Avasha Rambiritch
University of Pretoria

Validating the Test of Academic Literacy for Postgraduate Students (TALPS)

Abstract

The concepts of reliability and validity help to determine whether a test is a strong one or not, does what it is designed to do, tests what it is designed to test, and whether we can make inferences that are justified about the test takers, based on their score (Van der Walt & Steyn, 2007). Clearly, a lot depends on the reliability and validity of a test. This article will take the form of a validation argument of the Test of Academic Literacy for Postgraduate Students (TALPS). It will do so by providing a priori evidence collected before the test event (Weir, 2005) to support eight claims made about the reliability and validity of the test.

Keywords: academic literacy, validity, content validity, face validity, construct validity, reliability

Journal for Language Teaching, Volume 47, Number 1, pp. 175-193. 2013.
ISSN 0259-9570. DOI: http://dx.doi.org/10.4314/jlt.v47i1.8

1. Introduction

The Unit for Academic Literacy (UAL) at the University of Pretoria (UP), in addition to focusing on developing the academic literacy of undergraduate students, is also concerned with developing the academic literacy of postgraduate students. For the last seven years the Academic Writing for Postgraduate Students (EOT 300) module has focused on helping to develop the academic writing needs of postgraduate students. There has, over this period, been an increasing demand for the course, as supervisors recognised the poor academic literacy levels of their students. This has been the focus of a study conducted by Butler (2007). The study is concerned with the design of a course for academic writing at tertiary level. He states that the "immediate context of this study derives from the concern that a number of academic departments from a variety of disciplines at the University of Pretoria have expressed about the academic writing ability (and general language proficiency) of their postgraduate students" (2007:10). He explains that these students are unfamiliar with academic writing conventions, are often unable to express themselves clearly in English, and have not "yet fully acquired the academic discourse needed in order to cope independently with the literacy demands of postgraduate study" (2007:10). What became clear from Butler's study was the need for a "reliable literacy assessment instrument" (Butler, 2007:181) that would "provide one with accurate information on students' academic literacy levels" (2007:181). The need for a reliable testing instrument for postgraduate students had been identified and work on the test began. The story of the Test of Academic Literacy for Postgraduate Students (TALPS), i.e. its design and development is the focus of another article. The purpose of this article is to present a validation study of TALPS by making a number of claims about the test and providing evidence to support these claims. Details of the test and the administrations on which the validation exercise is based is provided below.

2. The design and development of TALPS

The first draft

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

Section 1 – Scrambled text (5 marks x 3)

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

Section 3 – Dictionary definitions (5 marks)

Section 4 – Academic vocabulary (40 marks)

Section 5 – Text types (5 marks)

Section 6 – Understanding texts (60 marks)

Section 7 – Grammar and text relations (22 marks)

Section 8 – T ext editing (10 marks)

With regard to TALPS, it was decided to include a section on argumentative writing. At postgraduate level it is essential that students follow specific academic writing conventions and it is important to test whether students were equipped with this knowledge. In addition to the question on writing there is a question that tests students' editing skills. At this stage, however, the test did not include a question requiring students to write an argumentative text. The concern of the developers at this early stage was in writing the multiple-choice questions. These would then be analysed using TiaPlus Test and Item Analysis (CITO, 2006) to determine which items did or did not test well. Items that did not test well were discarded.

The second draft

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

Section 1 – This section now included only one set of sentences. The total marks remained the same at 5 marks.

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

The third draft

The third draft version of the test became the first pilot for TALPS. This first pilot was completed in May 2007 with first year students at the University of Pretoria (UP). Students were given one and a half hours to complete the test. These were students who were taking the compulsory Academic Literacy module (EOT 110). These results were measured using TiaPlus Test and Item Analysis (CITO, 2006) which provides measures at item and test level. Before the first pilot, by which time the test had been reduced to 100, items were evaluated by the designers to determine the appropriateness/strength of the item. Most changes were made in the Understanding texts section. In the 100-item test (which was the first pilot) this section had 45 items; in the 88-item test (which was the second pilot) it had 33 items and then 28 items. The final version of the test has 21 items in this section. Justification for this decision was drawn from the analyses done using the TiaPlus Test and Item Analysis Build 300 (CITO, 2006). According to this, seven items had very high p-values, meaning that a high percentage of the test population got this answer correct.

