Acta Academica 2008 40(1): 161-182

Albert Weideman & Frans van der Slik

The stability of test design: measuring differences in performance across several administrations of an academic literacy test

First submission: August 2006

This study explores the stability or consistency across several administrations of a test designed to determine academic literacy levels. The reliability of two versions (one English, the other Afrikaans) of such a test used for placement purposes at three South African universities will first be analysed. Secondly, the number of potential misclassifications on the test is assessed (ie the extent to which the test does not measure fairly). Thirdly, the differences among the results of the various administrations of the test are explored, with a view to es-tablishing whether such differences are both significant and relevant.

Stabiliteit in die ontwerp van toetse: die meting van prestasieverskille in 'n akademiese geletterdheidstoets oor verskeie aanwendings heen

Hierdie bydrae verken die stabiliteit of konsistensie van 'n toets van akademiese geletterdheid oor verskeie aanwendings heen. Eerstens word analises aangebied van die betroubaarheid van twee weergawes (die een in Engels, die ander in Afrikaans) van toetse wat aan drie Suid-Afrikaanse universiteite vir plasingsdoeleindes gebruik word. Tweedens analiseer ons die getal potensiële misklassifikasies, dit wil sê die mate waarin die toets nie billik meet nie. Derdens ondersoek ons verskille in die resultate in afsonderlike aanwendings van die toets, en vra spesifiek of sulke verskille beduidend sowel as relevant is.

Prof A J Weideman, Unit for Academic Literacy, University of Pretoria, Pretoria 0002; E-mail: albert.weideman@up.ac.za; Dr F van der Slik, Research Associate, Unit for Academic Literacy, University of Pretoria, and Associate Professor, Dept of Linguistics, Radboud University of Nijmegen, P O Box 9103, 6500 HD Nijmegen, The Netherlands; E-mail: f.v.d.slik@let.ru.nl

This paper discusses the design of a test of academic literacy, and how that design may result in unfair measures of student performance when the test is administered in a variety of institutional contexts. A limited range of potential indicators of difference among the various administrations will be examined, such as analyses that show differential item functioning (DIF). Subsequent studies will investigate whether the tests demonstrate any gender bias, and whether they are stable across different years.

The context of the paper is many South African universities' current use of tests of academic literacy either as access mechanisms (cf Cliff *et al* 2003, Visser & Hanslo 2005) or for placement purposes, ie for determining what level of risk the student poses in terms of academic literacy. In some cases, institutions of higher education use a single test for both purposes. Although it might be argued that such a practice deserves critical consideration, since variations in test purpose or test use may influence design, we do not think that there is any theoretical or other reason why such tests should not be built on the same construct. One would, however, expect the former type to be more reliable, and therefore likely longer, since the use to which the results are put makes such tests high-stakes tests (permitting access to a university qualification, and the associated increased earning power). A test of academic literacy designed and used only for placement purposes, ie to determine what, if any, level of academic literacy support is required after the student has gained access — such as the *Test of Academic Literacy Levels* (*TALL*) or its Afrikaans equivalent, the *Toets van Akademiese Geletterdheidsvlakke* (*TAG*) — is not a high-stakes, but a medium- to low-stakes test. By the time students are required to take the *TALL*/*TAG*, the political questions — the issues of power, ie of access to higher education — have already been answered. The effects of the results of the *TALL*/*TAG* are at present limited to submitting to an intervention: a compulsory academic literacy course (Weideman 2003) intended to assist students in eliminating one of the factors most closely associated with lack of academic success and poor performance (Van Rensburg & Weideman 2002).

A test is an applied linguistic artefact (Weideman 2006), specifically an instrument of measurement, which both its designer and its users would expect to measure fairly, irrespective of whether its results were associated with high stakes, or with a medium-to-low impact. As Van der Slik (2006) has pointed out, the fairness with which a test measures is crucially dependent on its reliability, which can be defined either as its internal consistency or as its consistency across various administrations. If a test yields variable or inconsistent results when administered to more or less similar populations (in the present case: new first-year students at various South African universities), it may not be robust enough to yield fair results.

