This is an accepted manuscript of an article published by Taylor & Francis in *Southern African Linguistics and Applied Language Studies*,Volume 27, 2009 - Issue 3. Available online: http://www.tandfonline.com/doi/abs/10.2989/SALALS.2009.27.3.6.940.

Revisiting test stability: further evidence relating to the measurement of difference in performance on a test of academic literacy

Frans van der Slik[1] and **Albert Weideman**[2*]

[1]*Department of Linguistics, Radboud University*
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
and
Department of English, University of the Free State
P.O. Box 339, Bloemfontein, 9300 South Africa
E-mail: f.v.d.slik@let.ru.nl

[2]*Department of English, University of the Free State*
P.O. Box 339, Bloemfontein, 9300 South Africa
**Corresponding author, E-mail: albert_weideman@vodamail.co.za*

Abstract

In several earlier analyses of two tests of academic literacy, the *Test of Academic Literacy Levels* (TALL) and its Afrikaans counterpart, the *Toets vir Akademiese Geletterdheidsvlakke* (TAG), we have adopted an approach to the problem that tests may be abused, and used to harm people, by discussing various antidotes to this, to ensure fairness and consistency in the tests we use, as well as by demonstrating how the process of test development and implementation might be made more transparent. We have pointed out that a true measure of the stability of such tests may only become apparent when we have data stretching over a number of years. This article reports on an investigation of differing measures of performance on the two tests in question across several of their administrations between 2005 and 2008. We point to a number of ways in which empirical measures may be used to achieve benign, enabling effects on those tested.

Current emphases and fairness in language testing

The current emphasis on validating tests through the production of (mounting layers of) evidence and argument seeks to achieve a theoretical justification for the interpretation that attaches to test scores (following the massively influential work of Messick, 1980; 1981; 1988; 1989 on testing in general; cf. also Van der Walt & Steyn, 2007 for one of the two particular tests in question [TAG]). In addition to the validation of tests, there is a growing concern among test designers and users about becoming ever more accountable (Shohamy, 2001a, 2001b, 2004; McNamara & Roever, 2006). Both of these emphases make test development a contested terrain. Though seldom stated as bluntly as this, the implications in the literature are that test designers are all-powerful (Shohamy, 2004), specifically regarding the social impact their test might have. Indeed, the discussion of test abuse over the ages in McNamara and Roever (2006: 149-199) contains shameful examples of how specifically language tests have been employed in ways that are inimical to the interests of those submitting to them.

We should note in passing that, contrary to such examples and implications, the real experiences and lives of test designers are probably much more humdrum. Test designers, in our experience at least, live much more precarious lives than the powerful beings they are made out to be. Like many other professionals, they fret over quality and about ensuring fairness, expecting every new analysis to yield something new to be concerned about (Van der Slik & Weideman, 2005). Even at the micro (item) level, they have to deal with items or components of items that do not function well, by either not discriminating adequately, or by discriminating negatively, or by discriminating differently across different administrations of the test. At the subtest level, they may be concerned, as the two authors of this article were, about why a largely similar subtest may in the one instance exhibit gender bias, and in another administration not (cf. Van der Slik, 2009). At the test level, their best attempts may show up to be inconsistent and unreliable measures, a problem that is compounded when no ready explanations are to be found for such inconsistencies. At the theoretical level, the validity of their testing instrument may be challenged, as may the interpretations that they, and the administrators who use these, attach to the test scores subsequent to the writing of a test. There are therefore enduring worries about how much of what they are testing might be

irrelevant to the ability that is being tested. In some democratic countries, they have to concern themselves with steering clear of litigation, which may happen despite their every precaution and care. And so the list goes on.

There are, of course, various ways of ensuring that language tests do measure fairly, and that their results are not abused. This article reports in the first instance on the consistency of two tests of academic literacy over several administrations. In doing so, it replicates, to the extent that this is feasible and possible, the analyses done in Weideman and Van der Slik (2008), thus presenting a further opportunity for critical scrutiny of the stability of the tests of academic literacy in question, and a much sought after result: to see whether earlier analyses hold up when further data, of a similar sort, become available.

In addition to these concerns, the current article also attempts so situate the various considerations that test developers and administrators have within an emerging theoretical framework for applied linguistics (cf. Weideman, 2009 in this volume). How do concerns with reliability or test consistency and the technical validation of a test relate to conventional and current ideas of fairness? How does construct validity relate to the social impact of a test (if at all)? Where do transparency and accountability figure, conceptually, in test design and administration? We shall attempt to relate these as best we can to the various technical analyses that were done of the results of several administrations of both the *Test of Academic Literacy Levels* (TALL) and its Afrikaans counterpart *Toets van Akademiese Geletterdheidsvlakke* (TAG) at three South African universities.

It should be noted, however, that such technical analyses always fit into a larger framework, into a broader perspective against which we can justify theoretically the overall design of a test. While this article of necessity deals with only a limited number of technical analyses that will allow the fairness of the two tests in question to be improved, the overall concern remains with the broader picture, with how improvements to the design of the tests can ensure equity and justice, and how the interpretations of their results should be accomplished in such a way as to ensure that any negative social or ethical impact they may have, is either dealt with appropriately, or kept to a minimum. In articulating this perspective, we take care, too, to show that our work of designing and improving tests, of implementing and administering them,

and of interpreting and using the scores generated by these measuring instruments, is limited, and certainly not all-powerful. For us, any design has to be theoretically defensible and socially accountable, and our empirical analyses allow us to identify potential problems in this regard.

