Phuong Loan Le, Colleen du Plessis and **Albert Weideman**

http://dx.doi.org/10.4314/jlt.v45i2.7

Test and context: The use of the Test of Academic Literacy Levels (TALL) at a tertiary institution in VIỆT NAM

A B S T R A C T Concern for the academic success of undergraduate students with low literacy levels is by no means unique to South Africa. In response to initiatives by the Ministry of Education and Training to standardise and improve the quality of higher education in the country in accordance with international standards, the University of Đà Nẵng in Việt Nam recently tested out a pilot version of the South African designed Test of Academic Literacy Levels (TALL) at its College of Foreign Languages (CFL). This article examines the appropriateness of the TALL for use at an institution such as the CFL in Việt Nam and compares the literacy levels of students studying at the college with those of a similar group of students at the University of Pretoria. The research findings show that the TALL exceeds the set requirements for international language tests in terms of reliability and that it has high construct validity. In addition hereto the test has proven to be highly feasible to administer, particularly with due consideration to time constraints and practicality. Academic literacy levels were found to be lower amongst the cohort of Vietnamese students than amongst a comparable group of South African students. Recommendations for the continued employment of the test in Việt Nam include doing differential item function analyses to explore gender and cultural differences, the minor refinement of less productive task items, and a slightly shorter test version.

Different contexts of literacy assessment and of undergraduate course success

How robust are tests of language ability, and how sensitive are they to context? As Van der Walt and Steyn (2007: 138) point out, the validity of a test is today considered to be "contextual, local and specific, pertaining to a specific use of a test." Nonetheless, the widespread global use of standardised tests continues apace; literally tens of millions of tests from the international

corporate test-making enterprises are sold annually, especially in Asia. In the People's Republic of China alone, billions of learners have been subjected to English proficiency tests such as the College Entrance Test (CET) over the past 23 years (Cheng 2008). In comparison, since its inception in the 1960s, the Test of English as a Foreign Language (TOEFL), which has been developed by the Educational Testing Service (ETS), has been taken by more than 25 million individuals from all over the world (TOEFL 2011), while the tests of the International English Language Testing System (IELTS) are taken by more than 1.4 million candidates each year (IELTS 2011). Such tests provide accepted benchmarks of language ability for many education administrators. This trend, of relying on information from an internationally accepted measure of language ability to take local decisions about the possible access of students to tertiary education, seems well-established, and is counteracted only by national or local initiatives to design and develop tests for specific national, local or institutional contexts. The Test of Academic Literacy Levels (TALL), which, as its name implies, is a test of academic literacy, is an example of such a locally developed test. After seven years of use, it is today a well-accepted and authoritative test that has been employed at several of South Africa's major higher education institutions, with some 32 000 students having written the English version of TALL and its Afrikaans counterpart, the Toets van Akademiese Geletterdheidsvlakke (TAG), in 2011 alone (ICELDA 2011). These tests form the only set of academic literacy tests in South Africa that has been developed specifically for use in multilingual institutions, and, with more than two dozen articles having analysed the tests and their empirical properties in scholarly journals (ICELDA 2011), it is probably the most thoroughly scrutinised set of tests of language ability, and certainly of academic literacy, in South Africa. The development of the two tests is claimed to be the outcome of deliberate and careful design decisions (Van Dyk & Weideman 2004a, 2004b).

Although the National Benchmark Tests (NBTs), which are used to assess the literacy and numeracy skills of prospective first-year entry students at many tertiary institutions in South Africa, serve as a good example of locally developed tests, the development of TALL and TAG illustrate that there is a need for further localisation in respect of language testing. Apart from practical constraints which may necessitate the design of different tests, literacy tests also serve different purposes. TALL and TAG are specifically used for placement and intervention purposes. There are also financial considerations. Currently, along with their high reliability and construct validity, TALL and TAG are the most affordable academic literacy tests available in South Africa for administration at tertiary level. The test of Academic Literacy and Quantitative Literacy domains (AQL) costs R90 per test, as compared to the R56 of TALL and TAG (see HESA 2011 and ICELDA 2011). In fact, for members of ICELDA, the inter-university partnership that owns TALL and TAG, a number of rebates and discounts contribute to the actual cost per test never being more than R15, which amounts to a substantial saving for each partner. One may compare that cost to that of international tests such as those offered by IELTS and ETS, which can cost as much as R2000 per test.

Since their initial development in 2004 at the University of Pretoria's Unit for Academic Literacy (Van Dyk & Weideman 2004a), TALL and TAG have become part of an initiative incorporating 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, as well as North-West University. In addition hereto, the English version of the test, sometimes in the form of a derivative or alternative version of TALL, is spreading its wings across international borders to institutions in Namibia and Singapore (ICELDA 2011; Van der Slik & Weideman 2007). TALL has also, at international conferences in Singapore (2010) and Beijing (in 2011), attracted enquiries from Indian and Chinese test administrators. In light of the widespread international use of similar tests of language ability, it is therefore appropriate to ask (1) whether TALL would be robust enough to yield meaningful measurements of the ability to use academic discourse in other than South African environments; and (2) if so, whether adjustments would need to be made to it in order to make it contextually more appropriate.