Piloting the test

The first pilot

The first pilot of TALPS had 100 items and four sections: Dictionary definitions, Academic vocabulary, Understanding texts and Grammar and text relations. The test totalled 100 marks. It did not include the question requiring students to write an argumentative text.

The second pilot

The second pilot of TALPS was carried out on postgraduate students both at the University of Pretoria (UP) and the University of the Free State (UFS) in September 2007. This test comprised 88 items and totalled 120 marks. The TALPS second pilot was also carried out on a second batch of students from the University of the Free State (UFS).

The TALPS final draft version

The final draft version of TALPS was made up of 76 items and eight sections.This version of the test totalled 100 marks. The section on the Dictionary definitions was left out of this version of the test. According to the descriptive statistics of the drafts of TALPS, the Dictionary definitions question had p-values of 84.2 for both the 88 item pilots. Davies et al. (1999) explain that the higher the index, the easier the item. The closer the index is to 100% or 0%, the less differential information it can provide about candidates. They state that items that are excessively easy or very difficult are normally removed because they do not contribute to the test's discriminability (1999: 95). The pilot for this version of the test was carried out in September 2007 on two groups of students: postgraduate students from the North-West University (NWU) and postgraduate students from the University of Pretoria (UP).

3. Validity and the validation argument

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

There is evidently a distinction between the concept of validity and the act of validation. Davies and Elder (2005:799) make reference to this distinction when they speak of Messick's (1989) concern with "validity as a theoretical concept rather than with validation in its practical operation". In discussing this distinction further, their claim is that validity is an abstract and essentially empty concept, that it is through "validation that validity is established, which means that validity is only as good as its validation procedures"

(Davies & Elder, 2005:795). Van der Walt and Steyn label validation "an activity: the collection of all possible test-related activities from multiple sources" (2007:141). These 'multiple sources' of evidence may include what are traditionally conceived as content and construct validity, concurrent and predictive validity, face validity, reliability, as well as consequential validity (Davies & Elder, 2005:798). In a nutshell:

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

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

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

(Fulcher & Davidson, 2007:20).

What follows below is a brief discussion of the validation of the TALPS test.

4. A validation of the TALPS

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

According to Kurpius and Stafford, reliability can be defined as the "trustworthiness or the accuracy of a measurement" (2006:121). Bachman and Palmer state that reliability can be considered to be a function of the consistency of scores from one test, and test tasks, to another (1996:19). Another important point that Bachman and Palmer make is that reliability is an essential quality of test scores and that unless test scores are relatively consistent, they cannot provide us with any information at all about the ability we want to measure (1996: 20).

Kurpius and Stafford (2006) identify four types of reliability: test-retest reliability, alternate forms reliability, internal consistency reliability and inter-rater reliability. Test-retest reliability requires the test to be taken twice by the same group of students, alternate forms reliability is used when "you want to determine whether two equivalent forms of the same test are really equivalent" (2006:126), and inter-rater reliability is

used when “two or more raters are making judgements about something” (2006:129). The main reliability measure used for TALPS is internal consistency reliability. This type of reliability is obviously the most practical to use – it requires that students take the test once and reliability measures are calculated using statistical packages like TiaPlus, SPSS or Iteman.

The instrument/package we have used (TiaPlus: cf. CITO, 2006) provides us with two measures: Cronbach’s alpha and Greatest Lower Bound (GLB). All three pilots of the test have rendered very impressive reliability measures as indicated in Table 1 below. The first pilot had a reliability of 0.85 (Cronbach’s alpha) and 0.92 (GLB). The pre-final draft had measures of 0.93 (Cronbach’s alpha) and 1.00 (GLB). The final version of the test had measures of 0.92 (Cronbach’s alpha) and 0.99 (GLB). The measures for the final draft were based on the analysis done on the combined group (North-West University and the University of Pretoria).

Table 1: Reliability measures for the TALPS pilots

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

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

$Xo = Xt + Xe.$

- Xo = the score obtained by a person taking the exam (referred to as an obtained score or observed score)
- Xt = a person’s true score
- Xe = the error score associated with the obtained score

(Kurpius & Stafford, 2006:103).

The individual student's true score then would be the obtained score - (minus) 3.87 (2nd pilot done on UFS students) or 3.82 (3rd pilot done on UP and UFS students), which in each case is the standard error of measurement. Kurpius and Stafford explain that a smaller standard error of measurement reflects a smaller error score and that the goal in reliability is to control error (2006:133). A higher reliability is therefore an indication of a small error of measurement. The 1st pilot of TALPS that has a reliability of 0.85, had a standard error of 4.30. When the reliability measures in subsequent pilots improved, the standard error of measurement dropped to 3.82 and 3.87, respectively, as can be seen in Table 2 below.

