Ensovoort, 2007, 11(2): 126-137

Testing academic literacy over time:
Is the academic literacy of first year students deteriorating?

Frans W.P. van der Slik, Radboud University of Nijmegen & Research associate, Unit for Academic Literacy, University of Pretoria
Albert Weideman, Unit for Academic Literacy, University of Pretoria

Abstract

How much empirical evidence is there for the frequently expressed opinion that the academic literacy levels of first year students at South African universities are steadily deteriorating? Two tests of academic literacy used widely in South Africa, the Test of academic literacy levels (TALL) and its Afrikaans counterpart (TAG) may hold at least a partial answer to this question. We subject the administration, over the years 2005-2007, of one of these tests, the Toets van akademiese geletterdheidsvlakke (TAG) to an IRT analysis, using a One-Parameter Logistic Model (OPLM) package. The results show that, if we equalise the subsequent tests in terms of the first administration, there is evidence that is contrary to the popular opinion. More importantly, however, using an OPLM analysis enables us to make more responsible decisions derived from test results, and so make our tests not only theoretically more defensible, but also more accountable to a larger public.

Introduction

The debate about academic literacy in South Africa is situated at the interface between school and university education, or, in official terms, general and further education on the one hand, and higher education on the other. More often than not, questions are raised in terms of the readiness of students about to enter institutions of higher education, and specifically about their preparedness in terms of their ability to understand and use academic language within this new environment. It is therefore not surprising that there is already substantial experience in South Africa on the design and use of tests of academic literacy both for access and placement purposes (cf. Cliff, Yeld & Hanslo 2003; Cliff & Yeld 2006; Visser & Hanslo 2005; Weideman, 2003, Van der Slik & Weideman 2005, 2007; Van Dyk & Weideman 2004a, 2004b). In the discussion that follows, the measurement of academic literacy levels is understood to refer to the assessment of the ability by students to use language at the appropriate and desired level within the academic community, or their level of competence in academic discourse and its conventions, as this is defined in the work referred to here (especially Cliff & Yeld 2006 and Van Dyk & Weideman 2004a).

The interest in academic literacy levels is not confined to scholarly attention and investigation. It engages both experts and lay people in equal measure. A popularly expressed opinion would have it, for example, that the language abilities of our students are steadily decreasing over time. In South Africa, such attitudes are fed by occasional fairly sensationalist press reports of lower (and by implication lowering) literacy levels among pre-university learners (cf. for example Rademeyer 2007). Without much ado, 'low' scores are interpreted as decreasing ability. The question is almost never asked whether the scores have not perhaps been as low as this for some time. Phrased differently: a chronic problem is not the same as standards that are lowering. Furthermore, it often escapes the readers of these reports that some of them have as their origin the testing of academic and other forms of literacy by those producing commercially designed tests. Readers are not told, in other words, that those with whom the 'evidence' originates may have a financial interest in the results of these kinds of report. Dwindling language ability among the younger generation is an opinion akin to a number of those that Widdowson (2005: 15f.) discusses as 'folk linguistics'. In the perception of those involved, such strongly held opinions find more than adequate evidence in their everyday experience. It is of course so that theoretical analysis ignores naïve experience at its peril. Yet in the present case one would do well to ask: is it indeed a matter of something experienced intuitively as almost self-evident, or could these merely be deeply held prejudices and biases that are masquerading as observations that are backed up by sufficient evidence?

