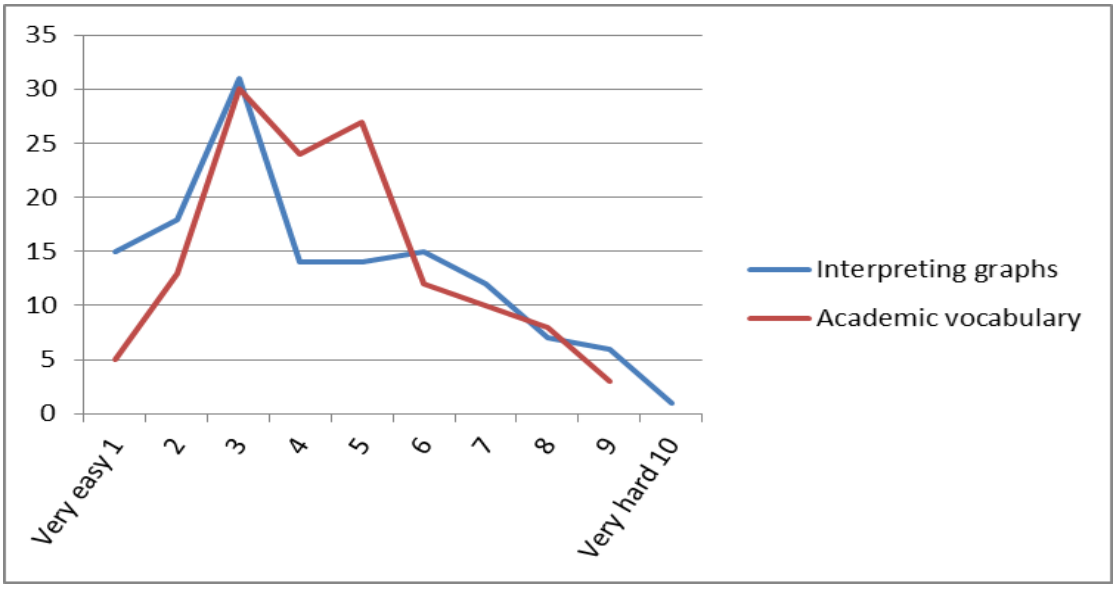


Figure 5.7: Student ratings of sections two and three of the TALPS

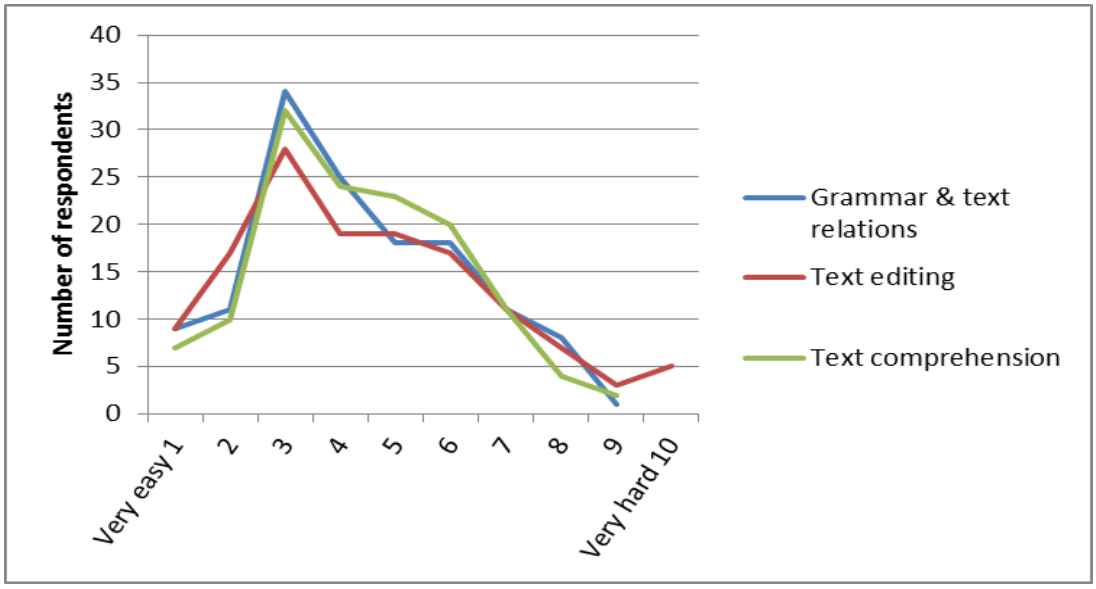


Figure 5.8: Student ratings of sections five, six and seven of the TALPS

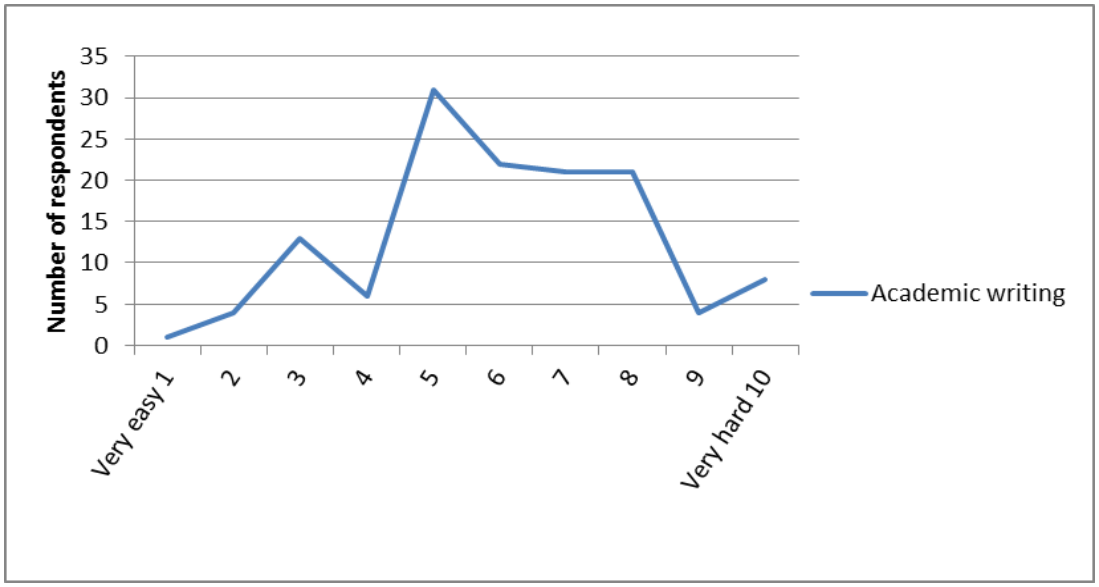


Figure 5.9: Student ratings of section eight of the TALPS

The responses from this section of the survey have been grouped together on the basis of similarities shown in the shapes of the graphs that were produced. All of the above graphs follow similar distributions, with the exception of the graph of the academic writing section (figure 5.9). The difficulty that the respondents experienced in producing an academic essay is clearly evidenced.

Questions 23 to 27 were included in the survey questionnaire for the purpose of raising awareness of the importance of academic literacy and to enable comparisons to be made between different test sections in a further study and will not be discussed in more detail here.