Before we turn below to the specific questions that we wished our subsequent and more extensive analyses (extensive in the sense that this time around we had data spread over several years at our disposal) to answer, we summarise in the next section the analyses and findings of the previous, preliminary study.

Analyses and findings: initial study

The main question in the preliminary study (Weideman & Van der Slik, 2008) was whether the tests (TALL and TAG, described below) used by the three institutions were robust enough to possess a stability across their various administrations in these three contexts. Moreover, we wished to ascertain whether the variation (if any) that was evident could be ascribed to differences among the various populations tested (first year students), or to the technical inconsistency of the test. The egalitarian assumption was that one should not have too much variation – or at least rationally inexplicable variation – in the composition of the first year intake of three South African universities.

As subsidiary questions, we first examined how fair the test was in terms of potential misclassifications, i.e. how many students who might have made the cut-off point, did not, or, conversely: how many who made the cut-off point, should not have. Second, through Differential Item Functioning (DIF) analyses (McNamara & Roever, 2006: 83-128), we looked at whether items were functioning differently when administered to these different populations. Third, we asked whether the scores achieved by the populations as a whole differed across the various populations, for both tests.

Our conclusions were, first, that the tests were very stable across various administrations in terms of at least two reliability measures. Both as regards Cronbach's alpha and as regards Greatest Lower Bound (GLB) (Jackson & Agunwamba, 1977), the two tests scored very satisfactorily on these reliability indices. The more conservative of the two is obviously Cronbach's alpha, but also on this measure the tests were well within an acceptable range, with the indices for TALL lying between .86 and .94, and

for TAG between .81 and .94, as can be seen in Tables 1 and 2 below (UP = University of Pretoria; US = University of Stellenbosch; NWU = Northwest-University):

Table 1: Reliability indices: TALL 2005

	UP	**US**	**NWU**
Cronbach's alpha	.91	.86	.92
GLB	.94	.91	.98[1]

Table 2: Reliability indices: TAG 2005

	UP	**US**	**NWU**
Cronbach's alpha	.81	.91	.83
GLB	.88	.94	.89

As regards potential misclassifications, the variations around the cut-off points ranged between .15 standard deviations to .41 for TALL (the latter being for the University of Stellenbosch, and expectable within that population), and between .13 and .30 for TAG. These could be explained and accounted for, and those potentially harmed by such misclassification could be dealt with fairly, by affording them a second opportunity, where they were allowed to write a similar academic literacy test. Such a second opportunity test presents a good example of how the results of empirical analyses may be employed to have a benign, enabling effect. Similarly, DIF analyses showed that the test designers should not have much concern, and, in addition, the acceptable discriminative power of the tests was yet another indication that the current design was doing what it should.

The answer to the third specific question, regarding differences in scores obtained among the different populations, was that there were indeed differences, but not nearly of as great an order as T-tests would suggest. In this instance, we used Cohen's *d* (cf. Cohen, 1988; 1992: 157) in order to gauge the effect-size of the difference, and found that for all of the differences that were expressed in this way one could find some reasonable explanation. For example, the only relatively strong variation in terms of this measure was the $d = 1.13$ effect-size obtained when comparing

test scores for the administration of TALL at the University of Stellenbosch with that of North-West University; the other variations were either weak or medium.[2] The large effect-size in the case of TALL at the University of Stellenbosch and Northwest-University was explicable in terms of the composition of the two populations, and by a strong sense that the two localities (the relatively affluent Western Cape as against the relatively poor Northwest) from which the institutions drew most of their prospective students, have both demographically and historically been associated, respectively, with high and low competence in English. We noted at the same time, however, that in this case we may have stumbled upon something that might, given our original assumption that one should ideally not expect too much variation among first year students, have consequences beyond the initial purpose of the test. In other words, if it were true, as seems to be indicated by this rather large effect-size, that the assumed similarity of these populations is a fiction, then that would mean that the universities in question both needed to be mindful of their access requirements: the first perhaps taking in too few students from disadvantaged backgrounds, and the latter perhaps taking in too many.

A final implication of our findings, unrelated to the initial research questions, but of some importance to test designers' work, was that we might be able to set parameters for some of the significant and non-trivial differences among testees' performance on the test, as a kind of warning light that something needed special or additional attention. For example, even though our analyses indicated some greater variation in this respect, we proposed that potential misclassifications generated by the test should ideally be between .1 and .2 SD for TAG, and between .1 and .3 SD for TALL. Should they lie outside of these parameters, this would cause the test designers concern.

In the latter example, we once again have an indication of how empirical analyses allow the test designer not to become all powerful and arrogant, but exactly the opposite: the results of the analyses work rather in the opposite direction, making the test developer more mindful, not less, of the meticulous attention that tests and their results, once employed, require, and of the increasing sophistication that empirical analyses bring to bear on test requirements.