This article therefore examines the case of employing the current design of TALL as a placement test to address the concern for the academic success of Vietnamese undergraduates of the College of Foreign Languages (CFL) of the University of Đà Nẵng in Việt Nam, in order to determine the appropriateness of its use in that context. At the same time a comparison will be made between the results of the academic literacy tests administered amongst the Vietnamese students and those written by students at the University of Pretoria, and to ask which adjustments might need to be made to the test and its design for Việt Nam.

High-school graduates in Việt Nam who wish to be admitted to college or university are required to write the national entrance examination. This is a standardised high-stakes test which is written by nearly two million prospective university students each year (Johnson 2007). Examinations of this nature tend to assess different kinds of skills and knowledge and do not necessarily provide the information needed to indicate academic literacy proficiency in a particular language. Success at tertiary level for foreign language students demands a mastery of a combination of skills such as critical thinking and reading, analytical and problem-solving ability, as well as visual literacy. Since academic language proficiency is now widely believed to be a prerequisite for academic success at university level (Van Rensburg & Weideman 2002), and even more so in the case of students who have to major in a foreign language, the assessment of academic literacy is essential if students are to be prepared better for their undergraduate studies.

Since the introduction of the *doi moi* 'open-door' policy in 1986 in Việt Nam, English proficiency is increasingly being hailed as a vital requirement for employment prospects and economic cooperation with mainly capitalist countries (Thinh 2006:1-2). English is the main foreign language that is being taught and used in Việt Nam and hundreds of language centres have been established in the country. This is in keeping with the Education Development Strategy of the Ministry of Education and Training to enhance competitiveness of the Vietnamese economy through policy initiatives aimed at improving quality in education. Entrance examinations constitute part of the strategy to standardise and modernise higher education (Ministry of Education and Training 2011).

In accordance with government priorities and in response to concerns about the performance of its undergraduate students, the University of Đà Nẵng is applying a credit-based training system. One of its main characteristics is that it grants students the autonomy to choose appropriate courses within their areas of interest and ability, while the university has the

right to review its courses periodically on the basis of practical considerations. This system has proven to be problematic, however, and a number of studies have highlighted some of its shortcomings (Sơn 2010: 127). Cường (2006) suggests that one remedy would be to ensure that the academic ability of new-entrant students meets the demands of the university. Nonetheless, he does not specify measures for the assessment of such academic ability, nor how students can know whether, for example, their language proficiency will be adequate for success in the courses they have chosen, keeping in mind the potentially detrimental consequences of a social and financial nature should a student fail to graduate (see Weideman 2003b: 56).

It is evident from the above that an appropriate assessment of the academic literacy levels of first-time entering undergraduate students in the University of Đà Nẵng would be likely to inform decision-makers much more exactly about the nature and level of support such students might need. Based on the successes achieved by the TALL in South Africa as an identification measure for those who are at risk as a result of too low a level of academic literacy, an initiative was therefore launched to pilot this test at the University of Đà Nẵng and determine the appropriateness of its continued use beyond that. In order to place the decision to employ the TALL at the University of Đà Nẵng for that purpose in perspective, the following section examines the various sets of validation evidence for its use.

Validation evidence for TALL

Traditionally in the nineteen forties 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). A later understanding, that has in fact become the current orthodoxy, is that validation depends on the "interpretations and inferences" (Angoff 1988: 24) drawn from the scores the test yields, and the decisions resulting from these inferences. Although today many language testers subscribe to Messick's (1988) unitary view of test validity, under which all other test requirements are subsumed as being secondary, others are of the opinion that each of those requirements is necessary and worthy of being considered individually, since each contributes towards not only the validation of a test, but to its responsible (technical) design (Weideman 2009, 2011).

In line with the above exposition, if literacy tests are to be consistent and theoretically justifiable, they should incorporate a multiplicity of evidence (Bachman & Palmer 1996, McNamara & Roever 2006, Weideman 2009, 2011) to back their validation. Since the TALL has already been subjected to thorough scrutiny and analyses, ample evidence is available that it meets the requirements for usefulness set by Bachman and Palmer (see Butler 2009, Van der Slik & Weideman 2005, 2009, Van der Walt & Steyn 2007, Van Dyk 2010).