5.4.2.8 Students' perceptions of their own academic literacy

Question 28 of the survey represents an initial attempt to investigate how students go about assessing their own academic literacy levels. More than half of the respondents (54%) stated that they used their undergraduate course marks to benchmark their literacy levels. Only 5% of the respondents understood academic literacy not to refer simply to the ability to use the language skills of listening, speaking, reading and writing and a resounding 82.6% affirmed that the completion of secondary schooling could be seen as an indication of having attained academic literacy. Surprisingly as many as 50.4% of the students did not consider language usage within the tertiary environment to differ from that outside academe. About two-thirds (67.6%) in fact stated that an extensive vocabulary was unrelated to academic literacy. All of these responses further support the postulation that many students have no understanding of a theoretical articulation of academic literacy.

5.5 Conclusion

Although much can be done to increase transparency about the nature and purpose of the postgraduate literacy test, the results of the survey do not support the initial hypotheses that students would be predominantly negative towards the TALPS and that its face validity would be low. Most of the respondents responded positively or neutrally towards being asked to take the test. Moreover, just over half of the respondents considered the test to be accurate in its ability to measure academic literacy. By far the majority of the

respondents (69.6%) viewed the test as fair and at least half agreed that all postgraduate students should be required to write the test. Most of the respondents also rated the test as difficult, which indicates that it was perceived to be set at the right level for postgraduate study, a more challenging level of study compared to the undergraduate course work of the cohort of the students.

Further efforts should nonetheless be made to augment the face validity of the TALPS. This can be achieved through providing more complete information on the format of the test beforehand. An example test is available on both the websites of the University of the Free State and ICELDA, but it seems that students were either not aware of this or did not consult the sample test as part of their preparation for the TALPS. The reason for the administration of the test also needs to be communicated to students, as well as how it will contribute towards the selection procedure of the respective academic departments. The prevailing consensus is that the test should not be allowed to contribute more than 15% in the light of the fact that it is a proficiency test and not an achievement test. However, should evidence unfold that the TALPS has a predictive ability, this percentage may be increased accordingly.

On the matter of the language of testing, the results of the survey suggest that an Afrikaans alternative needs to be developed, even if only a small percentage of students elect not to do the test in English. This would bring the TALPS into alignment with the parallel-medium language policy of the university in terms

of which students may choose to complete their studies in either English or Afrikaans and also provide a means to counteract the hegemony of English. The undergraduate academic literacy test, the TALL, has always had an Afrikaans counterpart, the Toets van Akademiese Geletterdheidsvlakke (TAG).

Further suggestions for test refinement emanating from the reception study include separating the essay section from the rest of the test, ensuring that the text passages in the test contain sufficient information to enable test takers to write an essay without having to rely on general knowledge of the essay topic and incorporating subject-specific questions related to the test-takers' fields of study. The latter aspect would be more difficult to address, since considerably more resources would need to be allocated to develop tests for particular groups of students.

The above discussion indicates that the response towards the TALPS is predominantly positive and that further exploration of literacy assessment at postgraduate level would be a worthwhile investment.

Chapter 6
Conclusions reached

6.1 Review and value of the research

The study corroborates the need for the development of a test of academic literacy for postgraduate students and shows how this can be executed by applying recognized design principles within the framework of applied linguistics. An examination of the test scores of the students who participated in the reception survey shows that at least a quarter of them have low literacy levels that place them at risk of not completing their degree courses. The actual situation is far more dismal if one consults the through-put rates at the University of the Free State over a five-year period (2006-2010). These statistics indicate the proportion of enrolments to graduations in a given year and are considered to be internationally accepted indicators of drop-out and failure rates in the higher education sector (Department of Education 1996). In South Africa only around 15% of university students graduate in the stipulated time for obtaining a degree (Strydom & Mentz 2010: 4). The reasons for this can be manifold, but inadequate academic literacy levels may be expected to feature amongst those. In this respect the research illustrates how a literacy test such as the TALPS can serve as a useful instrument both for identifying areas of weakness in academic literacy and for placing students in suitable postgraduate courses. This kind of test can also be used in conjunction with other indicators of academic success for access purposes.

The mere administration of the postgraduate literacy test can furthermore function as a teaching tool and raise student awareness of the importance of academic literacy. This effect is even more pronounced when the test is accompanied by a survey questionnaire which requires of the respondents to reflect on their own academic literacy levels and come to terms with deficiencies in certain areas of their English language proficiency. Much more, however, needs to be done at undergraduate level to give prominence to the importance of attaining high academic literacy levels.

The design of an academic literacy test for postgraduate students should be viewed as part of the investigation to resolve a language-related problem that exists within the context of an academic institution. By placing the design of an alternative version of the TALPS firmly within the framework of applied linguistics and by drawing on a comprehensive articulation of academic literacy for the blueprint of the test, considerable caution has been taken to ensure that the test construct is valid and that the resultant assessment instrument is theoretically founded and aligned with accepted contemporary testing practices. The definite move away from previous generations of 'scientific' and positivist thinking can be discerned in the kind of test tasks required of test takers and the way in which these emphasize the instrumental communicative function and mediating role of language as a social instrument within the material lingual sphere of academic discourse. Test takers are

required to engage with authentic texts dealing with relevant topics and carry out tasks typically required of postgraduate students, such as interpreting textual information and making meaning beyond sentence level. The emphasis in the test that has been designed falls on critical reading, in accordance with the view that it is the ability to interact with text (see Gee 1998, chapter 5) that provides the best articulation of academic literacy. The theme of the reading texts selected is the need to find viable alternative sources of energy. The subject reading matter forms part of current academic discourse and public debate and is not so discipline specific that test takers would find it inaccessible. Grammar usage is viewed as supportive to argumentative discourse, and not evaluated *per se*. There is thus a definite move towards communicative competence in the focus and formulation of the test items and an avoidance of norm-referenced testing on the basis of first-language knowledge of content.

An analysis of test responses can provide valuable insight into those areas of academic literacy that need development and whether any trends can be detected over a period of time. This in turn affords opportunities for the design of remedial interventions such as literacy courses, which may be offered at undergraduate and postgraduate levels. There is clearly a need to develop critical thinking skills in undergraduate course work, because many students who participated in the study indicated that they battled in this area. They also lacked writing skills, which suggests that these students could benefit from

language proficiency modules linked to their mainstream subjects. Once again it is the definitive element of design in applied linguistics that features prominently on both the side of literacy assessment and literacy instruction.

Note should be taken of the potentially negative role of test washback, which is generally defined as the effect that language testing has on actual language teaching and learning, as well as on educational practices, beliefs and systems (see Bachman & Palmer 1996: 30-31). This can be detrimental when tests become a means of prescribing curriculum and pedagogy. Like its predecessor, the alternative version of the TALPS is able to avoid any such manipulative effect on lecturers and students by being ability based rather than dependent on discipline-related content.