Research questions: follow-up study

Our research questions for this, the subsequent study, follow as closely as possible those of the initial analysis. We wished to ascertain:

- How stable are TALL and TAG across different administrations in terms of reliability, difficulty and discriminative power over four years (2005-2008)?
- How accurate are the cut-off points, and within which parameters do the potential misclassifications occur?
- Are there large variations, in terms of effect-size, among the scores of the different populations in each of the separate administrations?

If any of these had negative or undesirable answers, it would seriously impact for us, as test designers, on the credibility that we might attach to our tests, and on the level of confidence with which we would be able to promote their use. Returning once more to a sub-theme of this paper, these are technical limitations that test developers place upon their power. No doubt, there will be cases where dishonesty will reign, but this dishonesty is not caused by empirical analyses; rather, it might be prompted by either the lack or the unfavourable outcome of such.

Method

Population

Between 2005 and 2008, 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 have to take both tests. They normally submit to the Afrikaans test first, and one or more days later sit for the English test. Over this period 34 604 new students participated in the Afrikaans test (10 163 at UP; 13 852 at US; 10 589 at NW), while 29 763 first year students or new entrants took the English version (15 202 at UP; 13 886 at US; 675 at NW).

The tests: TALL and TAG

The versions of the *Test of Academic Literacy Levels* (TALL) and the *Toets van Akademiese Geletterdheidsvlakke* (TAG) that are being referred to here consist of between 62 and 66 items respectively, distributed over six subtests or sections (described in Van Dyk & Weideman, 2004a), all of which are in multiple-choice format (except for the 2005 test, that still had a seventh section on academic writing). The description below gives both the section (subtest) and, in brackets, the average number of items and marks normally allocated to each:

- 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 (7 items, 7 marks)
- Section 4: Text type (5 items, 5 marks)
- Section 5: Understanding texts (20 items, 47 marks)
- Section 6: Text editing (16 items, 16 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

For our analysis of the test results, we made use of two statistical packages: Statistical Package for the Social Sciences (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 shall present below 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) for both TAG and TALL. At the test level we make use of the reliability statistics Cronbach's α and *GLB* or Greatest Lower Bound reliability (the latter for the sake of brevity not reported here).

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). Again, for the sake of economy, we report only on one hypothetical set of misclassifications.

Another question, whether students from UP, US, and NW performed differently on the TALL and TAG items, is examined in more detail in the earlier paper (Weideman & Van der Slik, 2008). DIF-statistics like the Mantel-Haenszel test and Z-test are used in that analysis to determine whether individual items display a difference in sub-group (UP, US, NW) performance. In the current volume, Van der Slik (2009) examines whether the tests exhibit gender bias.

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

Descriptive statistics

Table 3 depicts the outcomes at the scale level for TALL and TAG. It is clear that, over the years, both TALL and TAG have been highly reliable in terms of Cronbach's alpha (for TALL averaging .90, for TAG .85). In addition, the average *Rit*-values, indicative of the discriminative power of the items, appear to be sufficiently high as well (TALL: .43; TAG: .36).

Table 3: Selected properties of the academic literacy test (2005-2008) (standard deviations in italics)

TALL	**UP**	**US**	**NWU**	**Overall**
N	15,202	13,886	675	29,793
Mean proportion correct (difficulty)	.65 *(.05)*	.69 *(.05)*	.49 *(.13)*	.61 *(.12)*
Mean Cronbach's alpha (reliability)	.92 *(.01)*	.88 *(.01)*	.91 *(.03)*	.90 *(.02)*
Mean Average *Rit* (discrimination index)	.45 *(.01*	.38 *(.01)*	.45 *(.02)*	.43 *(.04)*
TAG				

N	10,163	13,852	10,589	34,604
Mean proportion correct (difficulty)	.61 *(.06)*	.55 *(.05)*	.54 *(.06)*	.57 *(.06)*
Mean Cronbach's alpha (reliability)	.82 *(.01)*	.89 *(.01)*	.83 *(.01)*	.85 *(.04)*
Mean Average *Rit* (discrimination index)	.33 *(.01)*	.41 *(.01)*	.33 *(.01)*	.36 *(.04)*

Similarly, the tests appear to be neither too difficult nor to easy, as can be seen from the average proportion correct answers, though on TALL it is US students who score highest, and NWU students lowest, while on TAG UP students score highest, and NWU students lowest.

These outcomes are entirely consistent with the findings of the previous study (Weideman & Van der Slik, 2008).

Misclassifications

In Tables 4 and 5 we present the number of potential misclassifications based on two different criteria: reliability in terms of Cronbach's alpha, and correlation between test and hypothetical parallel test.