The current study deals with the consistency of *TALL* and *TAG* across several administrations and various contexts. We have commented before (Van der Slik & Weideman 2005) on the test developers' quest for continued refinements to their test designs, and the value of the various available empirical analyses. Specifically, we have concluded that different measures yielded by the statistical properties of a test do not conflict with, but complement present-day concerns about transparency and accountability (Weideman 2006). Although not all empirical analyses are directly accessible to the general public, a first level of accountability for any test design must remain the production of analyses such as this. The purpose of such a set of analyses is to submit the test to specific scrutiny by others within the academic community who are concerned about or interested in issues of language testing. This first level of transparency and accountability does not obviate the need to make such tests more generally transparent and accountable to the public at large; the opposite is true. We therefore agree with Bygate's (2004) notion that applied linguists, including language test designers, have a dual accountability: an academic, technical accountability and a public accountability. For further discussion of the interaction between transparency, accountability and a number of related notions, we refer to the analysis and arguments presented in Weideman (2006).

The current study therefore once again takes its cue from Shohamy's (2001) exhortation to "tell the story of a test" as a necessary first step in the process of becoming transparent and, subsequently, accountable as test developers. In the present case this is limited, however, to telling the story of a specific dimension of the test: its consistency or reliability from a test designer's point of view. As will be demonstrated, this perspective on reliability may have positive consequences for those who take the test.

The question that we wanted the analysis to answer was: How stable are these tests across various administrations? One would expect variation, of course, especially where, as in the current case, the test has been administered to differently-composed populations. The variations in the composition of the three populations may affect the reliability with which the tests measure academic literacy. Nonetheless, one would expect such variations to remain within certain limits, since the populations also share a number of crucial characteristics: they were all new undergraduates at the three South African universities (Northwest, Pretoria and Stellenbosch) and they were all taking the academic literacy test for the purposes of placement.

This study is one of a series of reports on further analyses of the results of *TALL* and *TAG*. These tests are administered annually to all new undergraduates at several South African universities — Northwest University's (NW) Potchefstroom and Vanderbijlpark campuses, the University of Pretoria (UP), and the University of Stellenbosch (US). In 2006, first-year students in the Faculty of Medicine at the University of Limpopo (Medunsa campus) were also assessed by means of *TALL*.

1. Method

1.1 Population

In February 2005, the academic literacy of new undergraduates at Northwest University (the Potchefstroom and Vanderbijlpark campuses) and the Universities of Pretoria and Stellenbosch was tested. New first-year students at Pretoria and Northwest may choose their test language: English or Afrikaans. The University of Stellenbosch's first-year students had to take both tests, the Afrikaans one first, and the English one a day or so later. In total, 6,924 new students took the Afrikaans test (2,701 at UP; 1,702 at US; 2,521 at NW) and 5,174 the English version (3,310 at UP; 1,729 at US; 135 at NW).

1.2 The tests: *TALL* 2005 and *TAG* 2005

The 2005 versions of the *Test of Academic Literacy Levels* (*TALL*) and the *Toets van Akademiese Geletterdheidsvlakke* (*TAG*) consist of 80 and 82 items, respectively, distributed over seven sub-tests or sections (described in Van Dyk & Weideman 2004a), six of which are in multiple-choice format:

- Section 1: Scrambled text (5 items, 5 marks)
- Section 2: Knowledge of academic vocabulary (10 items, 20 marks)
- Section 3: Interpreting graphs and visual information (*TALL* 6 items, 6 marks; *TAG* 7 items, 7 marks)
- Section 4: Text types (5 items, 5 marks)
- Section 5: Understanding texts (*TALL* 19 items, 49 marks; *TAG* 20 items, 48 marks)
- Section 6: Text editing (15 items, 15 marks)
- Section 7: Writing (handwritten; marked and scored only for certain borderline cases, 20 marks)

Students have 60 minutes to complete the test, and they may earn a maximum of 100 marks (approximately half of the items counting 2 or 3 instead of 1).

1.3 Analysis

Two statistical packages were used to analyse the test results of the UP, US and NW students: SPSS and TiaPlus (Cito 2005). TiaPlus is a detailed test and item analysis package, which contains statistical measures at the level of the item as well as the whole test. These statistics were used to evaluate the empirical properties of the tests in this study. Descriptive statistics are presented, such as the average difficulty of the items (average *p*-value) and the average discriminative power of the items (average *Rit*: or average item-to-test correlation). At the test level, the reliability statistics used were Cronbach's α and Greatest Lower Bound (GLB) reliability (cf Verhelst 2000).

Since an academic literacy test — or any test, for that matter — is never entirely reliable, some candidates may fail who should have passed, and *vice versa*. TiaPlus provides four outcomes relating to the total number of potential misclassifications that could have occurred due to imperfect measurement (cf also Van der Slik & Weideman 2005).

