Forthcoming in *Acta Academica* 40(1)

Albert Weideman & Frans van der Slik

The stability of test design: measuring difference in performance across several administrations of a test of academic literacy

Summary

There are various measures of fairness for a test. The current paper explores one such measure: the stability or consistency of a test of academic literacy levels across several administrations. Analyses are made, first, of the reliability of two versions (one English, the other Afrikaans) of tests of academic literacy used for placement purposes at three South African universities. Second, we analyse the number of potential misclassifications of the test, i.e. the extent to which it does not measure fairly. Third, we explore the differences among the results of the various administrations of the test, asking specifically whether such differences are both significant and relevant. The paper concludes that such analyses as are presented here have a potentially benign effect, in that they provide an empirical basis for changes in the way that a test is administered, which may counter stigmatisation and promote accountability.

Stabiliteit in die ontwerp van toetse: die meting van verskille in prestasie op ’n toets van akademiese geletterdheid oor verskeie aanwendings heen

Die billikheid waarmee ’n toets meet, kan op verskeie maniere bepaal word. Hierdie bydrae verken een dimensie hiervan: 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 van akademiese geletterdheid 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. Die slotsom is dat analises soos dié wat hier aangebied word ’n positiewe uitwerking behoort te hê op die manier waarop ’n toets aangewend word, aangesien dit empiriese gronde bied waarmee stigmatisering vermy en groter verantwoorbaarheid bevorder kan word.

Prof A J Weideman, Director, 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, Department of Linguistics, Radboud University of Nijmegen, PO Box 9103, 6500 HD Nijmegen, The Netherlands; E-mail: f.v.d.slik@let.ru.nl

Introduction

This paper discusses the design of a test of academic literacy, and how that design may result in unfair measures of student performance when it is administered in a variety of institutional contexts. We examine here a limited range of potential indicators of difference, such as analyses that show differential item functioning (DIF) and other indicators of difference among the various administrations, and intend to follow these up in subsequent studies that will investigate whether the tests show gender bias, and whether they are stable across different years.

The context of the paper is the current use, by many South African universities, of tests of academic literacy either as access mechanisms (cf Cliff, Yeld & Hanslo 2003, Visser & Hanslo 2005) or for placement purposes, i.e. for determining what level of risk the student shows as regards academic literacy. In some cases, institutions of higher education use a single test for both purposes. Even if one may argue that such a practice deserves critical consideration, since variations in test purpose or test use may influence the design, we nonetheless think that there is no theoretical or other reason why such tests cannot be built on the same construct. One would expect the former kind, however, to be more reliable, and therefore more likely longer, since the use of its results makes such tests high stakes tests (they give access to a university qualification, and the increased earning power associated with that). A test of academic literacy, such as the *Test of Academic Literacy Levels* (TALL), or TAG (*Toets van Akademiese Geletterdheidsvlakke*), as its Afrikaans version is called, which is designed and used only for placement purposes, i.e. to determine what level of academic literacy support is required (if at all) **after** the student has gained access, is for that reason not a high stakes, but a medium to low stakes test. Currently, at the time when students are required to sit for TALL/TAG, the political questions — the issues of power, i.e. who has gained access to higher education study — have been answered. The effects of the results of TALL and TAG are at present limited to submitting to an intervention, a compulsory academic literacy course (Weideman 2003), which is intended to assist students in eliminating one of the factors most closely associated with lack of academic success and performance (Van Rensburg & Weideman 2002).

A test is an applied linguistic artefact (Weideman 2006), and specifically it is an instrument of measurement, which the responsible designer and users would expect to measure fairly, irrespective of whether its results are 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 its consistency across various administrations. Should a test yield 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 perhaps not have the robustness to yield fair results.

The current paper 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 developer's quest to make continued refinements to their test designs, and the value, in doing so, of various empirical analyses that are available. 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). Even though all empirical analyses are not directly accessible to the general public, a first level of accountability for any test design must remain the production of analyses such as this one. The purpose of such a set of analyses is that it allows the test to be specifically scrutinised by others within the academic community who are concerned about or interested in issues of language testing. This first level of transparency and becoming accountable does not obviate the need to make the tests more generally transparent and accountable to the public at large; quite the opposite is true. We therefore agree with Bygate's (2004) notion that applied linguists, including language test designers, have a dual accountability, viz. an academic, technical accountability, as well as 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 paper 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 it 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 we shall see below, this perspective on reliability may have positive consequences also for those who take the test.