6.2 Meeting the constitutive and regulative conditions

Of central importance in the design of a literacy test is adherence to the constitutive and regulative conditions necessary for the process of validation. One of the limitations of the current study is that it has not been possible to replicate the full test amongst large cohorts of students to fulfil the validation process. It would have been irresponsible to unleash a newly developed test that has not attained a sufficiently high reliability index on a group of unsuspecting test takers, when the inferences made on the basis of the scores may disadvantage some of these students. More time is needed to re-pilot the test in non-threatening testing contexts before attempting to describe it as a

valid and reliable indicator of the academic literacy levels of postgraduate students. It should be borne in mind that the process of validation continues beyond the administration of the test and that it has both objective and subjective elements. Whereas the objective pertains to the multiple sources of evidence gleaned through the assessment instrument, the subjective refers to the interpretations, inferences and decisions based on the test scores and their social and political consequences. The test designer may have very little control over the latter, but can endeavour to monitor the transparency and fairness of the administration of the test by conducting quantitative and qualitative research in the form of surveys and interviews. This can help to ensure that test scores are used in a responsible and accountable manner (cf. Rambiritch, in preparation).

Rather than ascribing to the unitary view of construct validity propagated by Messick, this study draws on the concept of test usefulness expounded by Bachman and Palmer and articulates it in more definite terms through the constitutive and regulative conditions identified by Weideman. In effect Bachman and Palmer's notion of test usefulness is also of a unitary nature and in that sense can be just as problematic as the conflated view of Messick. Rather than emphasizing the overall usefulness of a test, as Bachman and Palmer suggest, or the all-encompassing construct validity that Messick seeks, the emphasis should fall on the individual qualities that define a test. A test may be weak in terms of one particular dimension, practicality for example, but

still be a valid and useful instrument. Note should be taken of the fact that the respective dimensions are mutually related and that none can be considered to be absolute.

The initial indications are that the test which has been designed for the present study meets the constitutive condition of reliability. Although the test was piloted on a small group of students, the internal consistency was acceptably high for a class test, with an *alpha* reliability of 0.748. The actual reliability may be expected to increase considerably in subsequent test administrations amongst larger groups of students and following the introduction of a number of refinements to test items. In addition to increasing the reading grade level and reducing the reading ease index of the main comprehension passage, test items with low facility or discrimination values in the pilot test version were amended or substituted. The impact hereof should be evident in the next administration of the test. Different weightings for the scoring of the test will only be allocated to the test items once the test has been re-piloted and the most productive items with high point-biserial correlations have been identified.

The construct validity of the refined test should be high, as the test employs the same construct used in the existing version of the TALPS. Usually test construct validity is reflected through the correlations between the subtests. However, in order to evaluate the subtest correlations using statistical

programmes such as Tiaplus, larger cohorts of test takers are recommended, something that was not possible in the present study.

The statistical analysis of the class test showed a negatively skewed distribution of test scores with a kurtosis of -0.184, which is well within the parameters required for a normal distribution. The low mean of 45% for the test explains the skewness of the distribution and confirms that the test was too challenging for the group of undergraduate students, but should be appropriate for a cohort of graduates and able to provide those additional sources of evidence needed to meet the constitutive conditions for the test.

With regard to the regulative conditions for an applied linguistics artefact, the construct used for the alternative test is substantiated by a theory of language that adopts an open and enriched view of language and makes a distinction between language skills and literate behaviour. The blueprint for the test – the articulation of the lingual dimension within the leading technical qualifying function of the test – shows that academic literacy skills cannot be compartmentalized and should not be measured in isolation, since more than one component of this ability is employed simultaneously when accessing a text. In the same vein, executing the test tasks requires the simultaneous application of language knowledge and metacognitive processes.

Attempting to meet all of the technical conditions that regulate a test can be rather challenging. In terms of current thinking in critical language testing, the social dimension requires that the test be administered in a context that would enable test takers to interact with the reading material as they would under normal non-testing circumstances. Eliminating unnecessary test anxiety and ensuring an optimal testing environment are thus of paramount importance. These can be achieved by giving careful consideration to the time scheduled for the writing of the test and by providing ample information beforehand on the format of the test and the kinds of test tasks that test takers will be expected to perform. Hereby the condition for transparency, which probably forms part of the juridical dimension of the test experience, can also be met.

The aspect of test fairness – another integral part of the juridically analogous moments within the technical sphere, that generate criteria or conditions for responsible test design – will always be subject to the legitimacy of the decisions and interpretations derived from test scores. Nonetheless, with the design of the current test an attempt has been made to exclude any unfair or unreasonable test tasks and to ensure alignment or harmony (the analogically aesthetic dimension of the test) with language-related tasks in a postgraduate environment.

Test designers also have a duty to make tests as economical as possible, both in terms of time constraints and resources available. Since the alternative version

of the TALPS that has been designed for the purposes of this study contains a writing section that requires manual marking, it is suggested that a separate fee perhaps be levied for this section to cover the costs of the marking. The current fee per test is already very low (R80) compared to tests offered by testing organizations such as the Educational Testing Service (ETS) and International English Language Testing System (IELTS), which can cost as much as R2000 per test. Moreover, the separation of the critical thinking and reading sections from the writing section of the test can have the further advantage that academic departments can choose whether their candidate students should complete all of the sections, or leave out the writing section. One of the criticisms levelled against the current version of the TALPS in the reception study was that it required all students to write an academic essay, even when subject fields such as Accounting did not require that. Further research is needed to determine whether the inclusion of writing sections in tests of academic literacy impacts negatively on test fairness and construct validity on the basis of the studies conducted by scholars such as Moore, Morton and Price (2007).

Ethical considerations in language testing also bear a measure of subjectivity. The main consideration here is that an academic literacy test for postgraduate students such as the TALPS should only be employed for the purpose for which it has been designed – the assessment of the current literacy levels of the test takers and the degree to which these pose a risk. Attempting to use the

assessment instrument as an achievement test would be an abusive act. The possibility cannot be ruled out that a student with relatively low initial literacy levels may have considerable latent potential and develop into a strong postgraduate student further down the line. The objective of the newly designed test is to be beneficent in its approach and serve both the interests of the test users and test takers. This can best be accomplished through ensuring compliance with each of the constitutive and regulative conditions for language testing, but these culminate in considerations of the technical care and concern that the test design and the application of the test as measuring instrument have for those persons who are subjected to being assessed by it.

6.3 Recommendations based on the reception study

The results of the reception survey show that the version of the TALPS being administered at the University of the Free State has an acceptable face validity for a recently introduced assessment tool. More than half of the survey respondents considered the test to be accurate and more than two-thirds stated that it was fair. Because the survey population was not representative of the general postgraduate sector, it is suggested that the reception study be continued until many more faculties and departments have been included. The perceptions of the test users also need to be ascertained, as well as the extent to which the TALPS or its alternative version are serving the needs of the academic departments. A similar survey or interview process can be used to garner the necessary information. This should be considered part of the

validation process of any new test. Furthermore, a reception study also provides a means of monitoring how tests are being administered and whether the regulative conditions are being adhered to. In order to ensure transparency attention also needs to be given to the amount of information provided to test takers prior to writing the test.