The current paper examines the question of decreasing levels of academic literacy obliquely, with reference to a number of tests conducted over time at North-West University (NW) and the Universities of Pretoria (UP) and Stellenbosch (US). It belongs to a series of investigations that we have done to determine the stability and robustness of the tests both across various administrations within the participating institutions (Weideman & Van der Slik 2007) and over time. The tests in question are the *Test of academic literacy levels* (TALL) and its Afrikaans counterpart, the *Toets van akademiese geletterdheidsvlakke* (TAG). The purpose of these investigations is to ensure in the first instance a measure of theoretical defensibility by telling, as Shohamy (2001) exhorts us to do, "the story of a test". This is a first step towards the eventual public accountability that a test must also achieve. In another paper (Weideman & Van der Slik 2007; cf. too Van der Slik and Weideman 2005), for example, we have already checked if the tests produce reliable outcomes when they are administered to different populations of newly arrived students. We plan to extend that investigation by performing a number of longitudinal analyses that will inform us about the ability of the tests to predict risk brought about by lower than adequate levels of academic literacy when a student enrols for study in higher education.

These tests of academic literacy have now been used at the three different universities mentioned above since 2005. Recently, the test has also been administered to new students of the Medical Faculty of the University of Limpopo. Since the outcomes of the tests for the years 2006 and 2007 have now also become available, we are currently in a position to give more serious consideration to the question raised in the subtitle of this article.

Though this is easier said than done, one way of testing if secondary schooling nowadays turns out students whose ability is growing worse as compared to students from previous years is to compare their competence in academic language. What is needed for such a comparison is some Archimedean point that can be used to compare students' language abilities, specifically their academic literacy, over the years. The tests of academic literacy levels referred to above might provide just such a fixed point. However, despite the fact that the TAG and TALL have been extensively pretested on groups with known academic language ability, there is no absolute guarantee that the difficulty of the tests has remained constant over the years. If, for example, the difficulty of the tests has increased over the years, one might arrive at the false conclusion that the academic literacy of first year students has deteriorated (while perhaps it has actually remained constant or has even risen). Needless to say, the outcomes of these analyses can have important consequences, both politically and for the lives of individual students.

The latter point deserves some further elaboration. Until now, the tests of academic literacy referred to here have been employed as low to medium stakes tests. That is, based on their outcomes, low performing students were compelled to enrol for an intervention programme at UP and NW, while in the case of US students there is in certain faculties a gradual phasing in of such programmes. In all of these cases, no major disaster occurs for the students if, as a result of having taken a more difficult test, their academic literacy is underestimated as compared to the academic literacy of students of previous years. Some of the students may in such a case be compelled to undergo additional tuition that they perhaps did not fully need. But the picture will change dramatically if the test should be used as an admissions test. By their nature, such tests are high stakes tests, since they partially determine access to university education, and the expected lucrative future earning power that follows on this. In such a case it seems imperative that some guarantee needs to be given that students with a given level of academic literacy will have the same likelihood of passing the test, independent of the year in which they took the test.[1]

One way of designing such a guarantee into the process of test development and administration is to make use of Item Response Theory (IRT) models rather than of classical test theory. A prerequisite for making use of the advantages of IRT modelling is that tests partly overlap, i.e. items in test of year (t) are to be found in exactly the same format in the test in year ($t + 1$) as well. Fortunately, this was the case for items of the TAG tests in 2005, 2006, and 2007, and these three administrations of the test will therefore provide the basis for our analysis. We are not yet in a position to do the same for the other

[1] Of course, an argument could potentially be made to back up a decision to vary the difficulty of the admission tests over the years, perhaps for reasons of capacity, or for other reasons (see discussion, below). But by seeking to build in a guarantee, one at least has control over difficulty levels.

(English) version of the test (TALL), since here overlap is still either too small or absent. Though some of the discussion and analysis below will therefore refer to both TAG and TALL, since they are parallel tests, we envisage doing similar analyses on TALL once the degree of overlap is sufficient for such an analysis to be made. In such a case, we would test whether the findings presented here present a similar pattern.