Table 4: 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

TALL	**UP**	**US**	**NWU**
Alpha based: Correlation between test and hypothetical parallel test			
2005	432 (13.0%)	246 (14.2%)	16 (11.8%)
	63 – 74 *(.31)*	63 – 74 *(.41)*	64 – 71 *(.18)*
2006	439 (12.0%)	432 (11.7%)	20 (13.7%)
	51 – 59 *(.25)*	52 – 58 *(.25)*	45 – 54 *(.26)*
2007	448 (11.5%)	604 (14.5%)	18 (12.8%)
	47 – 55 *(.19)*	54 – 61 *(.24)*	43 – 52 *(.19)*

2008	179 (4.1%)	152 (3.6%)	26 (10.0%)
	30 – 35 *(.15)*	34 – 42 *(.24)*	37 – 43 *(.15)*
Average % *(Average sd)*	(10.0%)	(11.0%)	(12.0%)
	(.23)	*(.28)*	*(.20)*

Table 5: Potential misclassifications on the Afrikaans 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

TAG	**UP**	**US**	**NWU**
Alpha based: Correlation between test and hypothetical parallel test			
2005	415 (15.4%)	192 (11.3%)	414 (16.4%)
	57 – 63 (*.30)*	46 – 55 *(.26)*	52 - 59 *(.25)*
2006	398 (15.6%)	590 (15.9%)	490 (18.5%)
	47 – 53 *(.27)*	52 – 59 *(.22*	46 – 53 *(.26)*
2007	372 (14.4%)	560 (13.5%)	464 (17.1%)
	42 – 48 *(.26)*	39 – 46 *(.21)*	41 – 47 *(.19)*
2008	117 (5.0%)	586 (13.7%)	429 (15.8%)
	29 – 35 *(.28)*	39 – 46 *(.21)*	37 – 42 *(.20)*
Average % *(Average sd)*	(12.6%)	(13.6%)	(17.0%)
	(.28)	*(.23)*	*(.23)*

Potential misclassifications occur because tests are never entirely reliable measuring instruments. As we have pointed our before, gauging such unreliability not only gives us a means of identifying who potentially may have been wrongly identified by the test as having risk in terms of academic literacy level, and making a second chance test available to these, but it also allows us to set parameters for the size of the group so potentially misclassified. The provisional parameters were, as was pointed out above,

between .1 to .3 standard deviations around the cut-off point for TALL, and between .1 and .2 standard deviations around the cut-off point for TAG.

We note that, for TALL, the average standard deviations in each case were within the proposed limits. What is also heartening is that in the case of the US students, where the standard deviation was .41 for the first test in 2005, the deviation has come within acceptable limits. For TAG, however, the UP misclassification has a variation lying outside the (initially stricter) limits of between .1 and .2. Were these limits too strict? Should the test designers be concerned that, on average over four years, there is too large a variation here? Perhaps, but there is also the possibility that the initial limits, set after the analysis of the results of a single test (TAG 2005), may simply have been too strict. Nonetheless, it is a potential concern, and, as test developers, we would recommend at least flagging it for future attention.

Overall, however, misclassifications occur more or less within the expected range of scoring points around the cut-off point, i.e. around .25 SD around the cut-off point.

Worst and best performing subtests

Throughout the years, two subtests (cf. Van Dyk & Weideman, 2004b) within TALL and TAG have stood out as, respectively, the worst and the best performing ones: the subtest Interpreting graphic and visual information (GVI) and Text editing (TE). Tables 6 (for TALL) and 7 (for TAG) below give an overview of how they fared, on average, over the years, within the various administrative contexts.

Table 6: Selected properties of the relatively worst (GVI) and best performing subtests of TALL (2005-2008) (standard deviations in italics)

Interpreting graphic & visual literacy	**UP**	**US**	**NWU**	**Overall**
Average number of items				6.5
Mean proportion correct	.74 *(.06)*	.79 *(.05)*	.63 *(.07)*	.71 *(.09)*
Mean Cronbach's alpha	.63 *(.09)*	.57 *(.14)*	.63 *(.08)*	.61 *(.10)*

Mean average *Rit*	.58 *(.03)*	.56 *(.02)*	.59 *(.03)*	.57 *(.03)*
Text editing				
Average number of items				16.5
Mean proportion correct	.61 *(.13)*	.64 *(.05)*	.47 *(.04)*	.57 *(.11)*
Mean Cronbach's alpha	.90 *(.06)*	.89 *(.01)*	.92 *(.01)*	.90 *(.04)*
Mean average *Rit*	.65 *(.07)*	.63 *(.04)*	.69 *(.01)*	.66 *(.05)*

Table 7: Selected properties of the relatively worst (GVI) and best performing subtests of TAG (2005-2008) (standard deviations in italics)

Interpreting graphic & visual literacy	**UP**	**US**	**NWU**	**Overall**
Average number of items				6.2
Mean proportion correct	.71 *(.08)*	.66 *(.07)*	.65 *(.08)*	.67 *(.08)*
Mean Cronbach's alpha	.53 *(.07)*	.62 *(.08)*	.53 *(.08)*	.56 *(.08)*
Mean average *Rit*	.54 *(.03)*	.58 *(.01)*	.54 *(.03)*	.55 *(.03)*
Text editing				
Average number of items				14.7
Mean proportion correct	.72 *(.09)*	.65 *(.07)*	.62 *(.09)*	.67 *(.07)*
Mean Cronbach's alpha	.87 *(.06)*	.88 *(.02)*	.90 *(.04)*	.88 *(.04)*
Mean average *Rit*	.63 *(.08)*	.63 *(.04)*	.67 *(.09)*	.65 *(.07)*

It is clear from the above that the subtest Text editing has superior reliability both for TALL and TAG, and that the subtest Interpreting graphic and visual literacy scores much less on the same measure.