The question that we wanted to answer in the analysis was: How stable are these tests across their different administrations? One would expect variation, of course, especially where, as in the current case, the test has been administered to populations that are composed differently. These 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 are new undergraduate students at three different South African universities (Northwest, Pretoria and Stellenbosch), and the test they take is a test of academic literacy, taken for the purpose of placement.

The paper we offer here is one of a series of reports on further analyses that we have done on the results of TALL and TAG. These tests are administered annually to all new undergraduate students 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.

Method

Population

In February 2005, the academic literacy of new undergraduate students of Northwest University (Potchefstroom and Vanderbijlpark campuses) and of the Universities of Pretoria and Stellenbosch was tested. New first year students of Pretoria and Northwest may choose which language they want to be tested in, i.e. either in English or in Afrikaans. The University of Stellenbosch first year students had to take both tests. They took the Afrikaans test first, and one or more days later they sat for the English test. In total, 6,924 new students participated in the Afrikaans test (2,701 UP; 1,702 US; 2,521 at NW), while 5,174 students took the English version (3,310 UP; 1,729 US; 135 at NW).

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 and 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, six 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 points (approximately half of the items counting 2 or 3 instead of 1).

Analysis

In order to analyze the test results of the UP, US and NW students, we made use of two statistical packages: SPSS and TIAPLUS (Cito 2005). TIAPLUS is a detailed test and item analysis package, which contains statistical measures at the item as well as the test level. These statistics have been used to evaluate the empirical properties of the tests in this study. We present descriptive statistics like 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 we make use of

the reliability statistics Cronbach's α and *GLB* or Greatest Lower Bound reliability (see Verhelst 2000).

Since an academic literacy test – or any test, for that matter – is never entirely reliable, some testees may fail were they should have passed, and vice versa. TIAPLUS provides four outcomes regarding the total amount of potential misclassifications that could have occurred due to imperfect measurement (see also Van der Slik & Weideman 2005).

One of our main questions was whether students from 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 display a difference in sub-group (UP, US, NW) performance. Finally, we used T-tests and Cohen's *d* (cf. Cohen 1988, 1992) in order to find out if the students from the three universities performed differently on the various administrations of TAG, and differently on the three administrations of TALL as a whole, and of their parts.

Results

Description of the population

Table 1 and 2 depict the outcomes at the scale level for TALL and TAG. Clearly, the TALL as well as the TAG are highly reliable, both in terms of alpha (for TALL: .91, .86, and .92, for TAG: .81, .91, and .83, respectively) and GLB reliability (for TALL: .94, .91 and .98, for TAG: .88, 94, 89, respectively). In addition, the average *Rit*-values, indicative of the discriminative power of the items, appear to be sufficiently high as well (TALL: .46, .37, .48; and .31, .43, .33 for TAG). It can be observed that approximately 34% of those who took the English test at UP fail (i.e. are indicated by the results of the test as being at risk in respect of their level of academic literacy), while around 23% of those who took this test at US did not pass. The number of students who did not pass at NW is rather high: 56%, despite the fact that at NW the cut-off point is one point lower than at UP and US.

The mean test scores are in line with the former observations. In addition, the variation around the mean is smaller at US than at UP and NW, implying that the academic literacy of those at US is more homogeneous than the academic literacy of those who took the English test at UP and NW. This may be explained by the US student population at present having fewer students from previously severely disadvantaged backgrounds, and a greater proportion from either formerly privileged or from only relatively disadvantaged backgrounds, and this is a first sign of variation that may be ascribed to the differences in the composition of the various student populations referred to above. 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 to see whether they yield the same results. 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 tests, which almost certainly may have had an influence on the results.

Again, however, we would need to look at their performance in subsequent years before we are able to present a more detailed argument and explanation.

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	.91	.86	.92
GLB	.94	.91	.98[1]
Standard error of measurement	5.64	5.39	6.11
Average *Rit*	.46	.37	.48
Cut-off point	68.5	68.5	67.5
Percentage failed	34.26	22.73	56.30

In the Afrikaans test, the picture is somewhat different. It can be seen that approximately 24% of the students at UP fail, while around 27% of those who took this test at US did not pass. The number of students who did not pass at NW is somewhat higher (31%), but it has to be noted that the cut-off point at US is lower than those at UP and NW (a discussion of the slight variations in cut-off points, and their justification, has been dealt with in some detail in Van der Slik & Weideman 2005). The mean test scores at US and NW are about the same size, while on average students at UP performed better than at US and NW. Clearly the variation around the mean is higher at US than at UP and NW, implying that the Afrikaans academic literacy of those at US is less homogeneous than the academic literacy of those who took the Afrikaans test at UP and NW. There is a fairly obvious explanation for this: students at US were not free to choose which language they would like to be tested in; they had to take both the TALL and the TAG, even in those cases where they were not proficient in Afrikaans. As we have 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.