Method

Population and context

In January and February of 2005, 2006, and 2007, the academic literacy of all new undergraduate students of the University of Pretoria, the Potchefstroom and Vanderbijlpark campuses of North-West University, and the University of Stellenbosch was tested through the administration of the *Test of academic literacy levels* (TALL/TAG). At the University of Pretoria and University of North-West, students are allowed to sit for either the English (TALL) or Afrikaans (TAG) test, and so have the freedom of choosing whichever language they feel more comfortable with in the academic environment. At the University of Stellenbosch, however, students have to take both tests. At this university, the English test was administered one day after the Afrikaans test. In total 17,659 students participated (but they do not necessarily represent *different* students: see: Table 1, note); 9,449 took the Afrikaans test, while the remaining 8,210 students participated in the English version. See Table 1 for a detailed description.

Table 1: Population of first year students

TALL	UP	US*	NW	Total
2005	3,310	1,729	135	5,174
2006	3,652	3,710	143	7,505
2007	3,905	4,165	140	8,210
TAG	UP	US *	NW	Total
2005	2,701	1,702	2,521	6,924
2006	2,547	3,703	2,650	8,900
2007	2,582	4,160	2,707	9,449

* Note: Stellenbosch students took the TALL the day after they took the TAG.

The tests: TALL and TAG, and their design

The 2005 and 2006 versions of TALL and TAG each consists of 120 marks, distributed over seven subtests or sections (described in Van Dyk & Weideman 2004a; 2004b, Weideman 2006), six of which are in multiple-choice format:

Section 1: Scrambled text
Section 2: Understanding graphs and visual information
Section 3: Understanding texts
Section 4: Academic vocabulary
Section 5: Text types
Section 6: Text editing
Section 7: Writing (handwritten; marked and scored only for certain borderline cases)

The 2007 versions of TALL and TAG each consists of 100 marks, distributed over the first six subtests or sections which are in multiple-choice format. Section 7 was omitted from 2007 on; borderline cases, who are identified by statistical means, are allowed to take another test, the results of which are used to decide if the student has passed or failed the test, and has risk in terms of academic language.

Students have 60 minutes to complete the test, and they earn a maximum of 100 marks (some items counting 2 or 3 instead of 1). In another paper (Van der Slik & Weideman 2005; cf. too Weideman & Van der Slik 2007), the determination of the cut-off point has been discussed extensively. We will return to the issue of cut-off points below, where we will evaluate them in light of the outcomes of the IRT-based analyses.

Analyses

In order to perform IRT analyses, we make use here of the One-Parameter Logistic Model (OPLM) package developed by Norman Verhelst and his colleagues at CITO in the Netherlands (Verhelst, Glas & Verstralen 1995). IRT analyses represent an ability of persons such as, for example, academic literacy, in a mathematical model. In an IRT analysis, the ability is usually denoted by the Greek letter theta (θ). Persons with a high θ (ability) are expected to have a high chance to give correct responses to difficult items, while persons with low ability are expected to have a low likelihood to answer difficult items correctly. The attractiveness of IRT modelling – as compared to, for example, Guttman scaling – is that persons who get difficult items correct, still have the likelihood to respond incorrectly to less difficult items. Similarly, less able persons have a chance to respond correctly to difficult items. Guttman scaling does not allow for such “inconsistencies”.

Various mathematical models may be used to represent the characteristics just described, but the critically important consideration for choosing the appropriate model would, no doubt, be the degree to which it would fit the data.

In respect of IRT modelling, various fit measures can be employed to evaluate the model against the data, but a logistic curve has the most attractive set of features. One of the main models in IRT analyses is the Rasch-model. A less attractive feature of this model, however, is that it assumes that all items measuring a specific ability have the same discrimination index. The One-Parameter Model that we are using in the current analysis relaxes this restriction by allowing discrimination indices to vary. It may thus represent the data better, since it is well known in classical test theory that test items may vary rather considerably as regards their discriminating power.

One decisive advantage of IRT analyses over classical test theory, however, is that they can cope with incomplete designs. That is: the program can deal with different persons responding to different sets of items (or tests for that matter). See Figure 1.

