Per Linguam 2005 21(1):23-35
http://dx.doi.org/10.5785/21-1-70

Journal Home Page:
http://perlinguam.journals.ac.za

Per Linguam

A Journal for Language Learning
Tydskrif vir Taalaanleer

THE REFINEMENT OF A TEST OF ACADEMIC LITERACY

Frans van der Slik & Albert Weideman
University of Pretoria

To ensure fairness, test designers and developers strive to make their instruments for assessing the language abilities of learners as accurate and reliable as possible, and have traditionally used a number of techniques to ensure this. From a post-modern, and especially critical perspective, however, these measures are not enough to ensure fairness. In these approaches, fairness is redefined and reconceptualised. This article demonstrates that it is still possible to use conventional techniques to achieve the goal of, for example, increasing the accessibility of a test. Using several statistical analyses of the results of a test of academic literacy as examples, the article concludes that traditional, quantitative measures enhance and complement, rather than undermine, current concerns.

INTRODUCTION

Test designers and developers engage with the tests that they produce with a number of conventional goals in mind. First among these is to ensure that the empirical aspects of the tests they produce, and the items that make up these tests, conform to certain standards. This is routinely achieved through a number of fairly sophisticated techniques involving statistical analysis of various kinds. However, since the standards for test reliability and empirical validity are generally expressed in numerical terms, a number of commentators that approach testing from a post-modern, critical point of view have pointed out that if one concerns oneself merely with 'measurement issues while overlooking the social and political dimensions of tests' (Shohamy 2004: 72), one may be obscuring rather than opening up the test instrument for scrutiny and criticism, i.e. making it transparent and, subsequently, accountable.

There is no denying that measures of the fairness of tests have traditionally been expressed in terms of reliability and validity, and the latter generally in numerical format, often as indices. That such numerical indicators may have had the intention of imbuing the results of tests with 'scientific' authority, and may have served to mislead, may of course in certain instances be true. But it is certainly unacceptable, and irresponsible from a test developer's perspective. Perhaps test designers have not always been open enough in communicating the meaning of various empirical measures both to their peers and to the community at large, since some of the measures indeed pre-suppose a more sophisticated level of understanding than can always be explained quickly and in simple terms. The more serious issue raised by critical language testing, however, is whether we can examine the purposes to which the results of language tests can be put (Shohamy 2004; 2001) so as to ensure transparency and accountability.

Measuring merely the internal effects (i.e. power or validity of a test to do what it should) or the consistency of language tests (how reliably a test accomplishes this measurement) is no longer sufficient: every external effect of a test, i.e. the uses to which its results will be put, must be scrutinised.

A variety of uses for refinements of tests

The argument of this paper will be that the ongoing refinement of a test, through various empirical means, can indeed continue to serve several responsible purposes. Not only may such refinement be the basis of a routine and ongoing enhancement of the empirical measures of the validity of a test, but it can also serve to inform responsible decisions taken on the basis of test results. An illustration of the latter is the example given below of how measures of test reliability can be used to increase the accessibility of a test. The thesis of this paper is therefore that there is a reciprocity between traditional, quantitative measures of test quality, and the (often politically defined) ends to which test results can be used (cf. Weideman 2005).

For the examples of refinement of both tests and test use discussed below, we use various statistical analyses of an early version of the Test of Academic Literacy Levels (TALL; TAG in Afrikaans, for 'Toets van akademiese geletterdheidsvlakke') used by the Universities of Pretoria, Stellenbosch, and Northwest, and show how these can be employed to refine the test further, and assist in taking decisions about the accessibility of the test. Though this is a low to medium stakes test, the lessons learned here may be useful for higher stakes tests, especially those that are currently being developed under the auspices of Higher Education South Africa (HESA) as national tests benchmarking academic literacy, quantitative literacy and mathematics, since these will almost certainly raise questions about the fairness of the judgements being made on the basis of the results, as well as the stigma that may attach to receiving a 'fail' grade on them. The test currently under discussion (TALL/TAG) is a lower stakes test since it is a placement, and not an access test. It is administered only after students have already enrolled, and its results are utilised to determine what level of institutional support may be necessary to bring to the appropriate level the academic literacy of the candidates that it identifies as being at risk.