1. The *GLB* is not entirely reliable in case the number of testees is lower than 200.

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	.81	.91	.83
GLB	.88	.94	.89
Standard error of measurement	5.91	6.00	6.18
Average *Rit*	.31	.43	.33
Cut-off point	60.5	50.5	55.5
Percentage failed	23.84	26.85	31.14

Misclassifications

In Table 3 and 4 we present the number of potential misclassifications based on four different 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 is applied. Remember, however, that approximately half of the misclassifications stem from testees who have passed where they could have failed. If we disregard this portion, i.e. give 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 have failed where they could have passed. In the latter case, between 216 and 128 testees who could have passed at UP may have failed. Applying the same logic to the testees at US and NW, potentially 76 to 123 testees may have undeservedly failed at US, and between 6 and 8 at NW. At NW these outcomes are somewhat unreliable, however, due to the low number of testees who took the English version there (n =135). In fact, the low number of testees at NW did not allow us to estimate the *GLB*-based number of potential misclassifications.

To give a clearer picture of these potential misclassifications, we provide additional information about the intervals around the cut-off points where these misclassifications may occur, both in terms of raw scores and in terms of standard deviations. For example, at UP the interval varies in between 3 to 6 points around the cut-off point of 68.5. In terms of standard deviations this variation is between .15 and .31 standard deviation.

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

	UP	**US**	**NW**
Alpha based:			
Correlation between test and hypothetical parallel test	432 (13.0%) *63-74 (.31)*	246 (14.2%) *63-74 (.41)*	16 (11.8%) *64-71 (.18)*
Correlation between observed and "true" scores	308 (9.3%) *65-72 (.21)*	176 (10.2%) *66-72 (.27)*	11 (8.4%) *64-71 (.15)*
GLB based:			
Correlation between test and hypothetical parallel test	360 (10.9%) *64-73 (.26)*	213 (12.3%) *66-72 (.27)*	not available
Correlation between observed and "true" scores	256 (7.7%) *66-71 (.15)*	152 (8.8%) *67-71 (.21)*	not available

The number of potential misclassifications on the Afrikaans test for those who have failed but may 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 (see Table 4).

Again, we provide additional information about the intervals around the cut-off points where these misclassifications may occur, both in terms of raw scores as well in terms of standard deviations. For example, at US the interval varies in between 3 to 5 points around the cut-off point of 50.5. In terms of standard deviations this variation is in between .15 and .26 standard deviation. In order to ensure fair treatment by the test, these measures should be used in some way to eliminate undesirable results.

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

	UP	**US**	**NW**
Alpha based:			
Correlation between test and hypothetical parallel test	415 (15.4%) *57-63 (.30)*	192 (11.3%) *46-55 (.26)*	414 (16.4%) *52-59 (.25)*
Correlation between observed and "true" scores	300 (11.1%) *58-62 (.22)*	137 (8.1%) *47-54 (.21)*	298 (11.8%) *53-58 (.20)*
GLB based:			
Correlation between test and hypothetical parallel test	349 (12.9%) *58-62 (.22)*	157 (9.2%) *47-54 (.21)*	343 (13.6%) *53-58 (.20)*
Correlation between observed and "true" scores	250 (9.3%) *59-61 (.15)*	112 (6.6%) *48-53 (.15)*	245 (9.7%) *54-57 (.13)*

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

Differential item functioning (DIF)

Using TIAPLUS, we have tested by means of DIF-statistics if students from the universities of Pretoria, Stellenbosch, and Northwest performed differently on the TALL and TAG. If the Mantel-Haenszel Statistic (Holland & Thayer 1988) is close to unity, the items are approximately equally difficult for these groups of students. If, however, this DIF-statistic is either close to zero or larger than unity, then the corresponding item performs differently for these groups of students. The associated Z-statistics show which item difficulties are considered to be statistically different ($p < .01$). We would like to emphasize, however, that due to the high numbers of testees who took the tests, even small differences in item difficulty are significant.

