Johann L. van der Walt
and
H.S. Steyn
North-West University (Potchefstroom Campus)

Towards standardisation: A comparison of two versions of an academic literacy test

Abstract

All versions of a standardised test should be at similar difficulty levels. In this article, we investigate whether two versions of TAG ("Toets van Akademiese Geletterdheidsvlakke") are valid and whether they are at the same difficulty level. A group of students wrote two versions of the test within a ten-week period. We first investigated their validity in terms of sampling, scoring and construct evidence. Before comparing the tests, we checked whether the classes the study population attended had any effect on the results of the second test. We then compared the scores of the tests by means of a Rasch analysis, an equipercentile measure and a Bland-Altman plot. Test 2 proved to be easier than Test 1. Various factors may have contributed to this, and although it is hard to achieve in practice, it was clear that further work is required to ensure that TAG tests are at more or less similar difficulty levels.

Keywords: standardisation, academic literacy, assessment

Journal for Language Teaching, Volume 48, Number 1, pp. 109 - 129. 2014.
ISSN 0259-9570. http://dx.doi.org/10.4314/jlt.v48i1.6

1. Introduction

Most tests come in more than one version. One such test, the *Toets van Akademiese Geletterdheidsvlakke* (TAG), is widely used at South African universities for purposes of placing students in an appropriate academic literacy course. Various versions of this test have been developed over a number of years, amounting to a battery of tests from which one can be selected. Some similarity and equivalence is assumed for all these versions: they are based on the same construct; have similar content; are administered under the same conditions; and are scored in the same manner. They have also been trialled before use. It is therefore presumed that they have at least satisfied some of the conditions normally associated with validation processes, and to ensure their reliability. They qualify to be called 'standardised' in terms of some definitions of 'standardised tests'.

The question, however, is whether these versions are indeed at the same level – are they equally difficult (or easy)? This is an important issue, as students can be expected to achieve lower marks in the more difficult versions. This would affect the fairness of the assessment, as it is usually used for placement in a course. The purpose of this article is to compare two versions of TAG in terms of their score characteristics after they were administered to the same study population – something that is often difficult to achieve in practice, as the tests are normally only administered to first-years just before the beginning of the academic year. This would provide an indication of whether the test could be regarded as a standardised one.

In this article we first investigate the validity of each version by advancing an interpretive argument based on empirical data, and arriving at a conclusion regarding their validity. It is necessary first to establish whether the two tests are valid, as it would make no sense to compare a valid test with an invalid one. We then compare the scores of the two tests in order to determine whether their standard is the same.

2. Establishing validity

The validity of any test can only be accessed by means of a validation procedure (Van der Walt & Steyn, 2007:141). The interpretation of test scores and their uses are validated in such a procedure. Validation therefore entails the making of inferences – what the scores of a test mean and how useful they are. It takes the form of an interpretive argument in which evidence is collected and systematically presented, and the case for validity is weighed and argued. A number of inferences are usually made in test score interpretations. Kane (2001:330; 2006:24) states that the inferences commonly used include *scoring, generalization, extrapolation and utilization*. He also makes provision for a theory-based or explanation (i.e. construct) inference (Kane, 2001:330). Chapelle (2012) adds an additional one, sampling, and includes it as the first step in any interpretive argument. This initial argument is then evaluated by means of a validity argument in order to arrive at a conclusion regarding the validity of a test.

Three inferences are relevant for obtaining an indication of the validity of the two tests, viz. sampling, scoring and explanation. The *sampling* inference provides a description of the targeted domain (academic study at university in the case of this article) (Chapelle et al., 2008:14) and ensures that the content of the test is valid. This typically takes the form of a description of all the tasks the learner has to perform in the specific domain (Bachman, 2002:15). A representative sample of tasks is then drawn from the list and included in the test. Tasks in an academic domain are usually very diverse, and this complicates their selection in an academic literacy test.

The *scoring* inference is a relatively simple one in a multiple choice test, as is the case here. The main issues are ones of reliability – that the test is reliable and internally consistent – and that as few learners as possible are misclassified (cf. Van der Walt, 2012:149).