The ongoing refinement of tests takes place at various levels and stages of test design. Before the first administration of a test, it is generally piloted thoroughly. This means that all items that will be used in the final draft of the test will be pre-tested, and their performance analysed. Through an item analysis, the test designer will normally determine how well each item discriminates, i.e. distinguish between those with high and low scores on the test as a whole, and how easy or difficult an item is. By setting certain parameters for acceptability, for example accepting, as in the case of TALL, only those items with a discrimination value lying higher than .25, and a facility value between .25 and .84, the test developer is able to sift through the items that will probably perform well, and those that will probably perform worse.

There are also a number of post-test refinements that can be brought into play, and the current article focuses mainly on those. Again, as in the pre-testing, piloting phase, both an item analysis and a test analysis can be made. The item analysis is likely to have the same format

Per Linguam 2005 21(1):23-35
http://dx.doi.org/10.5785/21-1-70

as that of the piloting phase, yielding measures of facility and discrimination for each individual test item. The test analysis is aimed at improving the reliability of a test, so may use one or several measures of reliability that are appropriate to the particular test. In addition, as we shall see below, the reliability indices also yield possible measurement errors. Additional test analyses involve factor analyses that may provide another measure of reliability, in determining the heterogeneity or homogeneity of a test, i.e. whether it measures only one ability or trait, or more than one. This will be a particularly interesting measure of the test we use here as an example, and will be discussed in more detail below.

METHOD

Population, test design and context

In January and February 2004, the Test of academic literacy levels (TALL/TAG) was used to test the academic literacy of all new undergraduate students of the University of Pretoria. The same test was used, a week or so later, at the Potchefstroom and Vanderbijlpark campuses of the University of Northwest. Since 2005, the University of Stellenbosch has also used the test. The analysis here, however, is restricted to the 2004 administration of the test. Since the administrators of the test were unhappy with various characteristics of the previous test (cf. Van Dyk & Weideman, 2004a, 2004b), a new test had been designed, and this was the first time that the test was administered. 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. These two languages are also the languages in which lectures are presented at the University of Pretoria. In total 6,310 students participated; 3,033 took the Afrikaans test, while the remaining 3,277 students decided to take the English version. As has been noted above, the test is a placement, and not an access test: the political decision of whether the candidate taking it should be allowed to enter university has already been answered (positively) when the test is administered.

The tests: TALL 2004 and TAG 2004

The 2004 versions of the Test of Academic Literacy Levels (TALL) and the 'Toets van Akademiese Geletterdheidsvlakke' (TAG) each consists of 71 items, distributed over eight subtests or sections (described in Van Dyk & Weideman, 2004a), seven of which are in multiple-choice format:

- Section 1: Scrambled text (5 items)
- Section 2: Understanding graphic information (6 items)
- Section 3: Dictionary definitions (5 items)
- Section 4: Register and text type (5 items)
- Section 5: Academic vocabulary (15 items)
- Section 6: Comprehension and numerical comparisons (21 items)
- Section 7: Text editing (14 items)
- Section 8: Writing (handwritten; marked and scored only for certain borderline cases)

Per Linguam 2005 21(1):23-35
http://dx.doi.org/10.5785/21-1-70

The English and Afrikaans versions were almost, but not entirely, equivalent. The content of the section 'text editing' was different for both tests. Students had 60 minutes to complete the test, and they could earn a maximum of 100 points (some 29 items counting 2 instead of 1).

Based on the results of intense pre-testing and past test experiences with the same population (Van Dyk & Weideman, 2004a), as well as an early, initial analysis of the results for this administration, the cut-off point for the English test was set at 58 points, while the cut-off point for the Afrikaans version was set at 52 points. Though the determination of the cut-off points is not the focus of this article, a brief explanation is nonetheless in order. Historically, i.e. on the previous test (ELSA Plus), between 24% and 27% of students who wrote the Afrikaans test were identified as being at risk, while between 31% and 38% of students who took the English test failed. Based on past experience, on average about 31% of the total number of students is identified as being at risk. The main argument for determining the cut-off point, however, relates to the comparative performance of the candidates who write a specific test on a specific day, and compares their performance with that of their peers on the same test. So, for example, the cut-off point for the Afrikaans test historically has been 10% below the group average, and that for the English version 4% below the average. In order to cross-check the validity of this historical criterion, the test administrators consider whether the use of this average identifies groups, that fail and pass, which are of similar size in proportion to the total as those mentioned above, and small adjustments may then be made. A further reason for the difference between the two cut-off points is that the Afrikaans group consists of predominantly first language users of the language, while the English group contains proportionately fewer first language users, since it is made up mainly of users of English as an additional language (i.e. English as a second, but more likely third or fourth language). A further criterion that may come into play at a later stage is the capacity of the institution to provide an intervention that will help to remedy the lack of ability that has been identified. So far, it has not been necessary to use such an additional criterion to determine the cut-off point for tests. However, just as it is irresponsible to identify a language development problem without providing a solution in the form of an intervention, it would be reasonable to argue that there may be limits to the scope of such an intervention.