The TALL has consistently measured test scores under different testing conditions and thus meets the quality of *reliability*, which is a function of score consistency between different tests and tasks (Bachman & Palmer 1996). This implies that test scores remain consistent from one set of tests and tasks to another. Inter-rater reliability for scoring academic writing tasks is not a problem with the TALL, since it does not require test takers to write texts. The motivation for excluding writing activities is to be found both in the objective to eliminate marker bias, as well as the evidence that exists of a positive correlation between reading and writing ability

(Flower 1990, Hirvela 2004, Mponda 2010). Carson (1993) furthermore considers reading to be the basis for writing.

Reliability is also referred to as the internal consistency of a test and is often linked additionally to the requirement of fairness (the notion of *impact*; see below). Reliability is harder to achieve when the construct is complex and covers a range of language ability components and topical knowledge, such as in the case of the test specifications of the TALL. Nevertheless, the TALL has managed to obtain a Cronbach's alpha and Greater Lower Bound (GLB; Jackson & Agunwamba 1977) well in excess of 0.9 in its many administrations in South Africa, attesting to its reliability (Weideman & Van der Slik 2008).

Reliability is, furthermore, a prerequisite for *construct validity*, which is considered to be at the core of test usefulness. 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). Weideman (2009) offers a third and complementary 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 of 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).

Many attempts have been made to define academic literacy, but not all are equally functional for the purposes of designing an assessment instrument. Compare, for example, the view of academic literacy in the 1990s as "a compound of linguistic, conceptual and epistemological rules and norms of the academe" (Van Schalkwyk 2008: 22), or the approach of Kern (2000: 23) that literacy is "an elastic concept; its meaning varies according to the disciplinary lens through which one examines it". Weideman (2003a) has suggested adopting a componential view of academic literacy, that includes cognitively demanding tasks typically required of students at tertiary level, and that is also functional, listing 10 main subcomponents. These include an understanding of academic vocabulary, a sensitivity to genre, an ability to compare and contrast, or see sequence and order, the mastery of text comparison (text and context), visual and graphic literacy, and so on. Where the test designs that use a similar construct, such as the Proficiency Test in English for Educational Purposes (PTEEP) that was developed by the Alternative Admissions Research Project (AARP) of the University of Cape Town, have fewer task types, TALL operationalises each or several of these components into actual task types, providing a blueprint for the TALL (Van Dyk & Weideman 2004b). At the same time it should be clear that this definition meets the requirement for *authenticity* which Bachman and Palmer (1996) define as the degree to which test task characteristics correspond with those of the target language use in a non-test domain, i.e. typical analytical thinking, reading and cognitive tasks of an academic nature that students have to engage in.

Interactiveness also relates to the characteristics of the test tasks, but more specifically in terms of the way test-takers interact with these tasks (Bachman & Palmer 1996). Students are

challenged to draw on topical knowledge and make meaning of an academic text "beyond the level of the sentence" (Weideman 2003a: xi-xii). Moreover, test theme and content are selected very carefully, so as not to evoke negative connotations or cause psychological trauma to test takers, hence the selection of rather bland topics such as "rubber" for the test that features in this investigation.

The expected *impact* of the test is yet another quality of a language assessment instrument that Bachman and Palmer (1996) foreground, especially considering that test scores are used for certain purposes and have definite consequences. At the University of Đà Nẵng, the TALL was in the present case not used for access purposes, but for placement, and in that sense the test cannot have an enduring negative impact on the test-takers. Reception studies will be needed, however, to deal with perceptions of the test and to increase test-taker involvement.

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 reference to logistical constraints. The need for test results to be available within a day or two obviously precludes time-consuming hand-marking procedures. In all its years of administration in South Africa, the TALL has proven to be a fully practical means of assessing academic literacy.

It should be noted that the TALL tests which ICELDA develops are subjected to ongoing refinement through funding provided by the partnering institutions as part of the validation process.

Research methodology

The research questions for this study relate mainly to determining the appropriateness of employing the TALL for students from the CFL and to establishing how the Vietnamese students performed in comparison to their South African peers. Such a comparison is relevant in the light of the objective of the Ministry of Education and Training in Việt Nam to base educational level and quality on international standards. Universities in other parts of the world are thus being used for benchmarking purposes and the establishment of research partnerships is also encouraged (Ministry of Education and Training 2009). A further objective, if the test is deemed to be sufficiently appropriate, is to identify test items which need refinement or modification in subsequent rounds of testing at the University of Đà Nẵng.

The test selected for administration consisted of 100 questions in multiple choice format for completion within 90 minutes and included tasks aligned with each subcomponent of academic literacy identified by Van Dyk and Weideman (2004a, 2004b). The version of TALL used in this pilot consisted of a set of seven subtests as indicated in table 1.