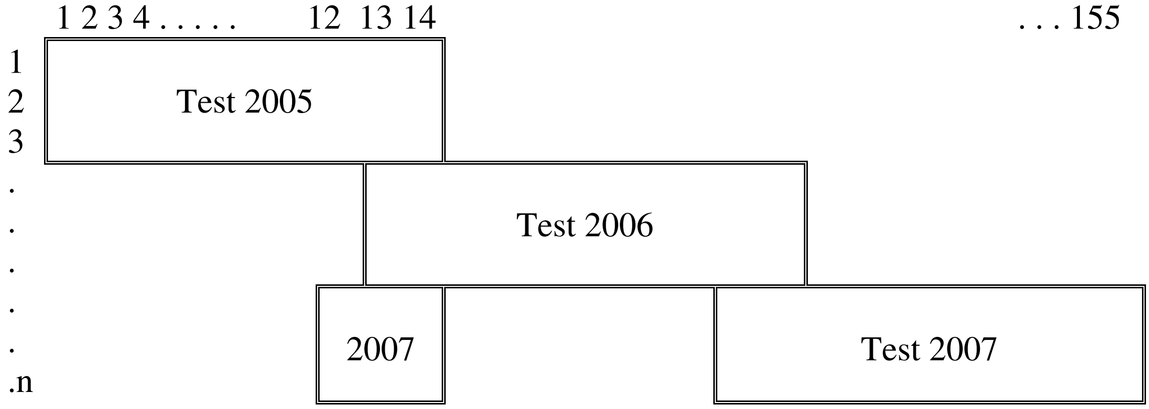

Figure 1: Persons by items matrix

In Figure 1, the rows represent persons taking the various tests, and the items are represented in the columns. This is a diagrammatic representation of three tests which overlap in part, i.e. items can be found in different tests and answered by different persons. For example, test item 12 is not just in Test 2005, but is found in Test 2007 as well, whereas items 13 and 14 are found in all three tests. Note that, for example, items 1 to 4 can only be found in Test 2005. In fact, Figure 1 represents the design we are working with in the present study. Note also that, since items may occur more than once, the number of unique items is smaller than the sum of the test items. In the present situation, 155 unique items are involved, whereas the total number of items is 62 + 62 + 63 = 187.

In case tests are partly overlapping (and therefore linked), the OPLM program is able to estimate an ability distribution in which the item parameters can be estimated independent of the characteristics of the population. Rather than the difficulty of the items, it is the likelihood of correct answers, taking into account a person's ability, that matters. As a consequence, the ability distribution can be used to equate different tests in such a way that cut-off points for the tests reflect equal ability, or, in this case, levels of academic literacy.

Results

Description of the sample

Table 2 depicts the outcomes at scale level for TALL. As can be seen, there is a general trend of decreasing mean scores for the three universities included in this study. Simultaneously, the cut-off points were set to a lower level each year. As a result, the percentages of students who failed to pass the TALL remained more or less constant for UP. In the case of North-West specifically, but also to some extent at the University of Pretoria, one of the deliberate reasons offered for a specific annual cut-off point is teaching capacity (cf. Van der Slik and Weideman 2005, 2007).

Table 2: Descriptive statistics of TALL

	UP	US	NW
MEAN (range 1 – 100)			
2005	71.75	76.89	59.70
2006	64.32	68.46	56.27
2007	61.11	64.98	50.44
Cut-off point			
2005	68.5	68.5	67.5
2006	55.5	55.5	49.5
2007	50.5	57.5	42.5
Percentage failed			
2005	34.26	22.73	56.30
2006	31.30	23.02	34.97
2007	31.50	32.58	40.00

Table 3 provides the outcomes for TAG. It can be seen that the general trends observed for TALL are also found in TAG. If one looks only at these numbers, the academic literacy of newly arrived students at the University of Pretoria, the University of Stellenbosch, and of North-West appears to have deteriorated over the years, thus providing an affirmative answer to the question posed in the subtitle of this paper.