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

Table 2: Descriptive statistics of the TALPS pilots

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

One other set of empirical information about the reliability of the test yielded by the TiaPlus Test and Item Analysis is the Coefficient Alpha of the test if it had a norm length of forty items. TALPS is made up of a number of short subtests. The reliability of a test or in this case a subtest will be compromised by its length – the longer a test is, the more reliable it usually is. Kurpius and Stafford explain that when a test is too short, the reliability coefficient is suppressed due to the statistics that are employed. The Spearman-Brown correction procedure can be used to make up for this (Kurpius & Stafford, 2006:129). One example of this is the Dictionary definitions subtest in the TALPS first pilot. This section has five items and is one of the shortest sections in the test. It has a Coefficient alpha of 0.36 and GLB measure of 0.42. The Spearman Brown correction procedure indicates a Coefficient Alpha of 0.82 if it had a standard norm length of 40 items.

The ideal then would be to design longer tests, thus ensuring higher reliability measures. But – and this is always the technical trade-off – such effectiveness may conflict with technical implementation constraints: there may not be enough time available to test. The test developer has to weigh up the advantages and disadvantages of lengthening a test, and take a responsible design decision.

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

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

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

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

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

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

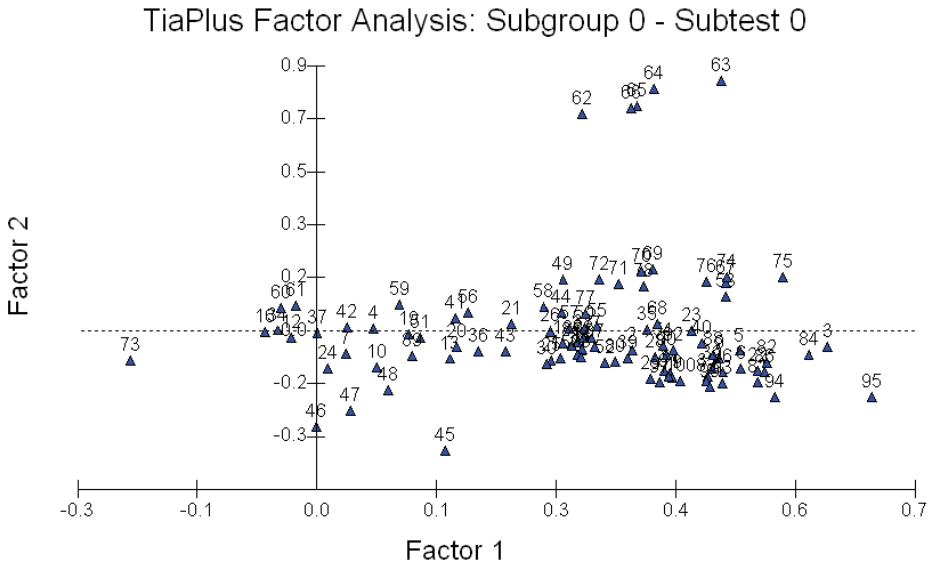

Figure 1: Measures of homogeneity/heterogeneity of TALPS first pilot (Geldenhuys 2007:73)

According to Geldenhuys (2007:73) the more heterogeneous items are, the less reliable the test can become. The factor analysis above indicates that in the case of TALPS there is a measure of heterogeneity: items 73 and items 62-66 are furthest away from the zero line. The test, however, still had a reliability measure of 0.85. The test designers chose to leave in these items, arguing that academic literacy is a "richly varied and potentially complex" (Weideman, 2009:237) ability and one would therefore have to "tolerate a more heterogeneous construct" (Weideman, 2009:237).

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

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

Table 3: Average Rit-values of the TALPS pilots

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

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

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

In terms of Kunnan's Test Fairness framework, construct validity is concerned with the representation of the construct/underlying trait (2004:37) that is being measured. Bachman and Palmer (1996:21) define a construct as the "specific definition of an ability that provides the basis for a given test or test task and for interpreting scores derived from this task." They explain further that construct validity is used to refer to the extent to which we can interpret a given test score as an indicator of the abilities or constructs we want to measure (1996:21). The construct for TALPS is based on the same construct as the Test of Academic Literacy Levels (TALL). The TALL has in many ways been the sounding board for TALPS. Moreover, the success of TALL has in part been the justification for TALPS. TALL and TALPS are designed to test the same ability – the academic literacy levels of students: undergraduate in the case of TALL, and postgraduate in the case of TALPS.