From the reception study it is further clear that many students have a restricted view of language and an erroneous perception of academic literacy. They fail to understand the important role played by language proficiency in all fields of study and how high academic literacy levels can contribute towards academic achievement and student success. This is a matter that needs addressing from the onset of tertiary study and should not be left until postgraduate level. The design of effective literacy interventions remains an area of concern, especially in the light of the constraints on time and resources that are available. Once tests such as the TALPS start to show a steady increase in the literacy levels of postgraduate students, it may be possible to infer that progress is being made at undergraduate level in addressing literacy challenges. In the meantime the integration of literacy development modules, effective pedagogical approaches and mainstream course modules needs exploration. Whether sufficient resources will ever be available to allow subject-specific literacy interventions remains to be seen. However, as long as literacy modules are viewed by students as ‘additional’ subjects unrelated to their mainstream courses and needed for credit purposes only, they may not invest sufficient effort and time

into developing their literacy levels and in this way may undermine the potential success of literacy interventions in the form of proficiency courses.

Although a number of survey respondents called for content-related testing relevant to their own fields of study, more research is needed on the desirability and feasibility of designing subject-specific tests. Apart from the additional resources that will be needed, such kinds of tests may evolve into the assessment of knowledge of subject content rather than the evaluation of academic literacy. Whereas it may be desirable to integrate subjective-specific literacy modules and mainstream course instruction, it may not be desirable to integrate mainstream course content and literacy assessment, as this could lead to a situation of 'teaching for testing', which is to be avoided. A second problem, of course, relates to how one would ensure test equivalence, should there be a number of subject-specific or discipline-oriented tests, and whether solving that complex problem will be worth the effort it will no doubt require.

The reception study also shows that more consideration needs to be given to the language of testing. A point of criticism raised by some of the respondents in the student survey was that the TALPS only recognized knowledge of the English language as a literacy tool and did not take into account levels of academic literacy that had been attained in other languages. The fact that 90.6% of the respondents had to write the test in a language that was not their mother tongue, but that only six students objected to this, suggests that many

students may have undergone a process of cultural assimilation in which they have come to view English as the powerful *lingua franca* of academe. The survey study also revealed that English as the language of instruction at university tended to increase incrementally by around 11% with each respective phase of education, i.e. from primary to secondary to tertiary level, while other languages showed a decline in usage.

Note should be taken of the fact that it takes several years to reach a level of proficiency adequate for tertiary study in a language that is not the first language of a student and that three or four years of instruction at university level may not be sufficient to enable a student to reach the required language ability to operate within the confines of academic language. A gifted student may under-perform in a literacy test as a result hereof. Having tests available in other languages where practically possible would therefore not only be more democratic, but could help to bring clarity to the assessment of a prospective postgraduate student. Ensuring test equivalence (see below) may in this case also be less problematic than in the case of subject-oriented or discipline-specific tests. However, making a test such as the TALPS available in English and Afrikaans will not be a sufficient remedy on its own. Adequate resources should be available to provide the necessary support to postgraduate students who experience language difficulties.

Shohamy (2006) draws attention to the manipulative role that language tests can play in determining which languages are given priority in society and education. In a multicultural country such as South Africa this is a particularly salient issue with a number of critics accusing the powers that be of making English the language of domination and hegemony contrary to the noble intentions of the constitution of the country which encourages language diversity and equality of status between the respective official languages. Since English and Afrikaans are currently the official languages of instruction at the University of the Free State and fully meet the requirements for being considered languages of academe, literacy assessment practices need to be brought into alignment with the language policy of the university. It is thus proposed that sections 1 to 5 of the academic literacy test for postgraduate students be retained in English, because most of the academic material available across all subject fields is accessed through the medium of English, irrespective of mother-tongue or first language. Sections 6 to 8, however, can easily be translated into Afrikaans. Postgraduate students who intend pursuing their studies through the medium of Afrikaans will then be afforded the opportunity to display their knowledge of Afrikaans grammar and textual relations and to show their proficiency in producing a written text in Afrikaans. A glossary containing the Afrikaans equivalents of technical terms used in the test can be provided to ensure that the test takers who are electing to write in Afrikaans do not have a vocabulary disadvantage compared to those students who are writing in English and who can draw on the terminology provided in

the reading texts. The design of a ‘bilingual’ type of test that incorporates English and Afrikaans in the proposed manner would be more in line with the target language use encountered at postgraduate level within the South African tertiary environment, and thus more authentic than the design of a full Afrikaans test version.

In summary, the research study shows that the design of a test of academic literacy at postgraduate level is a viable enterprise that presents several positive outcomes. Not only does it help to identify students at risk and assist them to gain an understanding of their current academic literacy levels and English language proficiency, but it can also be helpful to the tertiary institution in respect of resource allocation and enrolment planning. Further to this, the study confirms that it is possible to graduate from a tertiary institution and be admitted to postgraduate study with low levels of academic literacy, a matter that raises questions about assessment practices and teaching and learning at undergraduate level – part of the washback effect of language testing.

The amount of time invested in the design and refinement of a literacy test makes it clear that no one particular ‘best’ test exists for a specific language testing situation. Those who are of the opinion that any single model exists, fail to take into account that the complex and dynamic processes involved in language learning and literacy development are not the same for all learners, situations and purposes. There will thus always be an urgent need to design

more tests and to keep researching the effects of different task types. Just as applied linguistics has gone through different paradigms and traditions, so language testing may be expected to reflect alternating designs and approaches attuned to the prevailing needs of the tertiary sector.

Bibliography

Angoff, W.H. 1988. Validity: An evolving concept. In Wainer, H & Braun, I.H. (Eds). *Test validity*. New Jersey: Laurence Erlbaum Associates Inc. Pp. 19-32.

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

Bachman, L.F. 2004. *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

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

Bachman, L. &. Purpura, J.E. 2010. Language assessments: Gate-keepers or door-openers? In Spolsky, B. & Hult, F.M. *The handbook of educational linguistics*. Chichester, West Sussex: Wiley-Blackwell. Pp. 456-468.

Bhorat, H. & Oosthuizen, M. 2008. Determinants of Grade 12 pass rates in the post-apartheid South African schooling system. *Journal of African Economies*, 18(4): 634-666.

Blanton, L.L. 1994. Discourse, artefacts and the Ozarks: Understanding academic literacy. *Journal of second language writing*, 3(1): 1-16.

Butler, G. 2009. The design of a postgraduate test of academic literacy: Accommodating student and supervisor expectations. *Southern African linguistics and applied language studies*, 27(3): 291-300.

Carroll, J.B. 1968. The psychology of language testing. In Davies, A. (Ed.). *Language testing symposium: A psycholinguistic approach*. London: Oxford University Press. Pp. 46-69.

Carroll, J.B. 1993. *Human cognitive abilities: A survey of factor analytic studies.* Cambridge: Cambridge University Press.

Chan, D., Schmitt, N., DeShon, R.P., Clause, C.S. & Delbridge, K. 1997. Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of applied psychology*, 82(2): 300-310.

CITO. 2005. *TiaPlus, classical test and item analysis.* Arnhem: Cito Measurement and Research Department.

Cliff, A. & Hanslo, M. 2009. The design and use of 'alternate' assessments of academic literacy as selection mechanisms in higher education. *Southern African linguistics and applied language studies*, 27(3): 265-276.