Table 3: Descriptive statistics of TAG

	UP	US	NW
MEAN (range: 1 – 100)			
2005	70.16	63.15	63.08
2006	60.18	53.53	54.07
2007	56.66	51.78	51.14
Cut-off point			
2005	60.5	50.5	55.5
2006	50.5	50.5	49.5
2007	45.5	42.5	45.5
Percentage failed			
2005	23.84	26.85	31.14
2006	25.21	43.78	40.35
2007	24.98	33.25	38.27

These outcomes can also be visualized as in Figures 2 and 3:

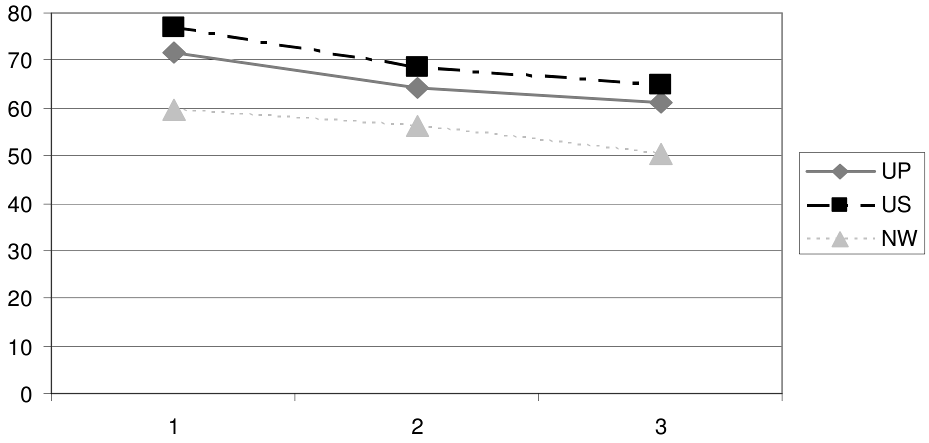

Figure 2: Mean scores on TALL for 2005(1), 2006(2), and 2007(3)

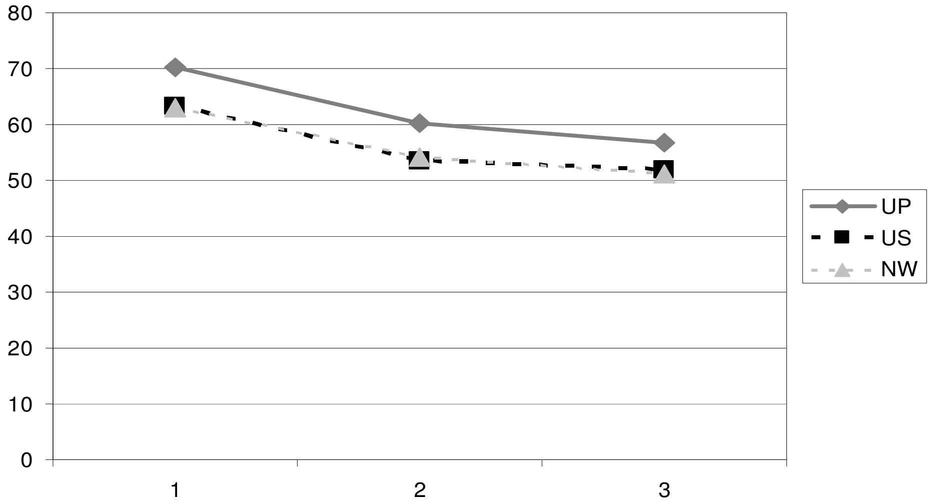

Figure 3: Mean scores on TAG for 2005(1), 2006(2), and 2007(3)

When Figures 2 and 3 are considered in isolation, one might be tempted to conclude that something is indeed wrong with the academic literacy of newly arrived students, since the outcomes consistently demonstrate a declining ability for the three universities involved, not just for the students taking the English test (TALL), but also for those who took the Afrikaans test (TAG).