The discussion here of the construct validity of TALPS demands further discussion of the construct on which the test is based. In deciding on a construct for TALL, Van Dyk and Weideman (2004:7) set out to answer the all important question of "what would a construct based on a theory of academic literacy look like?" In doing so they considered the work of Blanton (1994), Bachman and Palmer (1996) and Yeld (2000). Blanton's (1994) definition was important to Van Dyk and Weideman (2004) because it "described what proficient academic readers and writers should do" (2004:7). Importantly, it was a move away from an emphasis on vocabulary and grammar towards what Weideman has referred to as an "open view of language" (2003:58). When turning to Bachman and Palmer's (1996) definition of language ability, Van Dyk and Weideman (2004:8)

found that while it provided more detail, the "apparent seepage between categories" in the construct could be confusing. In addition, Bachman and Palmer (1996:66) point out that the construct would have to be reinterpreted for each testing situation. Yeld (2000) has done exactly this in the design of the academic literacy test developed at the Alternative Admissions Research Project (AARP), and it is this construct that Van Dyk and Weideman (2004) find most useful. Van Dyk and Weideman (2004:10) point out, however, that while the construct was useful, the AARP test was an admissions test, was part of a larger battery of tests, was a two and a half hour test and took more than three hours to administer. This would not be practical for the academic literacy test planned for the students at the University of Pretoria (UP). What was needed was a "reconceptualisation" (Van Dyk & Weideman, 2004:10) of how the test was designed. After much rationalising, re-ordering and reformulating (2004:10), the result was a "streamlined version" (2004:10) that made possible the testing of academic literacy within a much shorter time frame.

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

- understand a range of academic vocabulary in context;
- interpret and use metaphor and idiom, and perceive connotation, word play and ambiguity;
- understand relations between different parts of a text, be aware of the logical development of (an academic) text, via introductions to conclusions, and know how to use language that serves to make the different parts of a text hang together;
- interpret different kinds of text type (genre), and show sensitivity for the meaning that they convey, and the audience that they are aimed at;
- interpret, use and produce information presented in graphic or visual format;
- make distinctions between essential and non-essential information, fact and opinion, propositions and arguments; distinguish between cause and effect, classify, categorise and handle data that make comparisons;
- see sequence and order, do simple numerical estimations and computations that are relevant to academic information, that allow comparisons to be made, and can be applied for the purposes of an argument;
- know what counts as evidence for an argument, extrapolate from information by making inferences, and apply the information or its implications to other cases than the one at hand;
- understand the communicative function of various ways of expression in academic language (such as defining, providing examples, arguing); and

- make meaning (e.g. of an academic text) beyond the level of the sentence (Weideman, 2003:61).

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

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

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

Table 4: Table of subtest intercorrelations (TALPS 2nd pilot)

	Subtest	Total Test	Subtest(s) 1	Subtest(s) 2	Subtest(s) 3	Subtest(s) 4	Subtest(s) 5	Subtest(s) 6	Subtest(s) 7
Scrambled text	1	0.43							
Interpreting graphs	2	0.73	0.22						
Dictionary definitions	3	0.37	0.24	0.12					
Academic vocabulary	4	0.59	0.24	0.53	0.31				
Understanding texts	5	0.83	0.25	0.51	0.32	0.35			
Grammar & text relations	6	0.82	0.24	0.51	0.15	0.36	0.57		
Text editing	7	0.72	0.18	0.40	0.24	0.34	0.50	0.59	
Number of testees	:	117	117	117	117	117	117	117	117

	Subtest	Total Test	Subtest(s) 1	Subtest(s) 2	Subtest(s) 3	Subtest(s) 4	Subtest(s) 5	Subtest(s) 6	Subtest(s) 7
Number of items	:	88	5	10	5	10	33	15	10
Average test score	:	61.33	2.66	6.56	4.19	7.60	23.75	8.96	7.62
Standard deviation	:	14.19	1.92	2.89	0.93	1.83	4.89	4.52	2.85
SEM	:	3.82	0.75	1.22	0.75	1.24	2.30	1.51	1.00
Average, P-value	:	69.7	53.16	65.64	83.76	75.98	71.98	59.72	76.15
Coefficient Alpha	:	0.93	0.85	0.82	0.35	0.54	0.78	0.89	0.88
GLB	:	1.00	0.94	0.89	0.41	0.74	0.94	0.96	0.92
Asymptotic GLB	:	Na	Na	Na	Na	Na	Na	Na	Na