Section 1: Scrambled text
Section 2: Vocabulary knowledge
Section 3: Verbal reasoning
Section 4: Interpreting graphs and visual information
Section 5: Register and text type
Section 6: Text comprehension
Section 7: Grammar and Text relations

Table 1: Items with low facility values and discrimination indices in the TALL pilot

No.	Section	Item	Disc. Index	Fac. Value
1	1	3	.17	.09
2	2	5	.12	.07
3		6	.15	.14
4		7	.15	.09
5		8	.21	.13
6		9	.30	.15
7		10	.05	.04
8		11	.18	.13
9		13	.18	.16
10		14	.09	.04
11		15	.05	.02
12	3	16	.13	.03
13		17	.05	.04
14		18	.17	.05
15		19	.20	.15
16		20	-.00	-.06
17	4	21	.23	.16
18		22	.03	.03
19		23	.18	.16
20		25	.17	.12
21		26	.05	.01
22		27	.09	.06
23		28	.07	.03
24		29	.07	.09
25		30	.06	.01
26	5	31	.09	.08
27		32	.12	.08
28		33	.17	.15
29		34	.24	.14

No.	Section	Item	Disc. Index	Fac. Value
30	6	38	.12	.01
31		42	.13	.13
32		43	.18	.14
33		44	.24	.17
34		48	.11	.10
35		50	.20	.16
36		56	.10	.05
37		59	.08	.06
38		60	.16	.10
39		62	.21	.13
40		64	.12	.09
41		65	.07	.01
42		66	.16	.01
43		71	.22	.17
44		72	.10	.14
45		73	.06	.05
46		74	.10	.05
47		75	.18	.10
48		77	.08	.08
49		79	-.04	-.07
50	7	81	.22	.13
51		87	.22	.13
52		88	.28	.13
53		89	.12	.06
54		95	.24	.11
55		98	.16	.12
56		99	.03	.03
57		100	-.08	-.14

Test population

For the pilot study at the University of Đà Nẵng, 197 first-year students majoring in English wrote the TALL. Most of these students (56.9%) had learnt English as a foreign language in Việt Nam for seven years or more. One student had been studying English for under three years. The remainder of the students had received between four and seven years of English language tuition. Though gender differences did not form part of this pilot study, we should note here that there were substantially more female students than male students, as is typical of English language classes in Việt Nam (178 were female students). The gender discrepancy in the number of Vietnamese students studying foreign languages is an aspect that researchers have recently started to investigate (Court 2001, Carr & Pauwels 2006). Only three percent of the Đà Nẵng students had been exposed to other language tests such as those administered by the Test of English as a Foreign Language (TOEFL), International English language Testing system (IELTS) or Test of English for International Communication (TOEIC), prior to taking the TALL.

In order to undertake a comparative study between the literacy levels of students in Việt Nam and those in South Africa, the results of the same TALL written by a cohort of students at the University of Pretoria were used. A total of 1819 first-year students representative of a variety of language backgrounds, formed part of the study. These students were also from different socio-economic and academic backgrounds and in this sense more heterogeneous than their Vietnamese peers (for the complete study, see Le 2011).

Data analyses

Apart from examining the distribution of scores and comparing the averages obtained for the respective cohorts of test-takers, test developers employ numerous empirical procedures to establish quantitative measures of test reliability and validity. An item analysis is helpful to indicate the facility values and discrimination indices for each individual subtest item, while a factor analysis is generally used to indicate the heterogeneity or homogeneity of a test, i.e. if the test measures a single ability or more than one ability (see Van der Slik & Weideman 2005).

Item analysis

The *Iteman* program was used to examine the contribution of each item to the test as a whole and the extent to which each item distinguished between weaker and stronger test takers. The higher the discrimination index, the better the test item will be able to make this distinction and the greater the reliability of the test will be. Theoretically, the maximum discrimination index is the value of one. A discrimination index of zero indicates a very weak discrimination. Additionally, items in which weaker test takers perform better than their stronger peers would be identified as possessing a negative discrimination index, and would thus be undesirable. Facility values form part of the item analysis and refer to the percentage of correct answers for the whole of the test population. In the instance of a placement test such as TALL, the facility values should show a wide range of values instead of big gaps between scores (Hughes 2003).

Factor analysis

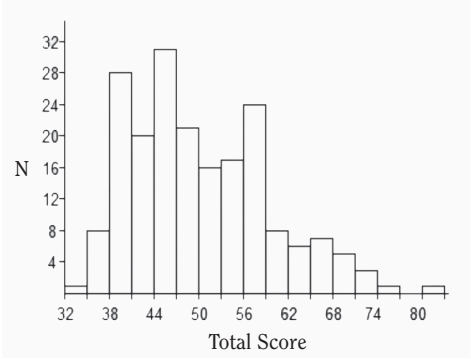
The *TiaPlus* program (CITO 2005) was utilised to indicate the possible existence of clusters of variables. The factor analysis it yields refers to a collection of statistical methods to study

the way underlying test constructs influence the responses based on measured variables (DeCoster 1998). Explained differently, this provides insight into the consistency of items in a test and whether they are homogeneous or one-dimensional in what they set out to measure, or heterogeneous with an adequate measure of homogeneity (Van der Slik & Weideman 2005). When working with a complex construct such as academic literacy, a degree of heterogeneity is to be expected.