The *explanation* inference is theory-based, and makes provision for the consideration of the construct validity of a test. It is notoriously difficult to arrive at a precise and agreed-on definition of a language ability construct (cf. Chapelle et al., 2010:4 and Purpura, 2010:55, for example). (This is also the case with academic literacy.) Because of this, Kane (2001:327) states that validation does not require any specific formal theory. However, many language test designers feel that a definition of the construct should be the starting point for all test design, and the explanation inference enables one to ascribe test performance to an underlying ability.

3. Method of Research

Our study population consisted of 1582 Afrikaans-speaking first-year students at the Potchefstroom campus of North-West University in 2012. As mentioned above, two versions of the TAG test were used. Both were based on the same blueprint used for all these tests (cf. Van Dyk & Weideman, 2004).

The students wrote the first test before classes commenced for the academic year (in January, during the university's Orientation Week for first-years), and the second one in April, ten weeks later. During the intervening period, the students attended lectures in academic literacy. In order to obtain evidence backing the interpretation of the three inferences, we analysed the targeted domain (academic study at university), and collected evidence of the scoring of the tests (reliability coefficients, correlations with Grade 12 English and Afrikaans, and misclassifications) as well as evidence of the construct being measured (internal correlations, principal component analysis and factor analysis). This enabled us to arrive at a conclusion regarding the relative validity of the two tests, before the two tests could be compared.

In our comparison of the two tests, we first had to establish whether the academic literacy classes that the group attended for ten weeks before taking Test 2 had any effect on the test scores. We then conducted a Rasch analysis, where the probability of a correct response is a function of the test-taker's ability, item difficulty and a chance of scoring or the guessing factor associated with each item (cf. Mohandas, 2007:5). Following this, we performed an equipercentile equating, which defines a non-linear relationship by equalising the percentile ranks for each mark/score point. Mohandas (2007: 3) explains this as follows: "Equipercentile equating is used when two test forms ... are equally reliable and parallel measures in the sense that both forms are measures of the same underlying trait and the percentile ranks of the two tests of scores ... can be considered equal". Finally, a Bland-Altman plot (cf. Bland & Altman, 1999) was drawn to obtain additional information on the relative difficulty of the two tests.

4. Validity of the two tests

We now present validity evidence for the TAG tests in terms of sampling, scoring and explanation (construct) evidence.

4.1 Sampling inference

The test domain is academic study at university. This means that academic tasks performed at university must be tested. The problem, however, is that this is a very broad and diverse field. Macro-categories, such as writing of assignments, reading textbooks, taking part in seminars and so on, are usually regarded as typical academic tasks at university. Ideally, these tasks should be assessed by means of a direct test, and relate to the field of study the students is to undertake. This, however, would make assessment extremely cumbersome. The approach adopted in the TAG is to compile test items that are based on an abstraction of university tasks. These include micro-level tasks such as ordering information, interpreting a text and graphic data, making inferences, understanding academic vocabulary and so on (cf. Van Dyk & Weideman, 2004: 10; Weideman, 2007: xi-xii for the original blueprint). These are generic skills that underlie successful university study. Further evidence of validity here is the fact that the test has generally been accepted as a test of academic literacy and is used at a number of universities for placement purposes, as mentioned above. (The ICDELDA website http://icelda.sun.ac.za provides further information on the contexts of use of the test.)

4.2 Scoring inference

Evidence for the scoring inference includes reliability and misclassifications. The latter were calculated because a cut-off mark was used in Test 1 and a pass mark applied in Test 2.

The reliability of the two tests as indicated by their Cronbach alpha coefficients was 0.78 and 0.88 respectively. Weir (2005:29) sets 0.80 as the generally-accepted criterion, so only Test 2 satisfied it, although the reliability of the first one was only slightly lower than the criterion.

In Table 1 the Pearson correlation coefficients between test totals and the Grade 12 marks for English and Afrikaans are displayed. The results are given for all students (numbers indicated by the n-values) for which scores and marks were available (3098 students wrote Test 1; 1902 students wrote Test 2; our study population was 1582). In the lower triangle of the table (displayed in boldface numbers) the correlations attenuated for the reliabilities of the Test 1 and Test 2 total scores are given (cf. Hunter & Schmidt, 2004:96).