Davies et al. (1999) explain that a correlation coefficient is a value showing the degree to which two variables are related, that a coefficient of zero indicates that there is no relationship between the two variables, a coefficient of -1 indicates a perfect negative correlation, and a coefficient of +1indicates a perfect positive correlation (1999:36). In terms of the correlation between each pair of subtests, these should fall between 0,3–0,5 (Alderson, Clapham & Wall, 1995:184). Alderson et al. explain that the reason for having different test components is that they all measure something different and therefore contribute to the overall picture of language ability attempted by the test. These correlations should be fairly low, in the "order of +.3 - +.5" (Alderson et al. 1995:184). If, however, these components correlate very highly (around +.9) one may wonder whether the two subtests are testing different traits or skills, or whether they are testing the same thing (1995:184). Of the 21 correlations in the table above, 9 fall below 0.3. Should this be adjusted, in line with the experience with TALL, these levels to 0.2 and 0.5, then 15 of the 21 are between the acceptable parameters. With regards the correlation between each subtest and the whole test, this should be "around +.7 or more since the overall score is taken to be a more general measure of language ability than each individual component score" (Alderson et al. 1995:184). Overall the average of the correlation between each subtest and the whole test, while not ideal, is an acceptable 0.64. The subtests that correlate best with the whole test are the Interpreting Graphs, Understanding Texts and the Grammar and Text Relations subtest. In the case of the TALPS final draft version (UP & UNW combined) (See Table 5 below), the average of the correlations between each subtest and the whole test is 0.66, indicating that the subtests correlate well, and more acceptably, with the test.

Table 5: Table of subtest intercorrelations (TALPS final draft version) UP& UNW Combined

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

Claim 7: The test displays content validity.

A factor analysis is also useful in determining the content validity of the test. According to the factor analysis above (Fig.1), not all items are related to a single construct underlying the test. As indicated earlier, it was the decision of the test developers not to exclude these outlying items, the reasoning being that "for an ability as richly varied and potentially complex as academic language ability, one would expect, and therefore have to tolerate, a more heterogeneous construct" (Weideman, 2009:237). Leaving out these outlying items would have increased the reliability of the test. The test, however, already has an excellent reliability measure of 0.85. The high reliability of the test allows the test developer the freedom to include these items without compromising the reliability of the test or the construct.

The one other method of determining the content validity of the test is to get "expert ratings of the relationship between the test items and the content domain" (Kurpius and Stafford, 2006:147). These experts judge each item to determine "how well it assesses the desired content" (2006:147). In the case of TALPS, members of the design team

were already familiar with the design of a test of this nature, having been involved in the design of the TALL tests. In addition to this, drafts of the TALPS were evaluated by other specialists within the academic institutions involved who were either interested in being involved in the process of design and development or were interested in using it on their students.

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

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

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

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

The face validity of TALPS is therefore an important consideration. Students who will be writing the test are from different faculties and disciplines. Their supervisors are not experts in the field of language testing and academic literacy. In having their students write the test they will have to believe that the test looks right, that it looks like a test that is testing the academic literacy of their students. In attempting to "speculate responsibly" (Butler, 2009:299) about the face validity of TALPS, Butler looked at supervisor perceptions of their students academic literacy, at students perceptions of their academic literacy abilities, and aligned this with the design of the TALPS. According to Butler's findings the potential face validity of TALPS indicates that it does meet the "expectations of prospective users" (2009:299).

5. Conclusion

It goes without saying that it should be the aim of test developers to design tests that are valid and reliable. Equally important is the need for validation at the a priori stage of test development (Weir, 2005). The evidence provided here indicate that TALPS is a highly reliable test – the final version of the test had measures of 0.92 (Cronbach's alpha) and 0.99 (GLB). Importantly, the test is based on a theoretically sound construct. Evidence has shown that while there are heterogeneous items in the test, this has not compromised the reliability of the test. The Rit-values indicate that the test discriminates well between test-takers. The internal correlations of the different test sections satisfy specific criteria and the face validity of the test meets the expectations of potential users. Based on the evidence collected, TALPS proves to be a highly valid and reliable test. What remains, however, is that further validation studies be conducted on the test for while "we can never escape from the need to define what is being measured, we are obliged to investigate how adequate a test is in operation" (Weir, 2005).