Research findings

Comparison of the academic literacy levels of the two test cohorts

The scores of the students from the Centre for Foreign Language (CFL) at the University of Đà Nẵng (UD) were lower than those of their peers at the University of Pretoria. This difference can be attributed to the fact that the Vietnamese students learn English as a foreign language, while the South African students use English as a second language or even as their language of instruction, which means that they have had significantly more exposure to the language than their peers in Việt Nam have had. The distribution of scores as provided by the Iteman programme for classical item and test analysis (Guyer & Thompson 2011) is depicted in figures 1 and 2 below.



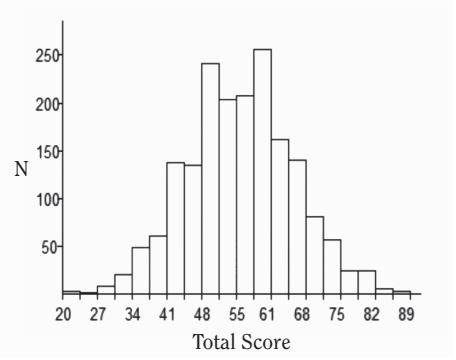


Figure 1: Frequency distribution of CFL scores *Figure 2: Frequency distribution of UP scores*

As can be seen, the distribution of the TALL test scores in the instance of the CFL students is skewed to the left (0.690) with a kurtosis of 0.037, whereas the scores of the UP students are skewed to the right (.031) with a kurtosis of -0.026. A comparison of the two sets of scores on the basis of a scale analysis generated by the *Iteman* programme will help to clarify the difference in the performances of the two cohorts of test takers.

Table 2: Scale statistics for scores of students from CFL and UP

	CFL	**UP**
Number of items	100	100
Number of examinees	197	1819
Mean	49.69	55.25
Standard Deviation	9.409	10.542
Variance	88.528	111.133

Minimum	30	20
Maximum	83	89
Alpha	0.774	0.831
Standard Error of Measurement	4.472	4.328
Skew	0.690	.031
Kurtosis	0.037	-0.026

Depending on where each institution wishes to set the cut-off point (indicating students at risk as regards academic literacy), that decision will indicate, either on the basis of historical or other data, what it considers to be the appropriate levels of academic literacy required for success at university. Of course, cut scores may be influenced also by logistical considerations, such as the availability of adequate interventions for potentially larger numbers of students than expected.

Measuring test reliability and construct validity

With regard to the process of validation and the reliability of the TALL, of particular importance in table 2 is the Cronbach's alpha of 0.774 for students from the CFL and 0.831 for students from the UP. These values are entirely acceptable since the TALL is not a high-stakes test that is used for access purposes. In instances where a test constitutes one of several types of information on the basis of which certain inferences are made, an alpha value of at least 0.8 is desirable. A value of 0.9 is considered to indicate a high level of reliability (see Hogan 2007 for a more detailed explanation). Small cohorts of test takers, such as in the case of the CFL, tend to produce lower alpha values. The Cronbach's alpha, which is considered the most conservative statistical test for reliability, provides the coefficient of internal consistency on a scale from 0 to 1, and also "depicts the degrees to which the observed scores represent the 'true' scores", in other words without measurement error (Van der Slik & Weideman 2005: 26).

Another important measure of reliability used in the analysis of multidimensional tests is Greatest Lower Bound (GLB). The GLB is a measure comparable to Cronbach's alpha, but one that does not assume homogeneity (Jackson & Agunwamba 1977). It is thus likely to be higher than Cronbach's alpha if the construct being measured is multidimensional (Van der Slik & Weideman 2005). Since academic literacy is a rich and complex notion, we may expect the GLB to be higher than the Cronbach's alpha and indeed it is. Using *TiaPlus* (CITO 2005) the GLB was calculated to be 0.91 for the combined scores of the CFL and UP students (compared to the combined Coefficient alpha of 0.82), a very satisfactory measurement for a pilot test with multidimensionality, as depicted in table 3. A certain minimum number of test takers is required to obtain acceptable results for the calculation of the GLB, hence the necessity of using the combined scores of the CFL and UP students. The high coefficient alpha compares well with that of the *Iteman* measurement.

In addition, the *TiaPlus* analysis shows that the number of potential misclassifications, i.e. the percentage of test takers who may have failed the test when they should have passed and vice versa, is low, varying under different scenarios between 0.6% and 1.05% of test takers

Table 3: Results from TiaPlus for CFL+ UP students' scores

Number of persons in the test	2016	Number of items	100
Average test score or Mean	53.73	Standard deviation	10.25
Greatest Lower Bound	0.91	Average Rit	0.24
Coefficient alpha	0.82	Standard error alpha	0.01
Average P-value	53.73	Standard error of measurement	4.31

who might need a second chance test. From the average P-value (53.73) it may be determined that the test was of average difficulty for the general ability of the pilot test takers: the higher the P-value, the easier the test items.