But is this actually what happened? We have tested this hypothesis by means of OPLM. As was remarked above, we could, unfortunately, perform these analyses only for the Afrikaans test, because only these tests were linked, by partial overlap, in the manner we have described above. We did so by taking the TAG 2005 outcomes as the reference or Archimedean point. That is: we took mean ability associated with the proportion that has passed the test as the starting point for each university separately. OPLM has, in other words, made it possible for us to show in detail how ability scores are associated with test scores in 2005, 2006 and 2007. By transposing the 2005 mean ability scores onto the 2006 and 2007 test scores, we were able to estimate the proportion of students that would have passed these 2006 and 2007 tests, assuming ability equal to that of 2005. In such an analysis, 2005 is thus accepted as the base year, and the scores of the subsequent years are interpreted with reference to the mean scores of the base year. In Figures 4 and 5 we present the outcomes.

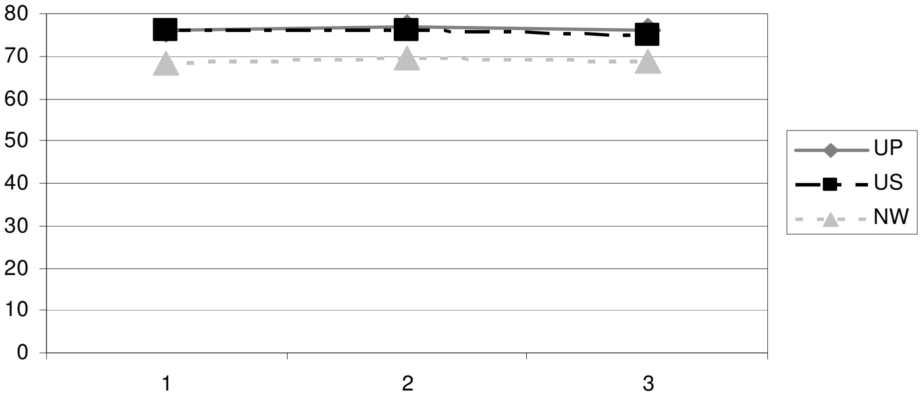

Figure 4: Proportions passed on TAG assuming equal difficulty for 2005(1), 2006(2), and 2007(3)

It seems quite obvious that Figure 4 leads to a completely different conclusion than Figure 3. Instead of a trend of declining ability, no apparent trend can be observed! The cause of this is that, instead of there being a declining ability in terms of academic literacy for students, the difficulty of the tests has in fact increased. And this increasing difficulty over the three years has not been fully compensated for by adjusting the cut-off points. Figure 5 graphically represents this.

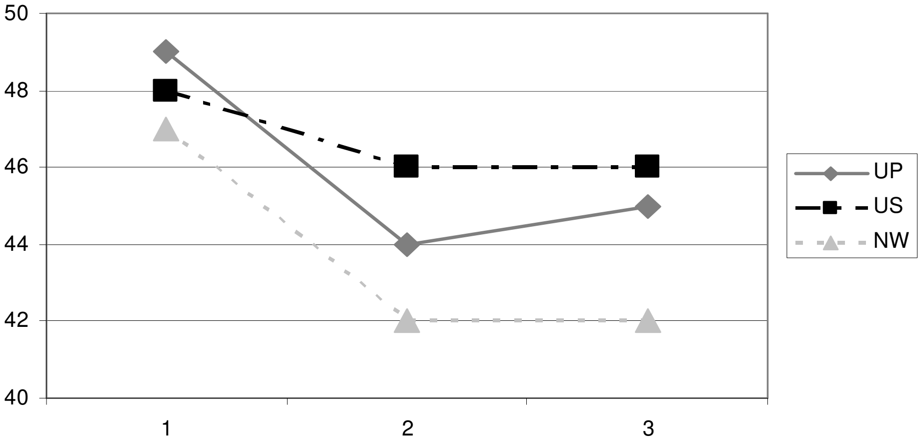

Figure 5: Cut-off points assuming equal difficulty for 2005(1), 2006(2), and 2007(3)