One of our main questions was whether students from the UP, US, and NW performed differently on the *TALL* and *TAG* items. DIF-statistics like the Mantel-Haenszel test and Z-test were used to determine whether individual items displayed a difference in subgroup performance. Finally, T-tests and Cohen's *d* (cf Cohen 1988, 1992) were used to establish whether the students from the three universities performed differently on the various administrations of *TAG*, or differently on the three administrations of *TALL* as a whole, or in part.

1.4 Results

1.4.1 Description of the population

Tables 1 and 2 depict the outcomes at the scale level for *TALL* and *TAG*. Clearly, both tests are highly reliable in terms of *alpha* (for *TALL* 0.91, 0.86, and 0.92; for *TAG* 0.81, 0.91, and 0.83) and GLB reliability (for *TALL* 0.94, 0.91 and 0.98; for *TAG* 0.88, 0.94, 0.89). In addition, the average *Rit*-values, indicative of the discriminative power of the items, appear to be sufficiently high (for *TALL* 0.46, 0.37, 0.48; for *TAG* 0.31, 0.43, 0.33). Approximately 34% of those who took the English test at UP failed (indicated by the results of the test as being at risk in respect of their level of academic literacy). The figure for US was around 23%. At NW the cut-off is one point lower than at UP and US; nevertheless, the failure rate was rather high: 56%.

The mean test scores are in line with previous observations. In addition, the variation around the mean is smaller at US than at UP and NW, which implies that the English academic literacy of those at US is more homogeneous than that of those at UP and NW. This may be explained by the fact that, of the three student populations, the US cohort at present includes fewer students from previously severely disadvantaged backgrounds and more from either formerly privileged or only relatively disadvantaged backgrounds. This is a

Table 1: Descriptive statistics of the English version of the academic literacy test

	UP	US	NW
N	3.310	1.729	135
Number of items	60	60	60
Range	0-100	0-100	0-100
Mean /average *p*-value	71.75	76.89	59.70
Standard deviation	19.31	14.57	21.97
Cronbach's *alpha*	0.91	0.86	0.92
GLB	0.94	0.91	0.98[1]
Standard error of measurement	5.64	5.39	6.11
Average *Rit*	0.46	0.37	0.48
Cut-off point	68.5	68.5	67.5
Percentage failed	34.26	22.73	56.30

first sign of the variation that may be ascribed to the differences in the composition of the student populations. Since the US student population may in future begin to show the same kinds of variation as other comparable populations, we intend to follow up these initial analyses. We should therefore be able to present a more thorough explanation later. Of course, as we have already indicated above, the US students wrote both the English and the Afrikaans tests, which may well have had an influence on the results. Again, however, we would need to consider their performance in subsequent years before we would be able to present a more detailed argument and explanation.

For the Afrikaans test, the picture is somewhat different. Approximately 24% of the students at UP failed, as compared to around 27% at US. At NW the figure was somewhat higher, at 31%. However, the cut-off at US is lower than those at UP and NW.[2] The mean test scores at US and NW are about the same, while on average

[1] The GLB is not entirely reliable in cases with fewer than 200 candidates.

[2] A detailed discussion of the slight variations in cut-off points, and their justification, appears in Van der Slik & Weideman 2005.

students at UP performed better than those at US and NW. Clearly the variation around the mean is higher at US than at UP and NW, which implies that the Afrikaans academic literacy of those at US is less homogeneous than that of those at UP and NW. There is a fairly obvious explanation for this: students at US were not free to choose which language they preferred to be tested in; they had to take both the *TALL* and the *TAG*, even if they were not proficient in Afrikaans. As has been shown in another analysis of the 2005 *TALL*/*TAG* data (Van der Slik & Weideman 2006), mother tongue significantly affects performance on a test of academic literacy.

Table 2: Descriptive statistics of the Afrikaans version of the academic literacy test

	UP	US	NW
N	2.701	1.702	2.521
Number of items	62	62	62
Range	0-100	0-100	0-100
Mean /average *p*-value	70.16	63.15	63.08
Standard deviation	13.55	19.50	15.07
Cronbach's *alpha*	0.81	0.86	0.92
GLB	0.88	0.94	0.89
Standard error of measurement	5.91	6.00	6.18
Average *Rit*	0.31	0.43	0.33
Cut-off point	60.5	50.5	55.5
Percentage failed	23.84	26.85	31.14

1.4.2 Misclassifications

Tables 3 and 4 present the number of potential misclassifications based on four criteria.