In analysing the TALL items, we found some interesting results. It appeared that the items of the final part of the text editing section (sub-test 6) were more difficult for UP students than they were for US testees (no differences were found with NW students, but this is not unexpected since there were only 135 NW students who took the TALL test). Additional analyses (not shown here) have underlined this conclusion. Item 53 up to item 60 remained unanswered by more than 10% of the UP students, whereas 2% or fewer of the US did not finish these last 8 items of the text editing sub-test. Does this mean that US students were more familiar with the *content* of the text editing section than UP students were, and that this might explain the occurrence of these differences? Perhaps, but we think

another explanation for the observed differences between UP and US students is more likely. Remember that US testees took the TALL only a few days after they did the Afrikaans version of the academic literacy test. 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, US students were already more acquainted with the *form* of this sub-test than 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, and the observation that US students wrote a test of roughly the same format and item types (TAG) a day or two before. Second, comparisons between students' higher levels of performance on various sub-tests when they have become increasingly familiar with the format of the 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, one should also consider whether the differences in composition of the test populations are not responsible, perhaps, for US students having less difficulty – as some other preliminary analyses that we have done with the TALL and TAG 2006 data seem to suggest – with the completion of sub-test 6.

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

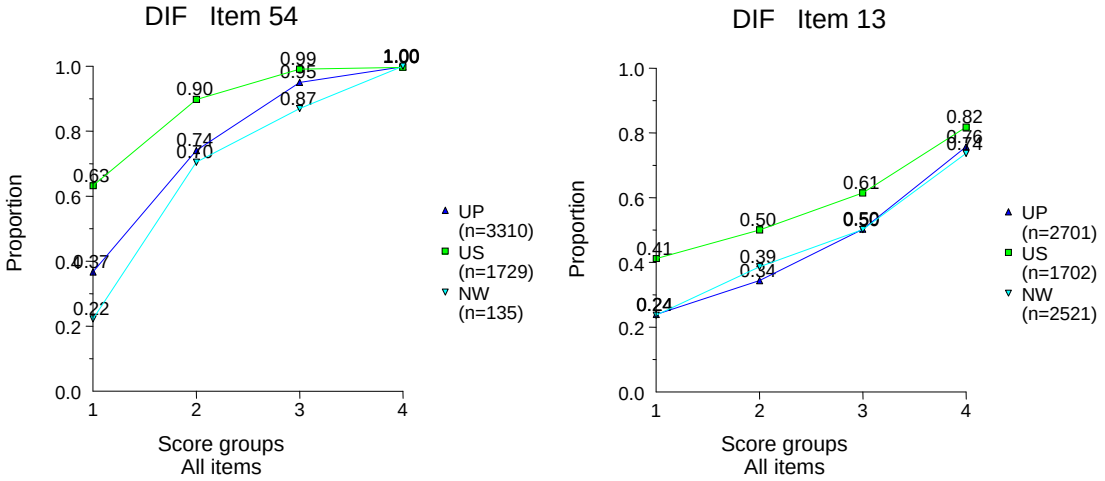

Figure 1: DIF-Graphics of TALL item 54 (left panel) and TAG item 13(right panel)

Figure 1 can be read as follows. The TIAPLUS package has divided the testees into four score groups. Score group 1 contains the 25% lowest scoring testees on all the 60 items (in the case of TALL), and on the 62 items of TAG, respectively. Score group 4 consists of those 25% who scored highest, while score group 2 and 3 fall in between. In Figure 1, left panel, it can be seen that TALL item 54 is more difficult for UP and NW students than it is for US testees. This can particularly be observed for the lower scoring groups. Where, for example, only 37% of score group 1 of UP

students had this item correct, 63% of score group 1 from US have given the correct answer on TALL item 54. The same pattern occurs in items 53 through 60.

Regarding TAG, we have found only a few indications that the constituting items perform differently for the students of UP, US, and NW. For example, we have found that TAG item 13 appears to perform differently for US students as compared to students from UP and NW (the corresponding Z-values are –3.84, and 4.06, respectively, and are highly significant: $p < .001$). In Figure 1, it can be seen that, on average, US students indeed perform better on item 13 than UP and NW testees, but this is a rare exception.