ANALYSIS

In order to analyze the test results of the University of Pretoria students, we made use mainly 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 were used to evaluate the psychometric properties of the tests in this study. We also used them to produce 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). In addition, we also produced a visual representation of distribution of the P-values. A uniform distribution between values of .20 and .80 is generally seen as advisable.

At the test level we made use of reliability statistics, Cronbach's α and GLB or Greatest Lower Bound reliability (cf. Verhelst, 2000). There are several ways to define reliability. On the one hand, a test is said to be reliable if the rank order of the testees is the same when the test is repeated a large number of times (parallel tests). On the other hand, Cronbach's α is

Per Linguam 2005 21(1):23-35
http://dx.doi.org/10.5785/21-1-70

defined as a measure that depicts the degree to which the observed scores represent the 'true' scores (i.e., without measurement error). GLB is a measure comparable to Cronbach's α. GLB, however, does not assume homogeneity, and will be higher than α in case the underlying concept is multidimensional (the test is heterogeneous). It is argued that GLB is a closer estimate of a test's 'true' reliability than Cronbach's α is (Jackson & Agunwamba, 1977). In the case of TALL/TAG, as we shall note below, the homogeneity or heterogeneity of the test is indeed an issue.

Since an academic literacy – or for that matter any other – test is never entirely reliable, some testees may fail in cases where they should have passed, and vice versa. TIAPLUS provides four measures of the total amount of potential misclassifications that could have occurred due to imperfect measurement. These measures are either GLB- or Alpha-based, and they can, in addition, either be based on the assumption of a hypothetical parallel test or based on the difference between observed and 'true' scores.

We also used factor analyses to gain insight into the question as to whether the tests are one-dimensional, i.e. are homogeneous in what they measure.

RESULTS

Overview

Table 1 depicts the outcomes at the scale level. Clearly, the TALL as well as the TAG are highly reliable, both in terms of alpha (.92 and .83, respectively) and GLB reliability (.95 and .90, respectively). In addition, the average Rit-values, indicative of the discriminative power of the items, appear to be sufficiently high as well (.43 and .30, for the TALL and TAG, respectively). It can be observed that approximately 37% of those who took the English test failed, while approximately 22% of those who took the Afrikaans version did not pass. It is important to remember, however, that the cut-off point for the TALL was 58, while the caesura for the Afrikaans version of the test was 52.

	English version (TALL)	Afrikaans version (TAG)
N	3,277	3,033
Number of items	71	71
Range	0–100	0–100
Mean / average P-value	63.89	60.39
Standard deviation	19.32	12.76
Cronbach's alpha	.92	.83
GLB	.95	.90
Standard Error of Measurement	5.37	5.22
Average Rit	.43	.30
Cut-off point	58	52
Percentage failed	37.29	21.94

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

Per Linguam 2005 21(1):23-35
http://dx.doi.org/10.5785/21-1-70

Critical analysis of the effects of a test would include a close examination of test reliability. As can be seen from the analyses tabulated above these tests have high reliability measures (both alpha and GLB). These values are higher than for many other high stakes tests.

It appears that the mean test score for the English version is higher than the average test score of those who took the Afrikaans version; we will return to this in a moment. In addition, the variation around the mean is smaller for the Afrikaans version than for the TALL, implying that the academic literacy of those who took the Afrikaans version is more homogeneous than the academic literacy of those who took the English version. The latter is to be expected, given the fact that the Afrikaans test population consists of predominantly first-language users of Afrikaans, while the English population, although made up of some first-language users of English, consists mainly of English as additional language users.

In Figure 1 below, the distribution of facility (P-) values is represented for both tests. Ideally, test designers look for a uniform distribution of values between .20 and .80.