As can be seen, the number of potential misclassifications on the English test varies between 432 and 256 at UP, between 246 and 152 at US, and between 16 and 11 at NW, depending on which criterion

Table 3: Potential misclassifications on the English version of the academic literacy test (percentage of this test population); the corresponding intervals around the cut-off points (in terms of standard deviations) are given in italics

	UP	US	NW
Alpha-based:			
Correlation between test and hypothetical parallel test	432 (13.0%) *63-74 (0.31)*	246 (14.2%) *63-74 (0.41)*	16 (11.8%) *64-71 (0.18)*
Correlation between observed and "true" scores	308 (9.3%) *65-72 (0.21)*	176 (10.2%) *66-72 (0.27)*	11 (8.4%) *64-71 (0.15)*
GLB-based:			
Correlation between test and hypothetical parallel test	360 (10.9%) *64-73 (0.26)*	213 (12.3%) *66-72 (0.27)*	not available
Correlation between observed and "true" scores	256 (7.7%) *66-71 (0.15)*	152 (8.8%) *67-71 (0.21)*	not available

is applied. It should be borne in mind, however, that approximately half of the misclassifications relate to candidates who passed when they could have failed. If we disregard this portion (giving them the benefit of the doubt), we need to concern ourselves only with the proportion (also approximately half) of the misclassifications that arise from those who failed when they could have passed. For UP, between 216 and 128 candidates who could have passed may have failed. Applying the same logic to the candidates at US and NW, 76 to 123 candidates may have failed undeservedly at US, and between 6 and 8 at NW. At NW these outcomes are somewhat unreliable, however, due to the low number of candidates who took the English version of the test there (*n* =135). In fact, this meant that we were unable to estimate the GLB-based number of potential misclassifications at NW.

Table 4: Potential misclassifications on the Afrikaans version of the academic literacy test (percentage of the test population); the corresponding intervals around the cut-off points (in terms of standard deviations) are given in italics.

	UP	US	NW
Alpha-based:			
Correlation between test and hypothetical parallel test	415 (15.4%) *57-63 (0.30)*	192 (11.3%) *46-55 (0.26)*	414 (16.4%) *52-59 (0.25)*
Correlation between observed and "true" scores	300 (11.1%) *58-62 (0.22)*	137 (8.1%) *47-54 (0.21)*	298 (11.8%) *53-58 (0.20)*
GLB-based:			
Correlation between test and hypothetical parallel test	349 (12.9%) *58-62 (0.22)*	157 (9.2%) *47-54 (0.21)*	343 (13.6%) *53-58 (0.20)*
Correlation between observed and "true" scores	250 (9.3%) *59-61 (0.15)*	112 (6.6%) *48-53 (0.15)*	245 (9.7%) *54-57 (0.13)*

To give a clearer picture, additional information about the intervals around the cut-off points where these misclassifications may occur, in terms of both raw scores and standard deviations, may be useful. For example, at UP the interval varies inbetween 3 to 6 points around the cut-off point of 68.5. In terms of standard deviations this variation is between 0.15 and 0.31.

On the Afrikaans test, the number of potential misclassifications of those who failed but might have passed the test varies between 208 and 125 at UP, between 196 and 56 at US, and between 207 and 123 at NW, depending on which criterion is applied (cf Table 4).

Again, additional information about the intervals around the cut-off points where these misclassifications may occur, in terms of raw scores as well as standard deviations, may be useful. For example, at US the interval varies inbetween 3 to 5 points around the cut-off point of 50.5. In terms of standard deviations this variation is between 0.15 and 0.26. In order to ensure fair treatment by the test, these measures should be used to eliminate undesirable results.

We return below to a discussion of how such analyses may affect the use of test results.

1.4.3 Differential item functioning (DIF)

Using TiaPlus, DIF-statistics were used to test whether students from the universities of Pretoria, Stellenbosch, and the Northwest performed differently on the *TALL* and *TAG* tests. If the Mantel-Haenszel Statistic (Holland & Thayer 1988) is close to unity, the items are approximately equally difficult for the different groups of students. If, however, this DIF-statistic is either close to zero or larger than unity, then the corresponding item performs differently for the different groups. The associated Z-statistics show which item difficulties are considered to be statistically different ($p < 0.01$). We would like to emphasise, however, that due to the high numbers of candidates, even small differences in item difficulty are significant.