Should this be of concern to designers and developers of the two tests? In the first instance, both appear to measure necessary components of the construct (Weideman, 2007; Van Dyk & Weideman, 2004a), and for that reason alone even the worst performing subtest cannot be excluded without diluting the rich construct of

academic literacy that is being tapped. That this same subtest exhibits some gender bias in favour of men in the one test, TAG, may be more reason for concern (Van der Slik, 2009). Yet the different scores on the same reliability index can almost wholly be ascribed to subtest length. In other words, since the Text editing subtest is more than twice the length of its lower performing counterpart, it is its length, the fact that its measurement is achieved over many more items, that gives it the edge. If one hypothetically had a subtest for Graphic and visual literacy that was also about 16 items long, its expected alpha would have risen from its current .61 (for TALL) and .56 (for TAG) to, respectively, .77 and .81.

T-Tests and effect sizes

Finally, we 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 8 and 9 we present the outcomes of T-tests for the entire tests. In addition, we present Cohen's *d* (Cohen, 1992: 157)[1] in order to find out whether differences between students from the three universities, though possibly highly significant, are nevertheless trivial.

Table 8: T-Statistics (and effect sizes) for TALL 2005-2008

	Max. score	**UP vs. US *(d)***	**UP vs. NWU *(d)***	**US vs. NWU *(d)***	**UP Mean (SD)**	**US Mean (SD)**	**NWU Mean (SD)**
2005	100	–10.60	6.28	8.94	71.75	76.89	59.70
		(–.29)	*(.62)*	*(1.13)*	(19.31)	(14.57)	(21.97)
2006	100	– 9.81	4.88	7.49	64.32	68.46	56.27
		(–.23)	*(.41)*	*(.68)*	(20.02)	(16.54)	(19.18)
2007	100	– 9.21	5.57	7.90	61.11	64.98	50.44
		(–.21)	*(.55)*	*(.75)*	(20.59)	(16.79)	(21.57)
2008	100	*Not available*	6.28	*Not available*	62.59	*Not available*	54.34
			(.41)		(20.15)		(20.30)

Table 9: T-Statistics (and effect sizes) for TAG 2005-2008

	Max. score	**UP vs. US *(d)***	**UP vs. NWU *(d)***	**US vs. NWU *(d)***	**UP Mean (SD)**	**US Mean (SD)**	**NWU Mean (SD)**
2005	100	12.98	17.79	.12	70.16	63.15	63.08
		(.44)	*(.49)*	*(.00)*	(13.55)	(19.50)	(15.07)
2006	100	15.98	14.74	–1.25	60.18	53.53	54.07
		(.39)	*(.41)*	*(–.03)*	(15.02)	(18.04)	(15.22)
2007	100	11.60	12.94	1.67	56.66	51.78	51.14
		(.28)	*(.36)*	*(.04)*	(15.41)	(18.80)	(15.56)
2008	100	*Not available*	17.49 *(.49)*	*Not available*	55.83 (14.23)	*Not available*	48.62 (14.96)

Gratifyingly, the large difference (d = 1.13) calculated when comparing the US and NWU results for TALL in 2005 has in the subsequent years eased in the right direction, away from a maintenance of that undoubtedly strong initial effect. In fact, except for a few slight rises, for example as in the US and NWU comparison on TAG between 2007 and 2008, the effect sizes generally seem to be remaining in the same ballpark, or even getting closer, which is heartening in another respect: it means that the various new intakes of first year students are at the very least remaining more or less stable in academic literacy ability over the years, a conclusion corroborated by the findings of Van der Slik and Weideman (2007).

Overall, therefore, we find differences among the various populations at the three institutions, but not as massively as suggested by the T-statistics. In the case of TALL, the differences between the UP and the US students' scores remain small, while the differences among NWU on the one hand, and UP and US on the other, are substantial. As regards TAG, differences between NWU and US are negligible, but rather large between UP on the one hand, and US and NWU on the other. Since the administration of the test is, by all accounts, subject to the same set of standardised administrative procedures for test implementation, the differences that we have again noticed here can be explained, no doubt, by differences in the composition of the various student bodies.

Conclusion

Our analysis has to conclude, therefore, that the robustness and stability noticeable in the initial study was no flash in the pan. The current outcomes are entirely consistent with those of the earlier study (Weideman & Van der Slik, 2008), and the analyses from which they derive shall probably be used, henceforth, only internally, for the improvement of the test designs, and as signals of which empirical limits are tolerable for such large scale test administration.