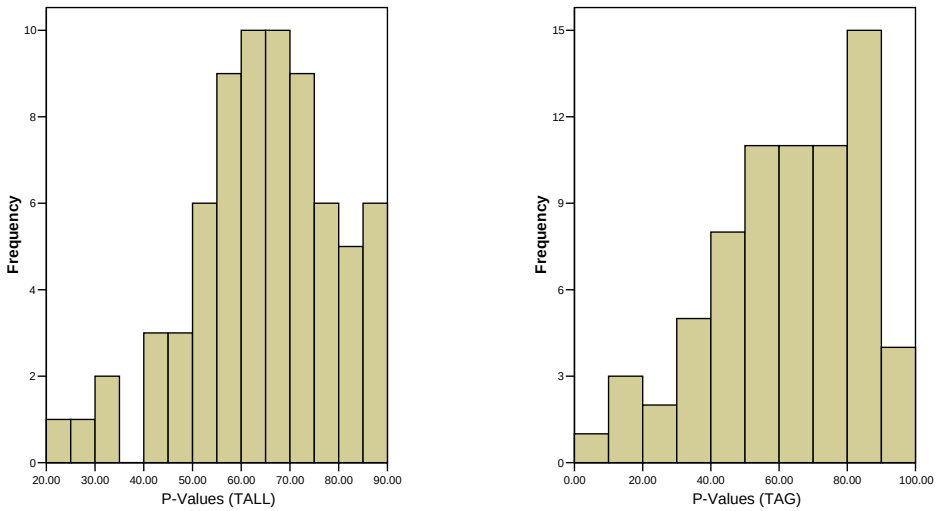

Figure 1: Distribution of P-values (English version left panel; Afrikaans version right panel)

Figure 1 provides a visual representation of the distribution of the P-values for both the English (left histogram) and the Afrikaans version (histogram to the right). It can be concluded that the P-values are not uniformly distributed. Low P-values refer to difficult individual items, and it seems that these items are underrepresented in both versions of the academic literacy tests. The Kolomogorov-Smirnov test proved that these distributions are not uniformly distributed (TALL: $Z = 2.40$, $p < .001$; TAG: $Z = 2.16$, $p < .001$). We return below to the implications of this for the refinement of subsequent versions of the test.

Misclassifications

Since no test is perfectly reliable, an important potentially negative effect of its administration may be that it fails to categorise correctly those who take it. In other words, because its reliability is never 100%, a test may misclassify a failure as a pass, and vice versa. Thus it is important for test designers and administrators to find a way of dealing with the

Per Linguam 2005 21(1):23-35
http://dx.doi.org/10.5785/21-1-70

extent to which a test may misclassify candidates who have taken it. In Table 2 we present the number of potential misclassifications based on four different criteria.

	English version (TALL)	Afrikaans version (TAG)
Alpha based:		
Correlation between test and hypothetical parallel test	397 (12.1%)	441 (14.5%)
Correlation between observed and 'true' scores	282 (8.6%)	318 (10.5%)
GLB based:		
Correlation between test and hypothetical parallel test	319 (9.7%)	357 (11.8%)
Correlation between observed and 'true' scores	226 (6.9%)	255 (8.4%)

Table 2: Potential misclassifications on the English and Afrikaans version of the academic literacy test

As can be seen, the amount of potential misclassification on the English test varies between 226 and 397, depending on which criterion is applied. Remember, however, that approximately half of the potential misclassifications stems from testees who have passed were they should have failed. So, between 113 and 199 testees who should have passed, may have failed.

Applying the same logic to the Afrikaans test, potentially 128 to 221 testees may have undeservedly failed.

There is a clear indication here that, in order to ensure fair treatment by the test, this measure should be used in some way to eliminate undesirable results. We return to this below.

DIMENSIONALITY

The question whether the academic literacy test is one-dimensional or multi-dimensional is of some interest, given the evolution of the test construct (Van Dyk & Weideman 2004a) and the particular selection of task types (Van Dyk & Weideman 2004b). In order to answer questions relating to how many dimensions of academic language ability the test measures, we performed a Factor Analysis on the various items in both tests. The outcomes are depicted in Figure 2. As can be seen in the top panel, the English test is virtually one-dimensional, i.e. most items cluster in more or less the same range. Interestingly, however, the items of section 7 (text editing) are situated in the lower right corner. If, of course, a test is not one-dimensional, it may be argued that it lacks homogeneity.