Analysis of the *TALL* items produced some interesting results. Items in the final part of the text editing section (sub-test 6) appeared to be more difficult for UP students than for US students (no differences were found with NW students, but this is not unexpected since only 135 NW students took the *TALL* test). Additional analyses (not shown here) have underlined this conclusion. More than 10% of the UP students failed to answer items 53 to 60, whereas only 2% or less at the US failed to finish these last 8 items in text editing. Does this mean that US students were more familiar with the content of the text editing section than UP students were? Might this explain the occurrence of these differences? Perhaps, but we think another explanation for the observed differences between UP and US students is more likely. The US candidates took the *TALL* only a few days after doing the *TAG*. It seems likely, therefore, that these outcomes may reflect a testing effect, rather than higher academic literacy in respect of text editing. To be more specific, the US students were already more acquainted with the form of this sub-test than the UP students were. This explanation is validated by two other observations. First, in the notes of those who take decisions on the cut-off points for various levels of results in the administrations of *TALL/TAG* at various institutions, there is a cautionary note regarding the 5% difference between the averages of the US and the UP administrations of *TALL* (to the effect that the test may be an easier test overall), as well as an observation that the US students had written wrote a test of roughly the same format and item types (*TAG*) a day or two before. Secondly, comparisons between students' higher levels of performance on various sub-tests when they have become increasingly familiar with the format of a test (cf Van der Slik & Weideman 2006) indicate that most learning takes place on this particular sub-test (text editing). In case this explanation turns out to be inadequate, it should also be considered whether the differences in the composition of the test populations may not perhaps have been responsible for the US students having experienced less difficulty in completing sub-test 6 — as some other preliminary analyses of the *TALL* and *TAG* 2006 data seem to suggest.

For illustrative purposes, we present in Figure 1 (left panel) the item functioning of *TALL* item 54.

Figure 1: DIF-graphics of *TALL* item 54 (left panel) and *TAG* item 13 (right panel)

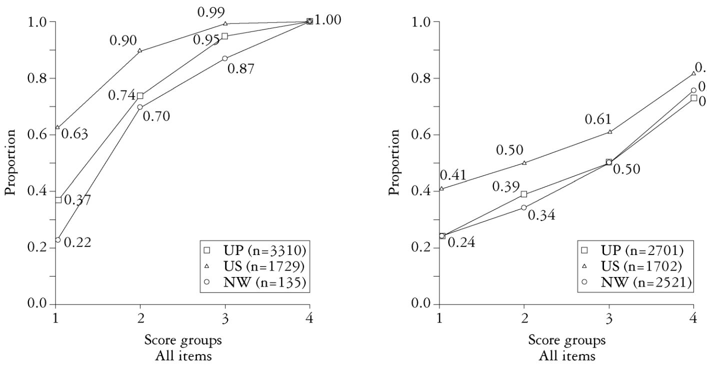


Figure 1 may be read as follows. The TiaPlus package has divided the candidates into four score groups. Score group 1 contains the 25% lowest scoring candidates on all 60 items (in the case of *TALL*) and all 62 items of *TAG*, respectively. Score group 4 consists of the 25% who scored highest, while score groups 2 and 3 fall in-between. In Figure 1 (left panel), it may be seen that *TALL* item 54 is more difficult for UP and NW students than for US students. This is particularly true of the lower scoring groups. Where, for example, only 37% of UP students in score group 1 had this item correct, 63% of US students in score group 1 gave the correct answer on *TALL* item 54. The same pattern was observed on items 53 through 60.

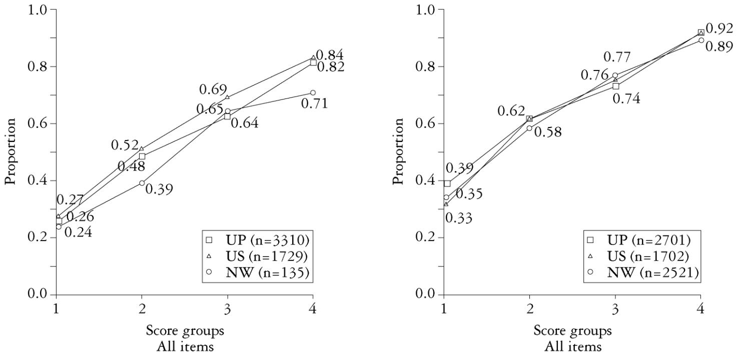
As far as *TAG* is concerned, there were only a few indications that the constituting items performed differently for the students of UP, US, and NW. For example, *TAG* item 13 appeared to perform differently for US students than for UP and NW students (the corresponding Z-values are -3.84 and 4.06, respectively, and are highly significant: $p < 0.001$). From Figure 1 it may be seen that, on average, US students performed better on item 13 than UP and NW students, but this was a rare exception.