Reference has already been made to the importance of the discrimination power of test items. Using *TiaPlus*, one finds that another measure of discrimination, the Average Item Rest Correlation (Average Rit) is indicated as 0.24 in table 3. The Rit may range between -1.0 to +1.0. If the Rit is negative, it indicates that test takers with a high total score tended to answer an item incorrectly, and vice versa. If, on the other hand, the Rit is 0, there is no correlation between the performance on the test as a whole and an individual item. Where the Rit is positive and high, it indicates that test takers with high total scores also tended to answer a particular item correctly, whereas those with low total scores would probably have given a wrong answer to the same item. The Rit thus indicates item consistency, which means the test items contribute substantially to the reliability of the test as a whole. In the case of this TALL pilot, the average Rit of 0.24 indicates that there is a satisfactory relation between performance on the items and the test as a whole. The following table shows how each subtest contributes to the reliability of the test and its construct validity on the basis of the *TiaPlus* analysis.

Table 4: Statistical results of subtests related to the CFL and UP students' scores on the TALL

No. of examinees	**2016**							
Test section	Total test	1	2	3	4	5	6	7
No. of items	100	5	10	5	10	5	45	20
Average test score	53.73	1.63	7.56	2.61	6.26	2.88	21.06	11.72
Standard deviation	10.25	1.45	1.77	1.04	1.77	1.10	5.48	4.44
SEM	4.31	0.81	1.25	0.99	1.25	0.77	2.95	1.81
Average Rit	0.24	0.76	0.44	0.47	0.43	0.66	0.27	0.50
Average P-value	53.73	32.60	75.64	52.24	62.61	57.58	46.81	58.59
Coefficient Alpha	0.82	0.69	0.50	0.10	0.50	0.51	0.71	0.83
Greatest Lower Bound	0.91	0.77	0.56	0.17	0.57	0.62	0.80	0.93

From the information in table 4 it is evident that the Rit values are all positive, an indication that there is a healthy correlation between test items and test performance. The only troublesome subtest is that of test section 3, Verbal Reasoning, with its coefficient alpha and GLB of 0.10 and 0.17 respectively. This strongly indicates that this section of the test should be omitted from a subsequent version or redeveloped and repiloted.

Subtest intercorrelations

Test construct validity is also reflected via the correlations between the subtests of the test. Van der Walt and Steyn (2008) advise correlation values of between 0.15 and 0.5 to maintain multidimensionality. If these values exceed the mentioned parameters, for example 0.8 or 0.9, the subtests may be measuring the same concepts. On the other hand, correlation values of lower than 0.15 may negatively affect the integrity of the test as a whole. Note should be taken that the correlation between the subtests and the total test should preferably be high, as "the total score is taken to be a more general measure of the attribute than is each individual section score" (Van der Walt & Steyn 2008: 196). Table 5 indicates the subtest correlations.

Table 5: Intercorrelations between the subtests and total test

	Subtest	**Total test**	**1**	**2**	**3**	**4**	**5**	**6**
Scrambled text	1	0.40						
Academic vocabulary	2	0.43	0.15					
Verbal reasoning	3	0.34	0.12	0.15				
Graphic and visual	4	0.47	0.16	0.09	0.14			
Register and text	5	0.35	0.15	0.15	0.10	0.14		
Text comprehension	6	0.82	0.27	0.28	0.24	0.29	0.22	
Grammar and text	7	0.63	0.08	0.10	0.07	0.16	0.10	0.24

The correlations between the subtests lie predominantly between 0.15 and 0.5, which is satisfactory. Once again it is clear that the Verbal Reasoning subtest does not perform well, which is another indication that it should be removed. Incidentally, this subtest is not normally part of the TALL, but was added for experimental design purposes, since this kind of subtest is reputedly used in other tests of academic literacy in South Africa that are claimed to be based on the same construct. The empirical indications, however, are to the contrary.

Dimensionality

Figure 3 displays the multidimensionality of the TALL pilot and the degree of homogeneity of test items.