Corder, S.P. 1973. *Introducing applied linguistics*. London: Penguin.

Coxhead, A. 2000. A new academic word list. *TESOL quarterly*, 34(2): 213-238.

Coxhead. A. 2011. The academic word list 10 years on: Research and teaching implications. *TESOL Quarterly*, 45(2): 355-362.

Crystal, D. 1981. *Directions in applied linguistics.* London: Academic Press.

Crystal, D. 1987. *The Cambridge encyclopedia of language.* New York: Cambridge University Press.

Currie, M & Chiramanee, T. 2010. The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language testing*, 27(4): 471-491.

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

Department of Education. 1996. Green paper on higher education transformation.

[Available online: http://www.info.gov.za/greenpapers/1996/highereduc.html] (Accessed on 2012-04-10).

Directorate for Institutional Research and Academic Planning (DIRAP). 2011. Through-put rates per faculty (HEMIS). Bloemfontein: University of the Free State.

Flower, L. 1990. Negotiating academic discourse. In Flower, L., Stein, V., Ackerman, J., Kantz, M.J., McCormick, K. & Peck, W.C. *Reading-to-write: Exploring a cognitive and social process*. New York: Oxford University Press. Pp. 221-252.

Fowler, F.J. (Jnr). 1995. *Improving survey questions.* Thousand Oaks, California: Sage Publications Inc.

Gee, J.P. 1998. What is literacy? In Zamel, V. & Spack, R. (Eds). *Negotiating academic literacies: Teaching and learning across languages and cultures.* Mahwah, New Jersey: Lawrence Erlbaum Associates Inc. Pp. 51-59.

Geldenhuys, J. 2007. Test efficiency and utility: Longer and shorter tests. *Ensovoort*, 11(2): 71-82.

Gronlund, N. & Linn, R.L. 1990. *Measurement and evaluation in teaching.* Sixth edition. New York: Macmillan.

Guilford, J.P. & Fruchter, B. 1978. *Fundamental statistics in psychology and education.* Sixth edition. New York: McGraw-Hill Book Company.

Guyer, R. & Thompson, N.A. 2011. *User's manual for Iteman 4.2*. St Paul Minnesota: Assessment Systems Corporation.

Hall, C.J., Smith, P.H. & Wicaksono, R. 2011. *Mapping applied linguistics: A guide for students and practitioners*. New York: Routledge.

Hardyck, C. & Petrinovich, L.F. 1975. *Understanding research in the social sciences: A practical guide to understanding social and behavioural research*. Philadelphia: W.B. Saunders Company.

Hirvela, A. (2004). *Connecting reading and writing in second language writing instruction.* Ann Arbor: University of Michigan Press.

Hogan, T.P. 2007. *Psychological testing: A practical introduction.* Second edition. Hoboken, New Jersey: John Wiley & Sons. Pp. 149-150.

Hunter, J.E. & F.L. Schmidt. 2000. Racial and gender bias in ability and achievement tests. *Psychology, public policy, and law,* 6(1): 151-158.

ICAS (Intersegmental Committee of the Academic Senates). 2002. *Academic literacy: A statement of competencies expected of students entering California's public colleges and universities.* [Available online: http://www.calstate.edu/AcadSen/Records/Reports/Academic_Literacy_Final.pdf] (Accessed on 2011-01-29).

ICELDA (Inter-institutional Centre for Language Development and Assessment). 2011. Language test development and research. [Available online: http://www.icelda.sun.ac.za/]. (Accessed on 2011-10-17).

Kaplan, R.B. (Ed.). 1980. *On the scope of applied linguistics.* Rowley, Massachusetts: Newbury House.

Kline, T.J.B. 2004. Gender and language differences on the test of workplace essential skills: Using overall mean scores and item-level differential item functioning analyses. *Educational and psychological measurement*, 64(3): 549-559.

Kumaravadivelu, B. 2006. *Understanding language teaching: From method to postmethod.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Lazaraton, A. 2010. Utilizing qualitative methods for assessment. In Shohamy, E. & Hornberger, N.H. (Eds). *Encyclopedia of language and education, second*

edition, volume 7: Language testing and assessment. New York: Springer. Pp. 197-209.

Loan Le, P. 2011. *Assessing academic literacy of first year Vietnamese students: How appropriate is the TALL?* Unpublished master's dissertation. Groningen: Rijksuniversiteit Groningen.

National Benchmark Tests (NBTs). [Available online: http://www.nbt.ac.za/] (Accessed on 2011-01-29).

McNamara, T. 2000. *Language Testing.* Oxford: Oxford University Press.

McNamara, T.F. 2003. Looking back, looking forward: Rethinking Bachman. *Language testing*, *20*(4): 466-473.

McNamara, T. 2005. Introduction. In Hinkel, E. (Ed.). 2005. *Handbook of research in second language teaching and learning.* Mahwah, New Jersey: Lawrence Erlbaum Associates. Pp. 775-778.

McNamara, T. & Roever, C. 2006. *Language testing: The social dimension.* Language Learning Monograph Series. Language Learning Research Club, University of Michigan: Blackwell Publishing.

Messick, S. 1980. Test validity and the ethics of assessment. *American psychologist*, 35: 1012-1027.

Messick, S. 1988. The once and future issues of validity: Assessing the meaning and consequences of measurement. In Wainer, H. & Braun, I.H. (Eds). 1988. *Test validity*. New Jersey: Lawrence Erlbaum Associates Inc. Pp. 33-45.

Moore, T. Morton, J. & Price, S. 2007. *Construct validity in the IELTS reading test: A comparison of reading requirements in IELTS test items and in university study*. IELTS Research Reports Volume 11, Report 4. [Available online: http://www.ielts.org/pdf/Vol11_Report_4_Contruct%20validity%20in%20the%20IELTS%20Academic%20Test.pdf] (Accessed on 2011-09-11).

Oates, B.J. 2006. *Researching information systems and computing*. London: Sage.

Paltridge, B. & Phakiti, A. (Eds). 2010. *Continuum companion to research methods in applied linguistics.* London: Continuum International Publishing Group.

Rambiritch, A. 2012. *Accessibility, transparency and accountability as regulative conditions for a post-graduate test of academic literacy.* PhD in preparation, University of Pretoria.

Ratangee, N. 2007. *Academic literacy, the PTEEP, and the prediction of academic success*. Unpublished master's dissertation. Johannesburg: University of the Witwatersrand.

Sealey, A. & Carter, B. 2004. *Applied linguistics as social science.* New York: Continuum.

Sharp, H., Rogers, Y. & Preece, J. 2007. *Interaction design*. 2nd edition. Chichester, England: John Wiley & Sons Ltd.

Shohamy, E. 1984. Does the testing method make a difference? The case of reading comprehension. *Language testing*, 1(2): 147-170.

Shohamy, E. 2006. *Language policy: Hidden agendas and new approaches*. New York: Routledge.

Spolsky, B. 1997. The ethics of gatekeeping tests: What have we learned in a hundred years? *Language testing*, 14(3): 242-247.

Spolsky, B. & Hult. F.M. (Eds). 2010. *The handbook of educational linguistics.* Chichester, West Sussex: Wiley-Blackwell.