Apart from these very few exceptions, however, we may conclude that, in general, the differences in the performances of the three test populations on individual items are negligible, since the scores of the four sub-groups remain close in almost all cases. Perhaps a more typical example than the two exceptional items referred to above is item 17 in *TALL* (on the left) and item 36 in *TAG* (on the right), given in Figure 2.

1.4.3 *T*-Tests and effect sizes

Finally, we have tested if the scores of UP, US, and NW students differ from each other in respect of the various administrations of *TALL* and *TAG*. In Tables 5 and 6 we present the outcomes of *T*-tests, not only for the entire tests, but also for the six sub-tests. In addition, we present Cohen's *d* (Cohen 1992: 157) in order to find out whether differences between students from the three universities, though possibly highly significant, are nevertheless trivial.

Figure 2: DIF-graphics of *TALL* item 17 (left panel) and *TAG* item 36 (right panel)



Apparently, candidates from UP and NW scored significantly lower on TALL than students from US (T = -10.60, $p < 0.0001$; T = 8.94, $p < 0.0001$, respectively). It has already been noted that, due to the large sample sizes, even trivial differences between scores might prove highly significant. For that reason we calculated Cohen's *d* (Cohen 1992: 157) in order to find out what the effect sizes actually were. It was found that the effect size (*d*) for the difference between the total score of UP and US was 0.29, which may be considered a rather weak effect size (Cohen 1992). The effect size for US against NW, however, was 1.13, which may be considered a strong effect. Clearly, the first difference may be called trivial; the latter far from trivial. A review of the effect sizes presented in Table 5 reveals that the differences between the scores of UP and US students are rather small, those between UP and NW students are medium, while those between US and NW candidates vary between medium and strong.

Table 5: *T*-Statistics (and effect sizes) for the English version of the academic literacy test and its parts

Sub-test	Max score	UP *vs* US	UP *vs* NW	US *vs* NW	UP Mean (SD)	US Mean (SD)	NW Mean (SD)
Sub-test 1	5	-6.41 *(0.18)*	3.86 *(0.36)*	5.62 *(0.62)*	4.02 *(1.43)*	4.27 *(1.21)*	3.50 *(1.56)*
Sub-test 2	20	0.04 *(0.00)*	4.87 *(0.43)*	4.79 *(0.44)*	13.83 *(3.96)*	13.83 *(3.83)*	12.12 *(4.01)*
Sub-test 3	6	-5.05 *(0.07)*	5.49 *(0.52)*	6.90 *(0.22)*	4.44 *(1.41)*	4.64 *(4.64)*	3.70 *(1.56)*
Sub-test 4	5	-6.16 *(0.18)*	2.43 *(0.22)*	4.36 *(0.40)*	3.83 *(1.30)*	4.06 *(1.28)*	3.54 *(1.34)*
Sub-test 5	49	-9.49 *(0.26)*	5.49 *(0.64)*	7.86 *(0.99)*	35.50 *(11.20)*	38.19 *(8.60)*	29.28 *(12.96)*
Sub-test 6	15	-16.86 *(0.44)*	5.91 *(0.57)*	10.03 *(0.96)*	10.12 *(4.48)*	11.90 *(2.94)*	7.56 *(4.95)*
Total test	100	-10.60 *(0.29)*	6.28 *(0.62)*	8.94 *(1.13)*	71.75 *(19.31)*	76.89 *(14.57)*	59.70 *(21.97)*

Table 6: *T*-Statistics (and effect sizes) for the Afrikaans version of the academic literacy test

Sub-test	Max score	UP *vs* US	UP *vs* NW	US *vs* NW	UP Mean (SD)	US Mean (SD)	NW Mean (SD)
Sub-test 1	5	2.90 *(0.09)*	7.44 *(0.21)*	3.65 *(0.11)*	3.40 *(1.81)*	3.23 *(1.88)*	3.02 *(1.89)*
Sub-test 2	20	6.64 *(0.21)*	13.34 *(0.37)*	4.39 *(0.14)*	13.65 *(3.81)*	12.79 *(4.37)*	12.21 *(3.95)*
Sub-test 3	7	5.27 *(0.11)*	10.27 *(0.28)*	3.36 *(0.16)*	5.59 *(1.39)*	5.43 *(1.64)*	5.17 *(1.56)*
Sub-test 4	5	5.76 *(0.19)*	6.67 *(0.19)*	-0.30 *(0.01)*	3.65 *(1.49)*	3.35 *(1.75)*	3.37 *(1.53)*
Sub-test 5	48	13.04 *(0.43)*	14.53 *(0.40)*	-1.87 *(0.06)*	31.82 *(7.88)*	27.93 *(10.61)*	28.51 *(8.55)*
Sub-test 6	15	13.63 *(0.46)*	14.13 *(0.39)*	-2.50 *(0.08)*	12.05 *(2.73)*	10.50 *(4.16)*	10.81 *(3.54)*
Total test	100	12.98 *(0.44)*	17.79 *(0.49)*	0.12 *(0.00)*	70.16 *(13.55)*	63.15 *(19.50)*	63.08 *(15.07)*