The two emphases in language testing that were referred to at the beginning of this article – the validation and accountability that are sought for such tests – are in our opinion related, respectively, as constitutive and regulative conditions for language testing (Weideman, 2009). Language testing belongs centrally to applied linguistics (McNamara & Roever, 2006: 255; McNamara, 2003), and tests are for that reason applied linguistic artefacts (Weideman, 2006). If, as Weideman (2008; 2009) argues in his articulation of a systematic, foundational framework for this field, applied linguistics itself is a discipline characterised by design, then test designers and developers are, in terms of that theoretical framework, doubly accountable for their tests. Their accountability is in the first instance a theoretical one, since, if tests are qualified by their technical function of design and founded upon their theoretical justification, as this framework suggests, then that justification presents us with a first articulation of what needs to be accounted for.

We may call the giving such a theoretical account the theoretical defensibility or justification of the artefact, which is in this case a set of two language tests. Conventionally, such justification, since it relates to a hypothetical ability of what the tests are measuring, is called the construct validity of the instrument. In the present case, this enriched notion of the technical power of the test is (partly) validated by a unity within a multiplicity of different sources of test data, as well as by the stability of these tests. Several such sets of data for test stability have been surveyed above. The notion of technical stability or consistency which is the main focus of the current study relates, within the framework proposed by Weideman (2008; 2009), to the leading technical function of the design connecting analogically with the kinematic dimension of experience, of which regular, consistent movement is the nuclear meaning. At the same time, the validation of the test, its technical power to test what it sets out to do –

measure academic literacy – relates to the coherence of the technical aspect, that leads the design, and the sphere of energy-effect, the physical dimension of our experience. And in augmenting that notion of technical power by connecting it also to the theoretical rationale for a design, we have enriched the concept of technical effect to a consideration also of the analytical defensibility of the test construct. For a more complete explanation, we refer to the exposition in Weideman (2008; 2009).

There is also a second kind of accountability for the test design, which contains echoes of a social and political nature (Weideman, 2006). This kind of accountability relates to the technical sphere that qualifies the test design reaching out, as it were, to the social and juridical dimensions of reality. Such technical accountability is nowadays referred to as the consideration of test consequences or test impact (Davies & Elder, 2005; Hamp-Lyons, 2001). The focus of this article has mainly been on test stability, but, since such technical stability is the basis for a consideration of accountability, we again have to note this in passing.

It needs to be said in this regard that this second kind of accountability is not the same as the first, its theoretical defensibility. Although this is not often enough stated thus in the literature on language testing, in order to achieve accountability, one needs much more than a set of professionally agreed upon standards (cf. e.g. AERA, 1999), which in the main give only a (mostly implicit) theoretical defence of the design, or of what might still become the design. Since such agreements for the most part refer to design principles for those wishing to construct theoretically sound tests, the professional formulation of and agreement about standards for language testing that we find in them are not realistically within reach of, and accessible to, either the public at large, or those closely affected by a test. Therefore, if test developers are to become socially and politically accountable for their designs, such designs of necessity have to be preceded by transparency, by having as much public information as possible freely available. For that, academic standard setting is not enough. Bygate's (2004) observation that applied linguists, which would in our definition include language test designers, have a dual accountability, viz. an academic, technical accountability, as well as a public accountability, is relevant here. All of these concerns, however, need further exploration, possibly in a study that deals more specifically with them.

As regards the need for transparency in the current case, the Unit for Academic Literacy makes information available on TALL and TAG to all prospective students at the University of Pretoria by first distributing some 17 000 brochures on the tests to them, and, second, by publishing this information, together with a sample test, on its website (Unit for Academic Literacy, 2008). The information given is about what is being tested – various components of academic literacy - and it is presented as far as possible in layman's terms. It is only on the basis of sufficient information (transparency) that the political goal of accountability can be achieved, otherwise a critique of a test may be motivated by nothing more than resentment, on the part of those who feel discriminated against by the test, or informed by fiction, or even myth. It is remarkable to see how quickly, when not enough information is available about a test, such myths can form. Before we dramatically sharpened our communication on the purposes of the tests at the University of Pretoria, for example, a story was doing the rounds among students who had to write it that it was better to write the English test, since, if you chose the Afrikaans one, you would be certain to fail. Of course, both in percentage and in absolute terms this was purely wrong, but, despite that, firmly believed.

We are of the opinion that a first level of accountability has to remain a theoretical one, such as the current report strives to achieve. But the challenge remains to convert these theoretical understandings into intelligible information for non-academics, and especially for those closely affected by these tests. Nonetheless, our theoretical analyses, far from giving us as test designers some kind of 'power', function rather as important limiting conditions for our test design, correcting it and potentially containing its excesses.

Notes

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

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

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

References

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. 1999. *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

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.

Cohen J. 1988. *Statistical power analysis for the behavioral sciences*. Second edition. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cohen J. 1992. A power primer. *Psychological Bulletin* **112**: 155–159.

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

Elder C, Brown A, Grove E, Hill K, Iwashita N, Lumley T, McNamara T & O'Loughlin K (eds). 2001. *Experimenting with uncertainty: Essays in honour of Alan Davies*. Cambridge: Cambridge University Press.

Hamp-Lyons L. 2001. Ethics, fairness(es) and developments in language testing. In Elder C, Brown A, Grove E, Hill K, Iwashita N, Lumley T, McNamara T & O'Loughlin K (eds). *Experimenting with uncertainty: Essays in honour of Alan Davies*. Cambridge: Cambridge University Press, pp 222–227.