Most of the items cluster in the upper left corner of the graph in figure 3, which shows that these items measure similar traits. The items in subtest 7 (Grammar and Text Relations) are situated in the lower left and middle right corners, which shows that the test has a multi-dimensional element. Van der Slik and Weideman (2005) explain that a possible reason for this is that these items measure different aspects of academic literacy and that the construct

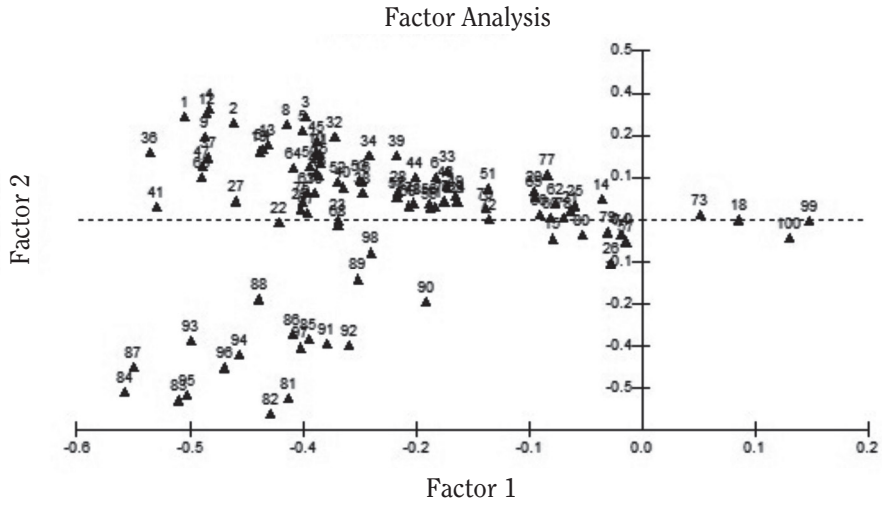


Figure 3: Factor analysis of the TALL pilot using TiaPlus

of the TALL cannot be reduced to a single homogeneous concept. Nonetheless, the level of homogeneity evident from the factor analysis is satisfactory for a test such as the TALL and the gull-wing shape found in figure 3 resembles that detected in other versions of the test.

The appropriateness of the TALL

The research findings reveal that the TALL is appropriate for administering to first-year students from the CFL at the University of Đà Nẵng. Not only was the test found to meet the requirements for tests in terms of reliability, but the multiplicity of evidence generated through the task types showed high construct validity as well. The average difficulty expressed in the P-value of the test also matched the general abilities of the students. From a practical perspective, the test was easy to administer and results could be made available soon afterwards, an aspect that is most conducive for placement purposes. The test also brings the university one step closer to ensuring that the academic literacy of the students meets the demands of the institution, one of the challenges posed by the new credit-based system. Nonetheless, as pointed out, the process of validation is ongoing and certain recommendations for the future employment of the TALL in Việt Nam are in order.

Further refinement of the test items is possible by studying more closely the Alpha-Rest values for each item. This measure indicates whether the item contributes positively towards the reliability of the test. Moreover, on the basis of *Iteman* results, items with low facility values and low discrimination can also be refined or substituted with more productive items. However, it would be unwise simply to remove these items until the number of test takers in Việt Nam has increased considerably and more empirical data are available. The removal of the Verbal Reasoning subtest is suggested as a necessary and immediate first process of refinement in the light of its low reliability coefficient and troublesome correlations with other subtests.

The ideal length of a test of proficiency is a matter of debate. In the case of the students at the University of Đà Nẵng the 90 minute time allocation was found to be longer than necessary

for the purposes of a placement test. It is thus recommended that the time be reduced to 55 minutes and the number of questions to 60, as in other current South African administrations of TALL. The effect hereof will have to be assessed carefully, however, as reducing the number of tasks can prove to have an adverse effect on both reliability and construct validity, since in the process the amount of evidence that can be generated will be reduced. Financial and practical considerations should never outweigh the importance of having sufficient evidence on which to base any decisions or inferences.

One of the limitations of the study was that it did not provide for investigations into gender and cultural differences. Differential item functioning (DIF) analyses may be considered in future research studies to ascertain whether any test items show bias towards any of the test takers. An ethnographic study could also yield information of importance, especially considering that the Vietnamese students did not perform as well as the South African students did. It is interesting to note that in the case of both cohorts of students, the TALL revealed that inadequate academic literacy levels posed a real threat to the successful completion by students of their college and university studies. Designing and tracking support measures of a developmental nature would constitute a logical further area of investigation.

In pursuance of the requirement of assessing the impact of the TALL, students should be familiarized with the test format prior to taking the test. On-line versions can be used for this purpose to ensure that candidates have the opportunity to perform at their best. Engaging the test-takers in a reception or qualitative study would further enhance their involvement in the testing process.

Finally, as regards the employment of a version of TALL abroad, this study has shown that it is not only robust enough to yield appropriate interpretations of results, but also has a design that is sufficiently flexible to develop a contextually appropriate test.

REFERENCES

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

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.

Carr, J. & Pauwels, A. 2006. *Boys and foreign language learning: Real boys don't do languages.* New York: Palgrave MacMillan.

Carson, J. 1993. Reading for writing. In Carson, J. & Leki. I. *Reading in the composition classroom: Second language perspectives*. Boston: Heinle & Heinle. Pp. 85-104.