The apparent lack of homogeneity that is revealed by items in Section 7 clustering away from the rest may indicate a number of different issues. For one thing, they may simply be badly performing items, and therefore in need of replacement. This eventuality is unlikely, however, since an item analysis for this, also done with the TiaPlus statistical package,

Per Linguam 2005 21(1):23-35
http://dx.doi.org/10.5785/21-1-70

indicates that the both their facility values (i.e. the degree of difficulty, or P-value in the terminology adopted here) and their discrimination values (Rit) fall within the parameters set by the test developers and administrators for TALL (see Table 3, below). The same applies, with two exceptions, to the values for the equivalent section in TAG. One of the reasons that the items perform well is that they were thoroughly pre-tested or piloted. Van Dyk & Weideman (2004b) provide a detailed, critical examination of the justification for using this type of test.

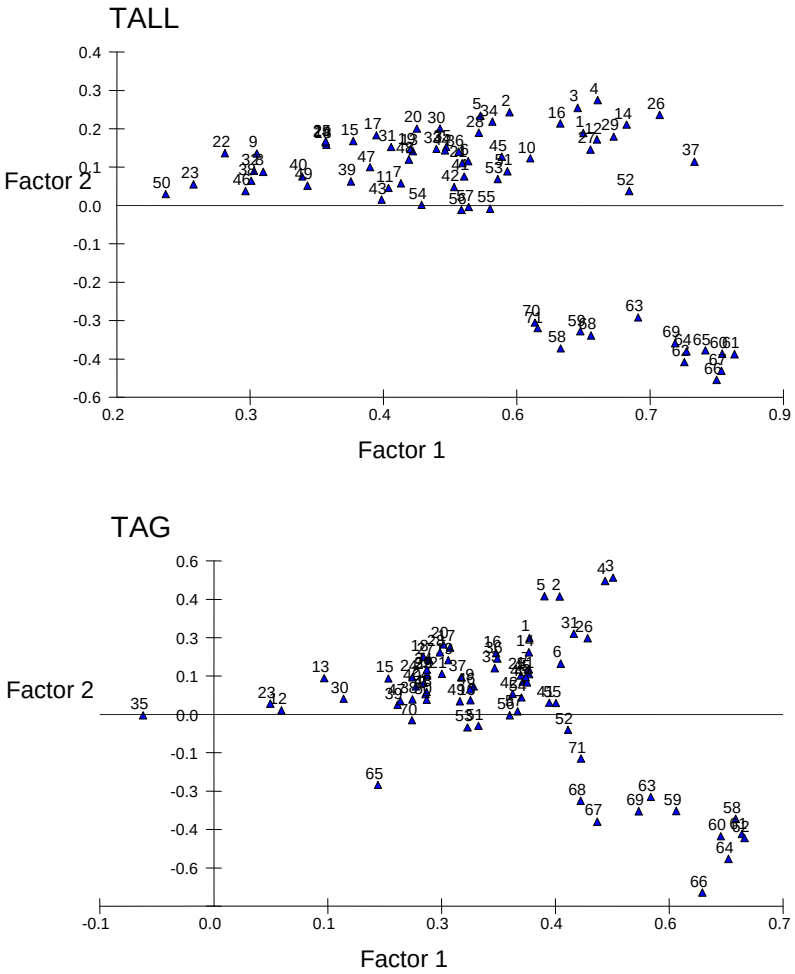

Figure 2: Factor analyses of the English (top panel) and Afrikaans (bottom panel) academic literacy test

However, if the items seem to perform adequately on their own, as is evident from Table 3 below, what other reasons can there be for the slight lack of homogeneity in this part of the test? A probable reason for the items in this section performing differently (in terms of measures of test homogeneity/heterogeneity) is that they are testing, in addition to academic language, some coding and decoding skills, since the format in which they are presented to testees requires a good measure of interpretation and understanding. If this is in fact the case, however, it may be worthwhile to keep them in the test, since on the face of it the abilities they test are indeed part of academic literacy. They may, in other words, measure a different aspect of academic literacy. Furthermore, the measure of heterogeneity that they introduce may be tolerable, since it does not appear to be that big.