References

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

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

Bachman, L.F & Palmer, A.S.1996. *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Blanton, L.L. 1994. Discourse, artefacts and the Ozarks: understanding academic literacy. *Journal of second language writing* 3(1): 1-16. Reprinted in: Zamel, V. & Spack, R. (Eds.) 1998. *Negotiating academic literacies: teaching and learning across languages and cultures.* Mahwah, New Jersey: Lawrence Erlbaum Associates. pp. 219-235.

Butler, H.G. 2007. *A framework for course design in academic writing for tertiary education.* Unpublished doctoral thesis. Pretoria: University of Pretoria.

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

CITO. 2006. TiaPlus, Classical test and item analysis ©. Arnhem: Cito M. & R. Department.

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

Davies, A & Elder, C. 2005. Validity and validation in language testing. In: Hinkel, E. (Ed.) *Handbook of research in second language teaching and learning.* New Jersey: Lawrence Erllbaum Associates. pp. 795-813.

Fulcher, G. & Davidson, F. 2007. *Language testing and assessment: An advanced resource book.* New York: Routledge.

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

Ho, R. 2005. *Handbook of univariate and multivariate data analysis and interpretation with SPSS.* Boca Raton: Chapman & Hall.

Huot, B. 1990. Reliability, validity, and holistic scoring: What we know and what we need to know. *College composition and communication* 41(20): 201-213.

Kunnan, A. 2004. Test fairness. In: Milanovic, M & Weir, C. (Eds.) *Studies in language testing 18.* Cambridge: Cambridge Univesity Press. pp. 27-45.

Kurpius, S.E.R. & Stafford, M.E. 2006. *Testing and measurement: A user-friendly guide.* California: Sage Publications.

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

Messick, S. 1989. Validity. In: Linn, R.L. (Ed.) *Educational measurement* (3rd ed.). New York: American Council on Education & Macmillan. pp. 13-103.

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

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

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

Weideman, A. 2009. Constitutive and regulative conditions for the assessment of academic literacy. *Southern African linguistics and applied language studies*, 27(3): 235-251.

Weir, C.J. 2005. *Language testing and validation: An evidence-based approach.* Houndmills, Basingstoke: Palgrove Macmillan.

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

APPENDIX A TALPS MARKING RUBRIC

Content and organisation		Poor	Average	Good
Introduction (5)	Statement of issue – angle to be argued	No clear statement of issue; no point of view to be argued; abrupt or no introduction	States issue and point of view weakly; not clear what relevance is	Clearly states issue and point of view, explains relevance and importance
	Framing of reader expectations	No or little interest in explaining clearly what will follow, or in guiding reader	Attempts unsuccessfully to frame reader's expectations of what will follow	Clearly sets out what is to follow, providing a frame for what reader can expect
Body (argument) (5)	Nature of problem/ issue	No or little discussion of the nature of problem/issue, or why it is necessary to deal with it	Unsuccessfully attempts to discuss nature of problem/issue and its importance in South Africa	Clear discussion of nature of problem/ issue, and necessity of addressing it in South Africa
	Discussion of pros and cons	Gives no or little indication that there is more than one side to an argument	Attempts to provide both pros and cons, but does so unconvincingly	Provides a comprehensive discussion of possible pros and cons
	Argue convincingly for specific point of view	Argumentation is weak, one-sided, unconvincing	Argument deals with some of the important issues, but not in any convincing way	Strong, balanced argumentation that leaves the reader convinced of point of view
Conclusion (5)	Emphasising again the point of view advanced – link with introduction	No connection between the issue/thesis introduced in the introduction and what is said in conclusion	Attempts to restate the issue/thesis, but does so unconvincingly	Clearly emphasises the thesis again without making it a word by word repetition of the introduction
	Clearly states again the most important issues	No attempt to highlight again the most important issues in the text	Attempts to again include the most important issues, but does so in an unconvincing and incomplete manner	Clearly emphasises the main issues again in a structured and non-repetitive manner (exact repetition of the sentences used in body)