As far as *TAG* was concerned, candidates from UP scored significantly higher than students from US and NW ($T = 12.98, p < 0.0001$; $T = 17.79, p < 0.0001$, respectively), while there was no evidence of any difference between US and NW students. Cohen's *d* values show, however, that the total scores of UP students are not as different from those of US and NW as the *T*-values might suggest, because in Cohen's terms these differences remain in the medium range (0.44 and 0.49, respectively). The remaining effect sizes, presented in Table 6, reveal that the differences between the scores of UP, US and NW students are in the range of weak to medium; a conclusion which would not be drawn if only the *T*-values were taken into account.

Again, the differences between *TAG* and *TALL* on these measures are an indication of variations either in the composition of the student population, or in the administration of the test. In respect of the first variation, it would come as no surprise to the lay observer that the level of English academic proficiency at US was generally higher than that at its two northern counterparts, with the institution that traditionally takes on the highest number of students from non-urban backgrounds faring the worst. As far as the second variation is concerned, it seems obvious that the compulsory administration of TAG to all students at US relates to the finding that this group was less proficient in Afrikaans than, for example, the UP first-years.

2. Conclusion

These analyses have several implications for the design and administration not only of the tests of academic literacy under discussion here, but also for those of similar tests of academic literacy, such as the National Benchmark Test of Academic Literacy currently being developed under the auspices of Higher Education South Africa (HESA).

First, the generally high reliability measures observed (in terms of both Cronbach's α and GLB) indicate that the tests as they are currently designed have an acceptable level of internal consistency. Similarly, the fairly good discriminative power of the tests, as measured in average terms across items, indicates that the current design is doing what it should. Subsequent preliminary analyses of the 2006 results, not reported here, show similar consistency levels.

Secondly, in the variations in test measurement, specifically as these are manifested in the different estimates of misclassifications, we have a clear indication of a need to provide an administrative or other solution to the measure of potentially unfair treatment by the test. Of course, everyone accepts that tests are never perfect. What matters is the way that we deal with the known imperfections. Even though the current tests are not high-stakes tests but medium-to-low stakes assessments of academic literacy, there is a need, as in the application of any measurement instrument, to treat candidates as fairly as possible. The calculation of the number of potential misclassifications at all institutions indicates, therefore, that an identifiable, limited proportion of the test population should be given another opportunity of demonstrating their level of academic literacy. The format of this second-chance or borderline-case test should be similar to that of the current test. Because the tests are so reliable, however, the number of students eligible for this second opportunity will not be high. In addition, the test administrators need not necessarily use the misclassifications as in Tables 3 and 4 above, based on the more conservative reliability estimate (Cronbach's α). We have pointed out elsewhere (Van der Slik & Weideman 2005; cf also CITO 2005: 18, Jackson & Agunwamba 1977) that for other than homogeneous tests, GLB is in any event the more appropriate estimate of reliability. So the number of candidates who qualify for a second-chance test will vary, in the case of UP, for example, between 128 and 180 for *TALL* 2005 (cf Table 3). For these relatively modest numbers it should not be too difficult, administratively, to arrange such an opportunity at any of the institutions concerned, and the results of this part of our analysis indicate that we should indeed make such a recommendation to those who administer the test.

What is also relevant in the elimination of unfair treatment in this case is that, even though the number and size of the misclassification will obviously vary from one administration to the next, or between the administration of a version in different years, we now have a set of benchmarks (between 0.1 and 0.2 standard deviation for *TAG*, and between 0.1 and 0.3 standard deviation for *TALL*) for the identification of such potential misclassifications, which could be applied to subsequent administrations of the tests.