Strydom, J.F. & Mentz, M. 2010. Focusing the student experience on success through student engagement. South African Survey of Student Engagement (SASSE) report. Pretoria: Council on Higher Education (CHE).

Truscott, J. 1996. Review article: The case against grammar correction in L2 writing classes. *Language learning*, 46(2): 327-369.

Van der Slik, F. 2009. Gender bias and gender differences in two tests of academic literacy. *Southern African linguistics and applied language studies*, 27(3): 277-290.

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. 2007. Testing academic literacy over time: Is the academic literacy of first year students deteriorating? *Ensovoort*, 11(2): 126-137.

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. *Southern 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.J. 2010. *Konstitutiewe voorwaardes vir die ontwerp van 'n toets van akademiese geletterdheid.* Unpublished PhD. Bloemfontein: University of the Free State.

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.

Weideman, A. 1987. *Applied linguistics as a discipline of design – a foundational study*. Unpublished D.Litt. Bloemfontein: University of the Orange Free State.

Weideman, A. 1988. *Linguistics: A crash course for students*. Bloemfontein: Patmos.

Weideman, A. 2002. *Designing language teaching: On becoming a reflective professional.* Pretoria: University of Pretoria.

Weideman, A. 2003a. *Academic literacy: Prepare to learn.* Second edition. Pretoria: Van Schaik.

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

Weideman, A. 2003c. Justifying course and task construction: Design considerations for language teaching. *Acta academica* 35(3): 26-48.

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 in a task-based approach. *Language matters*, 37(1): 8-101.

Weideman, A. 2007. The redefinition of applied linguistics: Modernist and postmodernist views. *Southern African linguistics and applied language studies*, 25(4): 589-605.

Weideman, A. 2009a. Constitutive and regulative conditions for the assessment of academic literacy. *Southern 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. 2012. Academic literacy tests: Design, development, piloting and refinement. Forthcoming in a special edition of the SAALT *Journal for Language Teaching* in honour of Prof. Johann van der Walt.

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.

Widdowson, H.G. 1984. *Explorations in applied linguistics 2.* Oxford: Oxford University Press.

Yeld, N. *et al.* 2000. The construct of the academic literacy test (PTEEP). Mimeograph. Cape Town: Alternative Admissions Research Project, University of Cape Town.

Abstract

A number of studies have indicated that the literacy levels of students at tertiary institutions in South Africa are lower than required for academic success. Nonetheless, the possibility exists that some of these students may in fact manage to graduate with low academic literacy levels, and even be admitted to postgraduate study although they lack the ability to negotiate the advanced academic discourse required to succeed at that level. This study confirms the need for the development of a test of academic literacy for postgraduate students and shows how this can be accomplished by applying recognized design principles within the framework of applied linguistics. The constitutive and regulative conditions for the validation of language tests are dealt with as the necessary foundation for the design of a test of academic literacy for postgraduate students. Whereas language tests have conventionally been required to show validity and reliability, in terms of more contemporary thinking such tests must also possess what is referred to as consequential validity, a notion that refers to the impact of tests. Literacy tests should thus not only be consistent and theoretically justifiable, but should incorporate a multiplicity of evidence to back up their validation.

The various phases involved in the design of a test are covered, including piloting and refinement. A theoretical articulation of academic literacy is provided, since this constitutes a crucial aspect of construct validity, with particular attention being given to the delineation of functional academic

literacy. The identified ability is reflected in the blueprint for the test construct and is further specified in the task types selected for inclusion in the test. The definite move away from previous generations of 'scientific' and positivist thinking can be discerned in the kind of test tasks required of test takers and the way in which these emphasize the instrumental communicative function and mediating role of language as a social instrument within the material lingual sphere of academic discourse. Task types are evaluated in terms of their ability to be productive, based on a quantitative system of measurement and the application of appropriate statistical procedures such as point-biserial correlations. The research concludes with a reception study which assesses the extent to which the current version of the Test of Academic Literacy for Postgraduate Students (TALPS) is well received and whether it meets the regulative conditions of fairness, transparency, utility and accountability. In addition to confirming that the test has an acceptably high face validity, the survey section reveals that many postgraduate students have a restricted view of language and an erroneous perception of academic literacy. Much more needs to be done at undergraduate level to raise awareness about the important role played by language proficiency in all fields of study and how high academic literacy levels can contribute towards academic achievement and student success. The design of effective literacy interventions remains another area of concern. Research is also needed on the desirability and feasibility of designing subject-specific tests and ensuring equivalence where different test versions are in operation.

Abstrak

Verskeie studies dui aan dat die geletterdheidsvlakke van studente aan tersiêre inrigtings in Suid-Afrika laer is as wat vereis word vir akademiese sukses. Desnieteenstaande bestaan die moontlikheid dat van hierdie studente wel daarin sal slaag om met hul lae akademiese geletterdheidsvlakke te gradueer, en selfs tot nagraadse studie toegelaat word, al ontbreek hulle die vermoë om die gevorderde akademiese diskoers te ontsluit wat benodig word vir sukses op hierdie vlak. Hierdie studie bevestig die behoefte wat bestaan vir die ontwikkeling van ’n akademiese geletterdheidstoets vir nagraadse studente en dui aan hoe dit bewerkstellig kan word deur die toepassing van erkende ontwerpbeginsels binne die bestek van die toegepaste linguistiek. Die konstitutiewe en regulatiewe voorwaardes vir die validering van taaltoetse word behandel as die nodige fundering vir die ontwerp van ’n akademiese geletterdheidstoets vir nagraadse studente. Terwyl daar konvensioneel vereis word dat taaltoetse geldigheid en betroubaarheid moet toon, moet hulle volgens meer kontemporêre denke ook konsekwensiële geldigheid besit, ‘n konsep wat verwys na die impak van toetse. Taalvaardigheidstoetse moet dus nie net konsekwent en teoreties verantwoordbaar wees nie, maar benodig ‘n menigvuldigheid van bewyse ter stawing van hulle geldigheid.

Die onderskeie fases betrokke in die ontwerp van ‘n taaltoets word bespreek, insluitend die van loodsing en verfyning. Die begrip akademiese geletterdheid word teoreties ge-artikuleer as ‘n kern aspek van konstrukgeldigheid, met