Hinkel E (ed.). *Handbook of research in second language teaching and learning*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Jackson PW & Agunwamba CC. 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.

Linn R (ed.). *Educational measurement*. Third edition. New York: American Council on Education/Collier Macmillan.

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

McNamara T & Roever C. 2006. *Language testing: The social dimension*. Oxford: Blackwell.

Messick S. 1980. Test validity and the ethics of assessment. *American psychologist* **35**(11): 1012–1027.

Messick S. 1981. Evidence and ethics in the evaluation of tests. *Educational researcher* **10**(9): 9–20.

Messick S. 1988. The once and future issues of validity: Assessing the meaning and consequences of measurement. In Wainer H & Braun IH (eds). *Test validity.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, pp 33–45.

Messick S. 1989. Validity. In Linn RL (ed). *Educational measurement.* Third Edition. New York: American Council on Education/Collier Macmillan, pp 13–103.

Norton B & Toohey K. (eds). *Critical pedagogies and language learning*. Cambridge: Cambridge University Press.

Shohamy E. 2001a. Fairness in language testing. In Elder C, Brown A, Grove E, Hill K, Iwashita N, Lumley T, McNamara T & O'Loughlin K (eds). *Experimenting with uncertainty: Essays in honour of Alan Davies*. Cambridge: Cambridge University Press, pp 15–19.

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

Shohamy E. 2004. Assessment in multicultural societies: Applying democratic principles and practices to language testing. In Norton B & Toohey K (eds). *Critical pedagogies and language learning*. Cambridge: Cambridge University Press, pp 72–92.

Unit for Academic Literacy. 2008. Sample test. [Online]. Available http://web.up.ac.za/default.asp?ipkCategoryID=2388&subid=2388&ipklookid=9. Accessed 6 September 2008.

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

Van der Slik F & Weideman A. 2005. The refinement of a test of academic literacy. *Per linguam* **21**(1): 23–35.

Van der Slik F & Weideman A. 2007. Testing academic literacy over time: Is the academic literacy of first year students deteriorating? *Ensovoort* **11**(2): 126–137.

Van der Slik F & Weideman A. 2008. Measures of improvement in academic literacy. *Southern African linguistics and applied language studies* **26**(3): 363–378.

Van der Walt JL & Steyn HS jnr. 2007. Pragmatic validation of a test of academic literacy at tertiary level. *Ensovoort* **11**(2): 138–153.

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.

Wainer H & Braun IH (eds). *Test validity*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

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

Weideman A. 2007. *Academic literacy: Prepare to learn*. Second Edition. Pretoria: Van Schaik.

Weideman A. 2008. Towards a responsible agenda for applied linguistics: Confessions of a philosopher. *Per linguam* **23**(2): 29–53.

Weideman A. 2009. Constitutive and regulative conditions for the assessment of academic literacy. Forthcoming in *Southern African linguistics and applied language studies* **27**(3).

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.

Note: See tables on following pages

Tables 1-9

Table 1: Reliability indices: TALL 2005

	UP	US	NW
Cronbach's alpha	.91	.86	.92
GLB	.94	.91	.98[1]

Table 2: Reliability indices: TAG 2005

	UP	**US**	**NW**
Cronbach's alpha	.81	.91	.83
GLB	.88	.94	.89

Table 3: Selected properties of the academic literacy test (2005-2008) (standard deviations in italics)

TALL	**UP**	**US**	**NW**	**Overall**
N	15,202	13,886	675	29,793
Mean proportion correct (difficulty)	.65 *(.05)*	.69 *(.05)*	.49 *(.13)*	.61 *(.12)*
Mean Cronbach's alpha (reliability)	.92 *(.01)*	.88 *(.01)*	.91 *(.03)*	.90 *(.02)*
Mean Average *Rit* (discrimination index)	.45 *(.01*	.38 *(.01)*	.45 *(.02)*	.43 *(.04)*
TAG				
N	10,163	13,852	10,589	34,604
Mean proportion correct (difficulty)	.61 *(.06)*	.55 *(.05)*	.54 *(.06)*	.57 *(.06)*
Mean Cronbach's alpha (reliability)	.82 *(.01)*	.89 *(.01)*	.83 *(.01)*	.85 *(.04)*
Mean Average *Rit* (discrimination index)	.33 *(.01)*	.41 *(.01)*	.33 *(.01)*	.36 *(.04)*

Table 4: 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

TALL	**UP**	**US**	**NW**
Alpha based: Correlation between test and hypothetical parallel test			
2005	432 (13.0%)	246 (14.2%)	16 (11.8%)
	63 – 74 *(.31)*	63 – 74 *(.41)*	64 – 71 *(.18)*
2006	439 (12.0%)	432 (11.7%)	20 (13.7%)
	51 – 59 *(.25)*	52 – 58 *(.25)*	45 – 54 *(.26)*
2007	448 (11.5%)	604 (14.5%)	18 (12.8%)
	47 – 55 *(.19)*	54 – 61 *(.24)*	43 – 52 *(.19)*
2008	179 (4.1%)	152 (3.6%)	26 (10.0%)
	30 – 35 *(.15)*	34 – 42 *(.24)*	37 – 43 *(.15)*
Average % *(Average sd)*	(10.0%) *(.23)*	(11.0%) *(.28)*	(12.0%) *(.20)*