All of the above is relevant, of course, not only for the necessary technical elements that ensure fairness (validity and reliability), but also to achieve the social acceptance of a test. For example, a test should not stigmatise, and making available second and further assessment opportunities is a way both of ensuring acceptance and of limiting stigmatisation. One way in which this may be achieved is by reflecting on, analysing and making public as much information about a test as possible (cf eg Unit for Academic Literacy 2006). As Weideman (2006) has pointed out, an applied linguistic instrument such as a test needs to possess not only a number of necessary technical elements (reliability, validity, and a theoretically justified construct) but also the additional components of transparency and accountability.

Thirdly, the analysis of the effect-sizes of the variations in the test performances of the different populations indicates to us that these could be the initial pointers to further parameters that the test designers may wish to set for variation in the way that the tests measure. While some of the variations on the total scores (as indicated by Cohen's *d*) are weak (as low as 0.29, for example; cf Tables 5 and 6) or medium (between 0.44 and 0.62), the relatively strong variation of 1.13 on two of the three administrations of *TALL*, though explicable in terms of the composition of the two populations, indicates a need for vigilance in relation to such differences in future. Should subsequent administrations of the test reveal growing differences in ability, especially where such differences can be explained in terms of the composition of the first-year student body, this may have implications beyond the initial purpose of the test.

In positive terms, however, these calculations indicate that the test designers may be able to set parameters for some of the significant and non-trivial differences between candidates' performances on the test. Should a test measure outside of these parameters, it would merit special attention.

Finally, we have reported here on only a limited number of measures of consistency. We intend to follow up these initial analyses with further analyses of the stability of the tests in question. We will only be able to carry out these further analyses, however, once we have made certain adjustments to the versions of the tests currently under construction, in order to make them amenable, for example, to different kinds of analyses beyond classical test theory. In general, we are at this point satisfied with the measures of stability reported on in this paper. These initial analyses indicate that we have a set of robust measuring instruments.

Bibliography

Bygate M

2004. Some current trends in applied linguistics: towards a generic view. *AILA Review* [Association Internationale de Linguistique Appliquée] 17: 6-22.

Cito

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

Cliff A F, N Yeld & M Hanslo

2003. Assessing the academic literacy skills of entry-level students, using the Placement Test in English for Educational Purposes (PTEEP). Bi-annual conference of the European Association for Research in Learning and Instruction (EARLI), Padova, Italy.

Cohen J

1988. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.

1992. A power primer. *Psychological Bulletin* 112: 155-9.

Holland P W & D T Thayer

1988. Differential item performance and Mantel-Haenszel. Wainer & Braun (eds) 1988: 129-45.

Jackson P W & C C Agunwamba

1977. Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I. Algebraic lower bounds. *Psychometrica* 42: 567-78.

Shohamy E

2001. *The power of tests: a critical perspective on the uses of language tests*. Harlow: Pearson Education.

Unit for Academic Literacy

2006. Compulsory academic literacy test. <http://www.up.ac.za/academic/humanities/eng/eng/unitlangskills/eng/fac.htm>

Van der Slik F

2005. Statistical analysis of the TALL/TAG 2004 results. Presentation to Test Development session, 1-3 June 2005. University of Pretoria.

2006. Language proficiency and fairness. Keynote address, Southern African Applied Linguistic Association 2006, Durban, 6 July.

Van der Slik F & A Weideman

2005. The refinement of a test of academic literacy. *Per linguam* 21 (1): 23-35.

2006. Measures of improvement in academic literacy. Submitted to *Southern African Linguistics and Applied Language Studies*.

Van Dyk T & A Weideman

2004a. Switching constructs: on the selection of an appropriate blueprint for academic literacy assessment. *SAALT Journal for Language Teaching* [South African Association for Language Teaching] 38 (1): 1-13.

2004b. Finding the right measure: from blueprint to specification to item type. *SAALT Journal for Language Teaching* 38 (1): 15-24.

Van Rensburg C & A Weideman
2002. Language proficiency: current strategies, future remedies. *SAALT Journal for Language Teaching* 36 (1 & 2): 152-64.

Verhelst N D
2000. *Estimating the reliability of a test from a single test adminis-tration*. Measurement and Research Department Reports 98-2. Arnhem: National Institute for Educational Measurement.

Visser A & M Hanslo
2005. Approaches to predictive studies: possibilities and challenges. *SA Journal of Higher Education* 19(6): 1160-76.

Wainer H & H I Braun (eds)
1988. *Test validity*. Hillsdale, NJ: Lawrence Erlbaum.

Weideman A
2003. *Academic literacy: prepare to learn*. Pretoria: Van Schaik.

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