Per Linguam 2005 21(1):23-35
http://dx.doi.org/10.5785/21-1-70

There is the likelihood, which merits some further investigation, that the performance of items in the section on text editing indicates that 'the construct of academic literacy being tested is indeed itself so varied that it cannot be reduced to a single, homogeneous idea' (Weideman 2005). If the idea of academic literacy that informs the test is itself so open, rich and varied – as the test developers indeed claim — that it cannot be reduced to a single, homogeneous ability, then the test makers are faced with a difficult choice: either tolerate the inclusion of a task type that makes the test less homogeneous, but more defensible on the grounds of contextual appropriateness, or omit such a set of questions on grounds of effectiveness.

Item number	Facility (P-value)	Discrimination index (Rit)
58	64	.46
59	59	.48
60	71	.59
61	74	.59
62	69	.56
63	67	.53
64	65	.57
65	62	.59
66	70	.58
67	70	.59
68	53	.49
69	56	.56
70	44	.45
71	44	.44

Table 3: Facility and discrimination values of TALL Section 7 (text editing)

The Afrikaans test appears to be equally close to uni-dimensionality, though, again, the text editing items appear to represent another aspect of academic literacy as well (Figure 2, bottom panel).

DISCUSSION

The analysis suggests a number of issues that either point to improvement of the technical aspects of the test, or that impact on the use to which the results of the test may be put.

As regards the former, it is clear that the test can be considerably improved technically if the distribution of difficulty (P-) values is made more uniform. The inclusion of a greater number

Per Linguam 2005 21(1):23-35
http://dx.doi.org/10.5785/21-1-70

of difficult items is indicated by the analysis, and is a pointer to the test developers about the design of a subsequent test. Setting narrower parameters for the facility value of items on the test, which is in any case desirable, may partly achieve, or assist in achieving this. Is this feasible? Since 2004, when this first version of TALL which is being reported on here, was used, the test has been re-administered at least three times, in three different institutional contexts, and to more than 14,000 students. To do this required at least twice as many fully fledged pilots. This means, in effect, that currently there are many more items stored in the TALL/TAG item bank, and, especially for items that are not text-dependent, both a selection of more difficult items and a narrower spread is possible.

Since some of the task types (especially for the sections on text comprehension and text editing) are dependent on texts, it may in those cases be less easy to find either more difficult items, or a narrower spread of items. There simply may not be enough items when a narrower set of parameters than the current .20 to .80 values is applied.

There are two remedies for this: one is to reconsider the detailed item analyses of items that have already been piloted, and see whether there is an obvious explanation for their not functioning as well as expected, try to fix that, and pilot again. The other is to pilot several further clusters of items to see, in light of the lessons learned, whether they may not yield improved results.

There are two remarks that should be made here, regardless of the solution sought. The first is that, though the items show too much variation in P-value, their average (63.89 for the English, 60.39 for the Afrikaans) is not entirely unexpected, given the high degree of selection of testees. That is, one may expect those who have already gained access to university generally to perform well. The reverse of this is, of course, the additional challenge to the test designers to set tests that are challenging enough by including more difficult items. The second is that, especially in the case of those items that fall outside the parameters on the current upper limit, i.e. are higher than 84, often function as the first, introductory items of each different task type. In order not to frighten testees by leading with questions that are difficult, test developers often tolerate an easy item or two to begin with.

Another technical aspect of the test that seems to require some discussion is the measure of heterogeneity that was revealed by the factor analysis described above (Figure 2). The discussion there has already indicated that the test designers are in this instance faced with a choice between the efficiency of having a completely homogeneous test and the appropriateness of the particular task type for measuring academic literacy. Such choices are often the basis of trade-offs that test developers have to make. In the present case, the arguments for a rich and varied construct are the most important consideration. The TALL/TAG designers claim to have designed an instrument that specifically acknowledges that academic discourse is an acquired, secondary discourse (cf. Gee 1998), with its own requirements and text types or genres, and have articulated the construct so as to capture this particular variety of language (Van Dyk & Weideman 2004a). If some degree of heterogeneity is what is required to ensure contextual appropriateness, it may have to be tolerated.

Per Linguam 2005 21(1):23-35
http://dx.doi.org/10.5785/21-1-70

In light of the test being less homogeneous, the indications are that the more appropriate measure of reliability is GLB. For both TALL and TAG, these are in any event fairly similar to the alpha measures yielded by the initial analyses.