Table 5: Potential misclassifications on the Afrikaans 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

TAG	**UP**	**US**	**NW**
Alpha based: Correlation between test and hypothetical parallel test			
2005	415 (15.4%)	192 (11.3%)	414 (16.4%)
	57 – 63 (*.30)*	46 – 55 *(.26)*	52 - 59 *(.25)*
2006	398 (15.6%)	590 (15.9%)	490 (18.5%)
	47 – 53 *(.27)*	52 – 59 *(.22*	46 – 53 *(.26)*
2007	372 (14.4%)	560 (13.5%)	464 (17.1%)
	42 – 48 *(.26)*	39 – 46 *(.21)*	41 – 47 *(.19)*
2008	117 (5.0%)	586 (13.7%)	429 (15.8%)
	29 – 35 *(.28)*	39 – 46 *(.21)*	37 – 42 *(.20)*
Average %	(12.6%)	(13.6%)	(17.0%)
(Average sd)	*(.28)*	*(.23)*	*(.23)*

Table 6: Selected properties of the relatively worst (GVI) and best performing subtests of TALL (2005-2008) (standard deviations in italics)

Interpreting graphic & visual literacy	**UP**	**US**	**NW**	**Overall**
Average number of items				6.5
Mean proportion correct	.74 *(.06)*	.79 *(.05)*	.63 *(.07)*	.71 *(.09)*
Mean Cronbach's alpha	.63 *(.09)*	.57 *(.14)*	.63 *(.08)*	.61 *(.10)*
Mean average *Rit*	.58 *(.03)*	.56 *(.02)*	.59 *(.03)*	.57 *(.03)*
Text editing				
Average number of items				16.5
Mean proportion correct	.61 *(.13)*	.64 *(.05)*	.47 *(.04)*	.57 *(.11)*
Mean Cronbach's alpha	.90 *(.06)*	.89 *(.01)*	.92 *(.01)*	.90 *(.04)*
Mean average *Rit*	.65 *(.07)*	.63 *(.04)*	.69 *(.01)*	.66 *(.05)*

Table 7: Selected properties of the relatively worst (GVI) and best performing subtests of TAG (2005-2008) (standard deviations in italics)

Interpreting graphic & visual literacy	**UP**	**US**	**NW**	**Overall**
Average number of items				6.2
Mean proportion correct	.71 *(.08)*	.66 *(.07)*	.65 *(.08)*	.67 *(.08)*
Mean Cronbach's alpha	.53 *(.07)*	.62 *(.08)*	.53 *(.08)*	.56 *(.08)*
Mean average *Rit*	.54 *(.03)*	.58 *(.01)*	.54 *(.03)*	.55 *(.03)*
Text editing				
Average number of items				14.7
Mean proportion correct	.72 *(.09)*	.65 *(.07)*	.62 *(.09)*	.67 *(.07)*
Mean Cronbach's alpha	.87 *(.06)*	.88 *(.02)*	.90 *(.04)*	.88 *(.04)*
Mean average *Rit*	.63 *(.08)*	.63 *(.04)*	.67 *(.09)*	.65 *(.07)*

Table 8: T-Statistics (and effect sizes) for TALL 2005-2008

	Max. score	**UP vs. US *(d)***	**UP vs. NW *(d)***	**US vs. NW *(d)***	**UP Mean (SD)**	**US Mean (SD)**	**NW Mean (SD)**
2005	100	–10.60	6.28	8.94	71.75	76.89	59.70
		(–.29)	*(.62)*	*(1.13)*	(19.31)	(14.57)	(21.97)
2006	100	– 9.81	4.88	7.49	64.32	68.46	56.27
		(–.23)	*(.41)*	*(.68)*	(20.02)	(16.54)	(19.18)
2007	100	– 9.21	5.57	7.90	61.11	64.98	50.44
		(–.21)	*(.55)*	*(.75)*	(20.59)	(16.79)	(21.57)
2008	100	*Not available*	6.28 *(.41)*	*Not available*	62.59 (20.15)	*Not available*	54.34 (20.30)

Table 9: T-Statistics (and effect sizes) for TAG 2005-2008

	Max. score	**UP vs. US *(d)***	**UP vs. NW *(d)***	**US vs. NW *(d)***	**UP Mean (SD)**	**US Mean (SD)**	**NW Mean (SD)**
2005	100	12.98	17.79	.12	70.16	63.15	63.08
		(.44)	*(.49)*	*(.00)*	(13.55)	(19.50)	(15.07)
2006	100	15.98	14.74	–1.25	60.18	53.53	54.07
		(.39)	*(.41)*	*(–.03)*	(15.02)	(18.04)	(15.22)
2007	100	11.60	12.94	1.67	56.66	51.78	51.14
		(.28)	*(.36)*	*(.04)*	(15.41)	(18.80)	(15.56)
2008	100	*Not available*	17.49	*Not available*	55.83	*Not available*	48.62
			(.49)		(14.23)		(14.96)