Cheng, L. 2008. The key to success: English language testing in China. *Language testing*, 25(1): 15-37.

CITO. 2005. *TiaPlus, Classical Test and Item Analysis*©. Arnhem: Cito M. & R. Department.

Court, K. 2001. Why are boys opting out? A study of situated masculinities and foreign language learning.

[Online]. Available: www.ling.lancs.ac.uk/groups/crile/docs/crile57court.pdf. Accessed on 10 June 2011.

Cường, D.Q. 2006. Bàn về mô hình các yếu tố tổ chức quản lý chất lượng đào tạo cử nhân ngoại ngữ

theo học chế tín chỉ. *Kỷ yếu hội thảo VUN Đà Nẵng.* [Online]. Available: http://www.cdspbrvt.edu.vn/portal/uploads/temp/vun2.pdf. Accessed on 11 June 2011.

(*Translation*: Discussion of elements of education quality management models in the training of undergraduates of foreign languages in a credit-based system. *Proceedings of VUN Đà Nẵng.*)

DeCoster, J. 1998. Overview of factor analysis. [Online]. Available: http://www.stat-help.com/factor.pdf. Accessed on 10 June 2011.

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.

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

HESA (Higher Education South Africa). 2011. The National Benchmark Tests (NBTs). [Online]. Available: http://www.nbtests.co.za. Accessed 16 November 2011.

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.

Hughes, A. 2003. *Testing for language teachers*. Second edition. Cambridge Language Teaching Library. Cambridge: Cambridge University Press.

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

IELTS (International English Language Testing System). 2011. [Online]. Available: http://www.ielts.org. Accessed 16 November 2011.

Jackson, P.W. & Agunwamba, C.C. 1977. Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I. Algebraic lower bounds. *Psychometrica*, 42: 567-578.

Johnson, K. 2007. The stresses of Vietnam's exam season. *Time World*, 12 July. Available: http://www.time.com/time/world/article/0,8599,1643014,00.html. Accessed 16 November 2011.

Kern, R. 2000. Notions of literacy. In Kern, R. (Ed.). *Literacy and language teaching*. New York: Oxford University Press. Pp. 13-41.

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

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. 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.

Ministry of Education and Training in Việt Nam. 2009. Report on the development of higher education system, the solutions to ensure quality assurance and improve of education quality. [Online]. Available: http://en.moet.gov.vn/?page=1.0. Accessed 16 November 2011.

Ministry of Education and Training in Việt Nam. 2011. Report on the 2001-2010 Education Development Strategy. [Online]. Available: http://en.moet.gov.vn/?page=6.7&view=4404. Accessed 16 November 2011.

Mponda, O.T. 2010. Academic literacy at a European university: Undergraduate students' perceptions of academic literacy in English as a second language at the University of Groningen, in the

Netherlands. Master's thesis in Applied Linguistics. University of Groningen.

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

Sơn, L.Q. 2010. Những vấn đề của quản lý đào tạo theo học chế tín chỉ ở trường Đại học Sư Phạm. *Tạp chí Khoa học và Công nghệ, Đại học Đà Nẵng.* Số 6(41). [Online]. Available: http://www.udn.vn/bankhcnmt/zipfiles/so41/20-lequangson.pdf. Accessed on 11 June 2011.

(*Translation*: Problems in the administration of credit-based training at the Colleges of Education. *Journal of science and technology, the University of Đà Nẵng*, 6(41).

Thinh, Do Huy. 2006. The role of English in *Việt Nam's foreign language policy: a brief history*. Unpublished paper delivered at the 19th Annual English Australia Education Conference. 15 September, Perth, Western Australia.

TOEFL (Test of English as a Foreign Language). 2011. [Online]. Available: http://www.ets.org/toefl. Accessed on 16 November 2011.

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. 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): 140-155.

Van der Walt, J.L. & Steyn, F. jnr. 2008. The validation of language tests. *Stellenbosch papers in linguistics*, 38: 191-204.

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.

Van Rensburg, C. and Weideman, A. 2002. Language proficiency: current strategies, future remedies. SAALT Journal for language teaching. 36 (1&2), 152-164.

Van Schalkwyk, S.C. 2008. Acquiring academic literacy: A case of first-year extended degree programme students at Stellenbosch University. PhD dissertation, Faculty of Education, Stellenbosch University.

Weideman, A. 2003a. *Academic literacy: Prepare to learn.* 2nd edition. Pretoria: Van Schaik.

Weideman, A. 2003b. 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* Special issue: Assessing and developing academic literacy (ed.: J. Geldenhuys) 27(3), 235-251.

Weideman, A. 2011. Academic literacy tests: design, development, piloting and refinement. Submitted to special edition of SAALT *Journal of language teaching*.

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.

ABOUT THE AUTHORS

Albert Weideman
Department of English
University of the Free State
Email: albert.weideman@ufs.ac.za