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

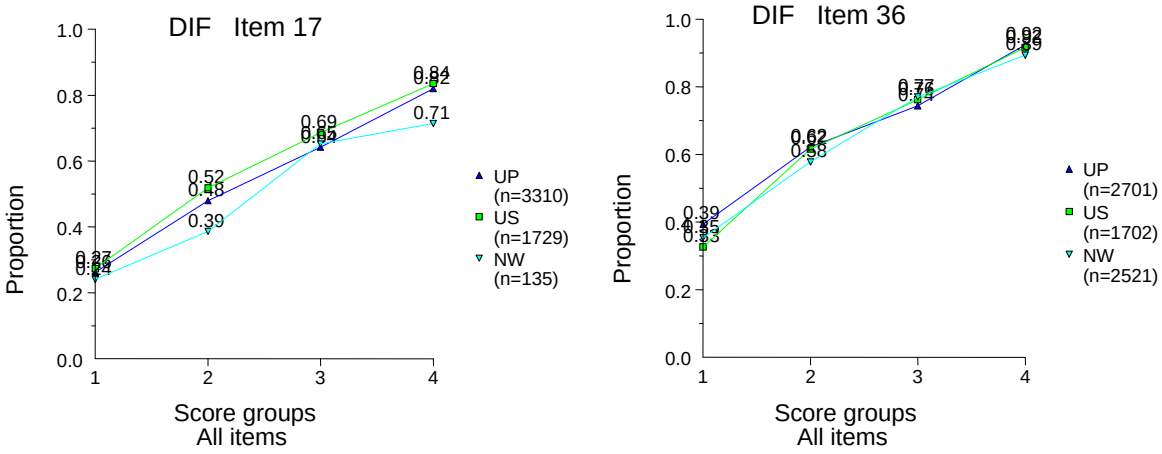

Figure 2: DIF-Graphics of TALL item 17 (left panel) and TAG item 36(right panel)

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 Table 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) [2] in order to find out whether differences between students from the three universities, though possibly highly significant, are nevertheless trivial.

2. Cohen's $d = (\mu_1 - \mu_2) / \sigma_{pooled}$, where $\sigma_{pooled} = (((n_1 - 1)\,\sigma_1^2 + (n_2 - 1)\,\sigma_2^2) / (n_1 + n_2 - 2))^{½}$

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

	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 *(.18)*	3.86 *(.36)*	5.62 *(.62)*	4.02 (1.43)	4.27 (1.21)	3.50 (1.56)
Sub-test 2	20	.04 *(.00)*	4.87 *(.43)*	4.79 *(.44)*	13.83 (3.96)	13.83 (3.83)	12.12 (4.01)
Sub-test 3	6	–5.05 *(.07)*	5.49 *(.52)*	6.90 *(.22)*	4.44 (1.41)	4.64 (4.64)	3.70 (1.56)
Sub-test 4	5	–6.16 *(.18)*	2.43 *(.22)*	4.36 *(.40)*	3.83 (1.30)	4.06 (1.28)	3.54 (1.34)
Sub-test 5	49	–9.49 *(.26)*	5.49 *(.64)*	7.86 *(.99)*	35.50 (11.20)	38.19 (8.60)	29.28 (12.96)
Sub-test 6	15	–16.86 *(.44)*	5.91 *(.57)*	10.03 *(.96)*	10.12 (4.48)	11.90 (2.94)	7.56 (4.95)
Total test	100	–10.60 *(.29)*	6.28 *(.62)*	8.94 *(1.13)*	71.75 (19.31)	76.89 (14.57)	59.70 (21.97)

Apparently, testees from UP and NW scored significantly lower on TALL than students from US did (T = –10.60, p < .0001; T = 8.94, p < .0001, respectively). We have already noted that due to the large sample sizes, even trivial differences between scores might prove to be 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 = .29, which is considered as a rather weak effect size (Cohen 1992)[3]. The effect size for US against NW, however, is 1.13, which is considered as a strong effect. Clearly, the first difference may be called trivial; the latter far from that. A review of the effects sizes presented in Table 5 reveals that the differences between scores of UP and US students are rather small, those between UP and NW students are medium, while those between US and NW testees vary between medium and strong.