The use to which the results of the test are put, viz. to channel students into various kinds of academic literacy support courses (cf. Unit for Language Skills Development 2005 for details) only after they have gained access to university, makes the test a low to medium stakes test, rather than a high stakes test, such as one whose results would determine or co-determine access to university. The measures of reliability, which for both tests remain around .9 or above across various versions, seem to indicate that the test is more than adequate for this purpose. Nonetheless, owing to a number of factors and prejudices present in the institutional context in which the test is administered, the application of the results of the test is not unproblematic. Most significantly in this respect perhaps is the degree to which some students who take the test feel that there is a stigma attached to being 'passed' or 'failed' by it.

In order to destigmatise the test results, the test administrators have devised several strategies. One is to make the results known not in two categories (pass or fail), but to grade results in terms of the measure of risk. In 2005, the results were published in five risk categories, from 1 (for very high risk) to 5 (little or no risk). It follows that for these categorisations to be made, one needs some foundation. At the present moment there is not yet a basis beyond arguments that say, for example, that achieving 10% or 20% below the test average constitutes very high or high risk. It is a challenge to give these some further empirical basis at some stage in the near future.

A more important destigmatising strategy, does, however, indeed already derive from some of the empirical measures that have been discussed in this paper. For those who have potentially been treated unfairly by the measuring instrument, a borderline cases test has been instituted, written about two to four weeks after the first. The statistical measures discussed here (cf. specifically Table 2) indicate that between 113 and 199 testees may have been misclassified by TALL, and a roughly similar number by TAG. These measures allow the test administrators to take a rational, responsible decision as to who should be allowed to write such a second-chance test. The measures have a further, economic aspect to them in that they eliminate the need to consider (and mark, with all the problems associated with ensuring inter-marking reliability) the writing section of the current test.

An additional idea that is being considered at present is to use these statistical measures to inform decisions on how to make further subsequent opportunities available to testees who rightly or wrongly perceive that they have been unfairly treated by the test. But the main point is that we now have an empirical basis on which to set parameters for such a decision. Of course, not every testee takes well to tests and particular test formats, such as multiple choice tests, for example. In fact, in the tests of academic literacy devised by the Alternative Admissions Research Project based at the University of Cape Town, in which the TALL/TAG developers also participate, female students have – contrary to some persistent prejudices and erroneous assumptions – consistently performed better than their male counterparts. For those who simply do not like the test format, there are a number of alternatives. Once on the academic literacy courses, they are subjected to continuous assessment of a variety of formats, and, upon outperforming on these, may demonstrate that

Per Linguam 2005 21(1):23-35
http://dx.doi.org/10.5785/21-1-70

their level of academic literacy is indeed adequate. There are, in other words, ways of demonstrating academic literacy other than through a single or second test.

In all of the decisions on the use of test results that have been discussed in this paper, it is clear that statistical data can inform responsible decisions, and provide ways of making tests more accessible. Of course, accessibility involves not only the opportunity to have more than one chance to write a test, but relates also to issues of transparency and accountability (cf. Weideman 2005). It is in that context that the notion of fairness should also be redefined, and remain the purpose of test refinement.

REFERENCES

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

GEE, JP. 1998. What is literacy? In Zamel, V & R Spack (eds), *Negotiating academic literacies: Teaching and learning across languages and cultures*. Mahwah, New Jersey: Lawrence Erlbaum, 1-59. Reprint of 1987 article in *Teaching and Learning: The Journal of Natural Enquiry*, 2:3-11.

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

JACKSON, PW & CC 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.

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

UNIT FOR LANGUAGE SKILLS DEVELOPMENT (2005). Academic literacy and sample test. [Online]. Available http://www.up.ac.za/academic/humanities/eng/eng/unitlangskills/eng/fac.htm. Accessed 14 April 2005.

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, ND. 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.

WEIDEMAN, AJ. 2005. Integrity and accountability in applied linguistics. Forthcoming in *South African Journal of Linguistics and Applied Language Studies*.

Biographic Note

Frans van der Slik teaches at the Radboud University of Nijmegen, and is also attached to the Unit for Academic Literacy of the University of Pretoria as research associate. Albert Weideman is director of the Unit for Academic Literacy. His email address is: albert.weideman@up.ac.za

Per Linguam 2005 21(1):23-35
http://dx.doi.org/10.5785/21-1-70