As can be seen in Figure 5, the cut-off points for the 2005 TAG test were around 48 marks. If we intend to measure a mean ability in 2006 and 2007 that is equal to the one measured in 2005, then students would have needed fewer marks (between 42 and 46) to make the cut-off point. Or put differently: If we equalise the tests in our analysis over the period in question, by holding their results steady in terms of the 2005 starting point, then we note that the 2005 test was in fact easier than the 2006 and 2007 tests, since a student who had scored around 48 marks on the 2005 test would have required a lower score (of between 42 and 46) on the 2006 and 2007 test in order to make the cut-off point. This implies that the 2005 test was easier than the 2006 and 2007 TAG tests.

The differences between the cut-off points of the three universities involved have to do with several factors. It can be seen, for example, that the drop off between 2005 and 2006 for North-West students is steeper than for Stellenbosch students. This may look odd, since they took identical tests; so the drop off may be expected to be of the same magnitude. Two comments can be made about this. Firstly, the observed differences may in part be coincidental, resulting from measurement error. Secondly, from the point of view of testing academic literacy levels, it may be that the students from North-West come from a population that is different from the population that the students of Pretoria and Stellenbosch are recruited from, in the sense that the universities involved may be employing different entry requirements. Thus different universities may land up with differences among their respective first year populations that are greater than one would expect at first glance.

Conclusion

In this article we have tested if the academic literacy of newly arrived students at three universities has deteriorated over the period 2005–2007. A superficial analysis may indeed indicate that this is the case. However, having performed analyses by means of the One-Parameter Logistic Model package (OPLM), we found nothing of this kind. On the contrary, rather than a decrease of academic literacy, the analyses have shown that the tests themselves have increased in difficulty. It is only when we do not fully compensate for this increased difficulty that it *appears* as if academic literacy has deteriorated. In fact, the academic literacy of newly arrived Afrikaans speaking students has proved to be remarkably constant over the past three years.

This study has several limitations which have to be addressed in future research. First, the time period under study may be too short to arrive at definite answers regards the possibility of a decline in academic literacy in South Africa. Second, we were only able to analyze the results of the Afrikaans test. This is unfortunate, because Afrikaans is taught mainly at formerly privileged schools, whereas English is taught at both formerly privileged and at formerly deprived and extremely deprived schools. Though the outcomes for the Afrikaans and the English test of academic literacy levels at first glance appear to be heading into the same direction, most of those with experience in research on language testing will be wary to rely solely on these impressions before they can be empirically tested and in some way quantitatively verified in a similar fashion. Thirdly, we have used only the TAG as an indicator of academic literacy levels. Other, similar tests may and should be used to study trends in academic literacy levels. This is not to imply that the results are due to the unreliability of the test. On the contrary, the TAG and TALL tests have proved to be highly reliable over the years. What we mean is that other measures than the standard paper-and-pencil tests might also be useful. Fourth, an issue that ties in with the second point above: our finding that academic literacy levels have remained more or less similar over the years applies to this, very specific and fairly select group of students. As one of the reviewers of this article has emphasized, we have worked here with a cohort of testees that is circumscribed in terms of language. What if the population is more varied in terms of first language, and if the ability that we have measured is not as evenly distributed across other test populations as we have assumed here? We would therefore need to test the kinds of conclusions we have reached here against a larger, and perhaps more diverse group. Finally, to take up another point of the same reviewer, the analysis challenges us as test developers and as users of the results of these kinds of tests to be careful about assuming equivalence among different versions of tests. When we interpret test results, we should refer to test difficulty, which implies a measure of comparison that we may not yet have or may not even have planned for. We cannot simply interpret such results at face value.