3. Cohen (1992) considers .20 a weak effect, .50 a medium effect, and .80 a strong effect.

Table 6: T-Statistics (and effect sizes) for the Afrikaans version of the academic literacy 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 *(.09)*	7.44 *(.21)*	3.65 *(.11)*	3.40 (1.81)	3.23 (1.88)	3.02 (1.89)
Sub-test 2	20	6.64 *(.21)*	13.34 *(.37)*	4.39 *(.14)*	13.65 (3.81)	12.79 (4.37)	12.21 (3.95)
Sub-test 3	7	5.27 *(.11)*	10.27 *(.28)*	3.36 *(.16)*	5.59 (1.39)	5.43 (1.64)	5.17 (1.56)
Sub-test 4	5	5.76 *(.19)*	6.67 *(.19)*	–.30 *(.01)*	3.65 (1.49)	3.35 (1.75)	3.37 (1.53)
Sub-test 5	48	13.04 *(.43)*	14.53 *(.40)*	–1.87 *(.06)*	31.82 (7.88)	27.93 (10.61)	28.51 (8.55)
Sub-test 6	15	13.63 *(.46)*	14.13 *(.39)*	–2.50 *(.08)*	12.05 (2.73)	10.50 (4.16)	10.81 (3.54)
Total test	100	12.98 *(.44)*	17.79 *(.49)*	.12 *(.00)*	70.16 (13.55)	63.15 (19.50)	63.08 (15.07)

Regarding TAG, testees from UP scored significantly higher than students from US and NW did ($T = 12.98$, $p < .0001$; $T = 17.79$, $p < .0001$, respectively), while we found no evidence of any difference between US and NW students on TAG. Cohen's *d* values show, however, that the total scores of UP students are not all that much 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 (.44 and .49, respectively). The remaining effects sizes, presented in Table 6, reveal that the differences between 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. As regards the first variation, it would come as no surprise to the lay observer that the level of English academic proficiency at US is generally higher than that of its two northern counterparts, with the group that traditionally takes on larger numbers of students from non-urban backgrounds faring worst. As regards the second variation, it seems obvious that the compulsory administration of TAG to all students at US has shown this group to be less proficient in Afrikaans than, for example, the UP first years.

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 similar tests of academic literacy, such as the National Benchmark Test of academic literacy now being developed under the auspices of Higher Education South Africa (HESA).

First, the generally high reliability measures (in terms of both Cronbach's α and GLB) observed give an indication 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.

Second, 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, as we have pointed out above, 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 testees 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 a second-chance or borderline cases test. This test should be of a similar format as the current test, and afford borderline cases another opportunity of demonstrating their level of academic literacy. Because the tests are so reliable, however, it appears that the numbers of those eligible for a second opportunity are not large. In addition, the test administrators need not necessarily use the misclassifications as in Tables 3 and 4 above that are based on the more conservative reliability estimate (Cronbach's α). We have pointed out elsewhere (Van der Slik & Weideman 2005; cf also the CITO 2005 manual, p. 18, and Jackson & Agunwamba 1977) that for other than homogeneous tests, GLB is in any event the more appropriate estimate of reliability. So the numbers of testees who qualify for a second-chance test need to vary, in the case of UP, for example, between 128 and 180 for TALL 2005 (cf Table 3 above). 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 analyses 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 one year and that of the next year, we now have a set of benchmarks (between .1 and .2 standard deviation for TAG, and between .1 and .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, viz. validity and reliability, but also to achieve the social acceptance of a test. A test should, for example, 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 is achieved is by reflecting on, analysing and making public as much information about a test as possible (cf e.g. 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.

Third, the analysis of the effect-sizes of the variations in test performance 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 .29, for example; see Tables 5 and 6) or medium (between .44 and .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 to be vigilant about 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, it 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 testees' performance 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 that were reported on in this paper. These initial analyses indicate that we have a set of robust measuring instruments.

References

BYGATE M

2004. Some current trends in applied linguistics: Towards a generic view. *AILA review* 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, New Jersey: Lawrence Erlbaum Associates.

COHEN J

1992. A power primer. *Psychological Bulletin* 112: 155-159.

HOLLAND P W & D T THAYER

1988. Differential item performance and Mantel-Haenszel. In H. Wainer & H. Braun (eds.), *Test validity*. Hillsdale, NJ: Lawrence Erlbaum. P 129-145.

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: 567-578.

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. [Online]. Available (http://www.up.ac.za/academic/humanities/eng/eng/unitlangskills/eng/fac.htm); Accessed 14 April 2006.

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.

VAN DER SLIK F

2006. Language proficiency and fairness. Keynote address: SAALA 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.

VAN DER SLIK F & A WEIDEMAN

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* 38 (1): 1-13.

VAN DYK T & A WEIDEMAN

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): 152-164.

VERHELST N D

2000. *Estimating the reliability of a test from a single test administration.* 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): 1160-1176.

WEIDEMAN A

2003. *Academic literacy: Prepare to learn*. Pretoria: Van Schaik.

WEIDEMAN, A

2006. Transparency and accountability in applied linguistics. *Southern African linguistics and applied language studies* 24 (1): 71-86.