spesifieke verwysing na die omskrywing van funksionele akademiese geletterdheid. Die vereiste taalvermoë word gereflekteer in die bloudruk vir die toetskonstruk en word verder gespesifiseer in die onderskeie toetstaaktipes. Die duidelike beweging weg van vorige generasies van ‘wetenskaplike’ en positivistiese denke word bespeur in die tipe toetstake wat van toetsafleggers verlang word en in die beklemtoning van instrumentele kommunikatiewe funksie en die bemiddelende rol van taal as sosiale instrument binne die materiale linguale sfeer van die akademiese diskoers. Taaktipes word geëvalueer na gelang van hulle produktiewe vermoë volgens ‘n kwantitatiewe metingstelsel en die toepassing van toepaslike statistiese prosedures, onder meer korrelasiekoëffisiënte. Die navorsing sluit af met ‘n resepsiestudie oor die mate waartoe die huidige weergawe van die Test of Academic Literacy for Postgraduate Students (TALPS) goed ontvang word en tot watter mate dit voldoen aan die regulatiewe voorwaardes van billikheid, deursigtigheid, bruikbaarheid en verantwoordbaarheid. Die opname bevestig die hoë gesigsgeldigheid van die toets en dui verder aan dat heelwat nagraadse studente ‘n beperkte siening van taal het en ‘n foutiewe persepsie van akademiese geletterdheid huldig. Daar sal baie meer gedoen moet word op voorgraadse vlak om bewustheid oor die belangrike rol van taalvaardigheid in alle studiegebiede te verhoog en oor hoe hoë akademiese geletterdheidsvlakke bydra tot akademiese vooruitgang en studentesukses. Die ontwerp van doeltreffende taalvaardigheidsintervensies is ‘n verdere bron van kommer. Daar is ook ‘n behoefte aan verdere navorsing oor die verkieslikheid en

doenlikheid van die ontwerp van vak-spesifieke toetse en ekwivalensie waar verskeie toetsweergawes in gebruik is.

Key words

Academic literacy

Applied linguistics

Consequential validity

Construct validity

Design principles

Language testing

Postgraduate literacy

Reliability

Validation

Validity

Annexure A: Through-put rates per faculty (HEMIS) - 1 August 2011 (DIRAP 2011)

	2006			2007			2008			2009			2010		
	HCs	**GRs**	**TPRs**	**HCs**	**GRs**	**TPRs**	**HCs**	**GRs**	**TPRs**	**HCs**	**GRs**	**TPRs**	**HCs**	**GRs**	**TPRs**
Post Graduate	**7386**	**2505**	**33.92%**	**7174**	**2241**	**31.24%**	**7224**	**2385**	**33.01%**	**7512**	**2491**	**33.16%**	**7365**	**2579**	**35.02%**
Eco & Man. Sciences	1036	374	36.10%	926	304	32.83%	887	304	34.27%	952	300	31.51%	981	309	31.50%
Education	1659	555	33.45%	1490	487	32.68%	1537	461	29.99%	1582	464	29.33%	1327	448	33.76%
Health Sciences	946	351	37.10%	919	205	22.31%	976	246	25.20%	1032	319	30.91%	1058	358	33.84%
Law	1420	518	36.48%	1395	523	37.49%	1351	524	38.79%	1420	483	34.01%	1287	490	38.07%
Nat. & Agric. Sciences	1357	404	29.77%	1477	418	28.30%	1540	527	34.22%	1614	578	35.81%	1743	598	34.31%
The Humanities	785	248	31.59%	784	248	31.63%	756	277	36.64%	744	311	41.80%	803	317	39.48%
Theology	183	55	30.05%	183	56	30.60%	177	46	25.99%	168	36	21.43%	166	59	35.54%
Under Graduate	**15428**	**2921**	**18.93%**	**15970**	**2458**	**15.39%**	**16828**	**2728**	**16.21%**	**17108**	**2888**	**16.88%**	**18882**	**2965**	**15.70%**
Eco. & Man. Sciences	4638	654	14.10%	4513	596	13.21%	3992	641	16.06%	3997	654	16.36%	4169	667	16.00%
Education	3043	1101	36.18%	3160	666	21.08%	4090	717	17.53%	3903	838	21.47%	4466	707	15.83%
Health Sciences	1413	250	17.69%	1417	260	18.35%	1415	284	20.07%	1375	263	19.13%	1391	261	18.76%
Law	975	100	10.26%	991	106	10.70%	898	122	13.59%	908	101	11.12%	937	124	13.23%
Nat. & Agric. Sciences	2617	410	15.67%	2817	438	15.55%	2985	462	15.48%	3248	483	14.87%	3607	553	15.33%
The Humanities	2649	389	14.68%	2974	381	12.81%	3333	485	14.55%	3574	532	14.89%	4210	631	14.99%
Theology	93	17	18.28%	98	11	11.22%	115	17	14.78%	103	17	16.50%	102	22	21.57%
Grand Total	**22814**	**5426**	**23.78%**	**23144**	**4699**	**20.30%**	**24052**	**5113**	**21.26%**	**24620**	**5379**	**21.85%**	**26247**	**5544**	**21.12%**

HCs = Headcounts

GRs = Graduates

TPRs = Through-put rates

Annexure B: Item analysis of ENG 104 class test

Sequence	Item ID	Key	Scored	NumOptions	Domain	N	P	Total Rpbis	Total Rbis	Domain Rpbis	Domain Rbis	Alpha w/o
1	1	C	Yes	5	Scrambled	150	0.733	0.299	0.403	0.387	0.521	0.74
2	2	A	Yes	5	Scrambled	150	0.26	0.325	0.44	0.595	0.804	0.739
3	3	E	Yes	5	Scrambled	150	0.347	0.16	0.207	0.516	0.665	0.746
4	4	B	Yes	5	Scrambled	150	0.307	0.212	0.278	0.549	0.721	0.744
5	5	D	Yes	5	Scrambled	150	0.253	0.419	0.57	0.608	0.826	0.735
6	6	C	Yes	4	Understar	150	0.067	-0.006	-0.012	0.03	0.057	0.75
7	7	D	Yes	4	Understar	150	0.673	0.139	0.181	0.286	0.371	0.748
8	8	B	Yes	4	Understar	150	0.333	0.142	0.184	0.172	0.224	0.747
9	9	D	Yes	4	Understar	150	0.667	0.092	0.119	0.225	0.292	0.749
10	10	C	Yes	4	Understar	150	0.367	0.11	0.141	0.153	0.196	0.749
11	11	D	Yes	4	Understar	150	0.427	0.199	0.251	0.264	0.333	0.745
12	12	A	Yes	4	Understar	150	0.453	0.118	0.149	0.199	0.25	0.749
13	13	A	Yes	4	Understar	150	0.273	0.156	0.209	0.138	0.185	0.746
14	14	C	Yes	4	Understar	150	0.513	0.191	0.239	0.255	0.32	0.744
15	15	B	Yes	4	Understar	150	0.44	0.276	0.347	0.297	0.373	0.741
16	16	C	Yes	4	Understar	150	0.26	0.015	0.02	0.079	0.107	0.753
17	17	B	Yes	4	Understar	150	0.413	0.219	0.277	0.192	0.243	0.743
18	18	D	Yes	4	Understar	150	0.6	-0.005	-0.007	0.067	0.085	0.754
19	19	C	Yes	4	Understar	150	0.72	0.213	0.284	0.294	0.392	0.744
20	20	C	Yes	4	Understar	150	0.32	0.288	0.375	0.273	0.356	0.741
21	21	B	Yes	4	Understar	150	0.347	0.371	0.479	0.392	0.505	0.736
22	22	C	Yes	4	Understar	150	0.393	0.513	0.651	0.446	0.566	0.729
23	23	A	Yes	4	Understar	150	0.407	0.101	0.128	0.154	0.195	0.75
24	24	A	Yes	4	Understar	150	0.6	0.144	0.183	0.137	0.173	0.747
25	25	D	Yes	4	Understar	150	0.127	0.145	0.232	0.177	0.284	0.746
26	26	A	Yes	4	Understar	150	0.287	0.114	0.151	0.097	0.128	0.748
27	27	B	Yes	4	Understar	150	0.38	0.123	0.157	0.18	0.229	0.747
28	28	D	Yes	5	Text types	150	0.387	0.366	0.465	0.568	0.722	0.735
29	29	C	Yes	5	Text types	150	0.827	0.198	0.293	0.433	0.639	0.743
30	30	E	Yes	5	Text types	150	0.447	0.327	0.411	0.553	0.695	0.738
31	31	B	Yes	5	Text types	150	0.86	0.235	0.367	0.455	0.709	0.741
32	32	A	Yes	5	Text types	150	0.847	0.31	0.472	0.524	0.797	0.739
33	33	A	Yes	4	Vocabular	150	0.58	0.293	0.369	0.401	0.506	0.739
34	34	C	Yes	4	Vocabular	150	0.56	0.326	0.41	0.418	0.527	0.739
35	35	B	Yes	4	Vocabular	150	0.367	0.33	0.422	0.413	0.528	0.738
36	36	D	Yes	4	Vocabular	150	0.487	0.218	0.273	0.322	0.404	0.742
37	37	D	Yes	4	Text editi	150	0.413	0.139	0.176	0.18	0.227	0.746
38	38	D	Yes	4	Text editi	150	0.507	0.15	0.188	0.135	0.169	0.745
39	39	C	Yes	4	Text editi	150	0.413	0.158	0.2	0.259	0.328	0.745
40	40	D	Yes	4	Text editi	150	0.273	0.258	0.345	0.263	0.353	0.741
41	41	D	Yes	4	Text editi	150	0.687	0.119	0.156	0.249	0.325	0.743