Notwithstanding these limitations, the outcomes of our analyses have made it quite clear, we think, that IRT modelling is a useful tool to get a better understanding of the difficulty of tests and test items. In that sense, it may enable us to arrive at responsible decisions that do more justice to those who take the tests, in that they are doubly accountable, both in being defensible in a theoretical or empirical sense, and in being accountable to a larger public (Weideman 2006, 2007). This is not only helpful for low stakes tests such as TALL and TAG, but also of critical importance when the stakes are higher, for example in the case of access or admission tests. There should really be no dispute about the condition that the likelihood of passing a test should depend on students' ability as expressed in terms of a score that captures their level of academic literacy. IRT modelling gives us one way of doing justice to the expression of that ability in the measurements that are made with similar instruments over time. The exceptions to this condition in our context occur generally in the case of tests that are used for access to higher education. In such a case a political decision may be taken where, say, those belonging to a specific, previously disadvantaged group who perform in the top three deciles of a test of academic literacy compared with their peers in that same group, may be granted access to university study, even though, in comparison with others that do not belong to their group, they may not have made it. The groundbreaking work for making such a decision responsibly has been done in South Africa by the Alternative Admissions Research Project of the University of Cape Town (cf. Cliff & Yeld 2006, Cliff, Yeld & Hanslo 2003, Visser & Hanslo 2005). Though such decisions clearly need to be taken on some political cue, and not one based solely on a measure of ability, they would need their own arguments to be defensible and justifiable. But even in this case, IRT modelling might prove to be a valuable tool for underpinning such arguments. It is conceivable, for instance, that if there are differences in academic literacy that are associated with membership of different cultural or other groups, such differences will not remain constant over the years. IRT modelling will also be particularly suitable to notice and monitor such changes.

Acknowledgements

We are indebted to two anonymous reviewers, and to Tobie van Dyk of the University of Stellenbosch, for suggesting improvements both to the consistency of the text and to the substance of the argument.

References

Cliff, A.F. & Yeld, N. 2006. Test domains and constructs: Academic literacy. In H. Griesel (ed.) 2006: 19-27. *Access and entry level benchmarks: The National Benchmark Tests project*. Pretoria: Higher Education South Africa.

Cliff, A.F., Yeld, N. & Hanslo, M. 2003. Assessing the academic literacy skills of entry-level students, using the Placement Test in English for Educational Purposes (PTEEP). Paper read at Bi-annual conference of the European Association for Research in Learning and Instruction (EARLI), Padova, Italy.

Rademeyer, A. 2007. SA onderwys stuur op ramp af, maan kenner. *Beeld*, 17 August: 12-13.

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

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

Van der Slik, F. & Weideman, A.J. 2007. 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.

Verhelst, N.D., C.A.W. Glas, & H.H.F.M. Verstralen. 1995. *One-parameter logistic model OPLM.* Arnhem: Cito.

Visser, A. & Hanslo, M. 2005. Approaches to predictive studies: Possibilities and challenges. *South African journal of higher education* 19 (6); 1160-1176.

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

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

Weideman, A.J. 2007. A responsible agenda for applied linguistics: Confessions of a philosopher. Keynote address, joint LSSA/SAALA/SAALT 2007 conference. Submitted to *Per linguam*.

Weideman, A.J. & Van der Slik. 2007. The stability of test design: Measuring difference in performance across several administrations of a test of academic literacy. Forthcoming in *Acta academica*.

Widdowson, H.G. 2005. Applied linguistics, interdisciplinarity, and disparate realities. In Bruthiaux, P., Atkinson, D., Eggington, W.G., Grabe, W. & Ramathan, V. (eds.) 2005: 12-25. *Directions in applied linguistics: Essays in honor of Robert B. Kaplan*. Clevedon: Multilingual Matters.