Annexure C

THE RECEPTION OF A TEST OF ACADEMIC LITERACY FOR POSTGRADUATE STUDENTS

Instructions and additional information for students

FOR OFFICE USE

For the purposes of this study the term "academic literacy" refers to the integrated language ability of students to read and critically reflect on academic texts and to undertake academic tasks typically required at postgraduate level, including the production of written texts.

Where necessary circle the appropriate block to indicate your choice. Only **one** answer per question may be selected. Where requested provide a short answer.

1 Write your student or ID number in the blocks below:

													Q1			1-3

2 What is your field of study?

Economic & Management Sciences	1	Q2 4
Education	2	
Humanities (psychol., languages, etc.)	3	
Law	4	
Medicine & Allied Professions	5	
Natural & Agricultural Sciences	6	
Theology	7	

3 What level of study are you presently engaged in?

Undergraduate	1	Q3 5
Postgraduate diploma/certificate	2	
Honours	3	
Masters	4	
Not studying at present	5	

4 How old are you? _____ Q4 6-7

5 What is your gender?

Female	1	Q5 8
Male	2	

6 What is your home language? _____ Q6 9-10

7 In what language did you study at primary school? _____ Q7 11-12

8 In what language did you study at secondary school? _____ Q8 13-14

9 What was your language of instruction at university? _____ Q9 [ ][ ] 15-16

10 Did you do any language development courses at university?

Yes	1
No	2

Q10 [ ] 17

11 Was English one of your mainstream subjects at university?

Yes	1
No	2

Q11 [ ] 18

12 Which of the following best describes your reaction when you were asked to write the test of academic literacy?

Negative	1
Angry	2
Concerned	3
Neutral	4
Positive	5

Q12 [ ] 19

13 Was the format of the test discussed with you before the test date?

Yes	1
No	2

Q13 [ ] 20

14 Did you have access to an example test?

Yes	1
No	2

Q14 [ ] 21

15 Would you prefer to do the test on computer rather than on paper?

Yes	1
No	2

Q15 [ ] 22

16 Did you experience any anxiety while taking the test?

1	2	3	4	5	6
None					A lot

Q16 [ ] 23

17 In your opinion can anxiety affect test performance?

1	2	3	4	5	6
Not at all					A lot

Q17 [ ] 24

18 How did you find the test?

1 Very easy	2	3	4	5	6 Very hard

Q18 [ ] 25

19 How would you describe the time limit for completing the test?

1 Too little time	2	3	4	5	6 Too much time

Q19 [ ] 26

20 How accurately do you think the test can measure your level of academic literacy?

1 Not at all accurately	2	3	4	5	6 Very accurately

Q20 [ ] 27

21 Should all undergraduate students who wish to register for postgraduate study be required to write the test?

Yes	1
No	2

Q21 [ ] 28

22.1 Would you say that the test is fair to students?

Yes	1
No	2

Q22.1 [ ] 29

22.2 If you answered no, explain why you think the test is not fair:

_____
_____
_____
_____
_____
_____

Q22.2 [ ][ ] 30-31
[ ][ ] 32-33

23 Which **one** of the following do you find the **most** difficult in your course work?

Reading with understanding	1
Critical thinking/reasoning	2
Writing academic texts	3

Q23 [ ] 34

24 How would you rate your ability to read with understanding?

1 Very poor	2	3	4	5	6 Very good

Q24 [ ] 35

25 How would you rate your ability to reason critically?

1 Very poor	2	3	4	5	6 Very good

Q25 [ ] 36

26 How would you rate your ability to write academic texts?

1 Very poor	2	3	4	5	6 Very good

Q26 [ ] 37

27 Rate how you found the following sections of the test:

		Very easy									Very hard		
27.1	Scrambled text	1	2	3	4	5	6	7	8	9	10	Q27.1	38
27.2	Interpreting graphs	1	2	3	4	5	6	7	8	9	10	Q27.2	39
27.3	Academic vocabulary	1	2	3	4	5	6	7	8	9	10	Q27.3	40
27.4	Matching text types	1	2	3	4	5	6	7	8	9	10	Q27.4	41
27.5	Text comprehension	1	2	3	4	5	6	7	8	9	10	Q27.5	42
27.6	Grammar/text relations	1	2	3	4	5	6	7	8	9	10	Q27.6	43
27.7	Text editing	1	2	3	4	5	6	7	8	9	10	Q27.7	44
27.8	Academic writing	1	2	3	4	5	6	7	8	9	10	Q27.8	45

28 Answer the following questions on how **you** assess your **own** academic literacy levels:

28.1	Do you compare yourself to other persons?	Yes 1	No 2	Q28.1	46
28.2	Do you use your undergraduate course marks to assess how literate you are?	Yes 1	No 2	Q28.2	47
28.3	Do you understand literacy to refer to the language skills of listening, speaking, reading and writing?	Yes 1	No 2	Q28.3	48
28.4	Do you think that if you have a big vocabulary you are literate?	Yes 1	No 2	Q28.4	49
28.5	Do you believe that since you have had the chance to study further after school you must be literate?	Yes 1	No 2	Q28.5	50
28.6	Do you believe that the kind of language used in university courses is the same as the kind of language used outside the university?	Yes 1	No 2	Q28.6	51

29 Please write down any further comments on the test that you would like to make.

Q29 [ ] [ ] 52-53