Table 1: Correlations for Test 1, Test 2, Gr 12 English and Afrikaans marks

	Gr 12 English	Gr 12 Afrikaans	Total Test 1	Total est 2
			(n=3098)	(n=1902)
Gr 12 English			0.48 (n=2869)	0.57 (n=1596)
Gr 12 Afrikaans			0.54 (n=2868)	0.60 (n=1596)
Total Test 1	**0.54**	**0.61**		0.63 (n=1582)
Total Test 2	**0.61**	**0.63**	**0.76**	

There was a high correlation between Test 1 and Test 2 scores (0.76), indicating test-retest reliability. Also, good correlations were obtained between the Grade 12 English and Afrikaans marks on the one hand and the Test 1 score (0.54 and 0.61) on the other, and slightly higher ones for the correlations with the Test 2 score (0.61 and 0.63). These correlations suggest good predictive validity of both tests, using past language performance. They serve to confirm the findings on the reliability of the two tests.

Test measurements are never entirely accurate, since no test can be 100% reliable. Therefore, it is anticipated that misclassifications might occur when a cut-off score is used, so that an examinee who deserves to pass might fail, or vice versa. An estimate of the number of potential misclassifications that occurred in a test administration is a function of the overall test reliability (Cronbach's alpha), the standard error of measurement and the cut-off score. This function attempts to correlate the observed test scores to hypothetical parallel test scores. The number of students who did not deserve to fail cannot be too high (although there is no definite criterion for this), as this would affect the validity of the test. The number of misclassifications that might have occurred in the administration of Test 1 and Test 2 was obtained through the use of TiaPlus (2008) software and are displayed in Table 2.

Table 2: Potential misclassifications

	Test 1	Test 2
% Misclassified	17.6	18.2
No of persons misclassified	543	346

Taking 0.3 standard deviations from the cut-off score as an informal criterion for misclassification, 21.1% for Test 1 and 21.8% for Test 2 should have been misclassified – well beyond our results. The misclassifications in Table 1 are in line with those reported by Weideman and Van der Slik (2008:170) in a TAG study involving students from UP, Stellenbosch and NWU. They found that 16.4% (414 out of 2521) NWU students were misclassified in their study. A smaller percentage of misclassifications occurred in Test 1 than Test 2, even though the former had a smaller relative reliability coefficient. This might be explained by the fact that the cut-off score for Test 1 was 35% while that for the second test was 50%.

4.3 Explanation (construct) inference

The respective internal correlations of the test sections were calculated (Tables 3 and 4) (Kok, 2012) and measured against three specific criteria. These provide evidence of construct validity (cf. Bachman, 1990:258; Alderson et al. 2005:184). The various sections of a test are intended each to measure a different aspect of the construct, so their correlations can be expected to be low – between 0.3 and 0.5. The correlations between each section and the whole test (displayed in the last row of the tables), however, can be expected to be high (0.7 or more) (cf. Van der Walt & Steyn, 2007:148). The inclusion of the individual section score in the total score for the test inflates the correlation (Alderson et al., 2005:184), and the test sections must therefore also be correlated with the test total minus the section in question (given in the last column of the tables).

Table 3: Internal correlations Test 1

Section	1	2	3	4	5	6	Total excluding section
1							0.40
2	0.15						0.37
3	0.13	0.23					0.31
4	0.11	0.23	0.18				0.49
5	0.18	0.28	0.27	0.33			0.66
6	0.08	0.13	0.14	0.20	0.31		0.26
Total	**0.48**	**0.47**	**0.41**	**0.59**	**0.79**	**0.51**	

Table 4: Internal correlations Test 2

Section	1	2	3	4	5	6	Total excluding section
1							0.32
2	0.20						0.36
3	0.34	0.37					0.79
4	0.17	0.26	0.52				0.66
5	0.18	0.11	0.31	0.35			0.43
6	0.15	0.18	0.35	0.41	0.44		0.44
Total	0.42	0.46	0.87	0.73	0.50	0.65	

Table 3 indicates that all 15 sections meet the first criterion, with 13 lower than 0.3. Table 4 shows that 12 of the 15 sections meet the criterion, with 7 lower than 0.3. The correlations between each section and the test totals are only high for section 5 of Test 1, but in the case of Test 2, sections 3 and 4 indicate high values. As can be expected, the same pattern exists, but with lower values, when test totals excluding a section are correlated with the sections (the third criterion).

A principal component analysis also provided data on construct validity. This analysis extracts the main factors that underlie the constructs being assessed. The variation (i.e. information) explained by each factor is indicated as eigenvalues in a scree plot, where any sharp drops indicate that subsequent factors are relatively less important.

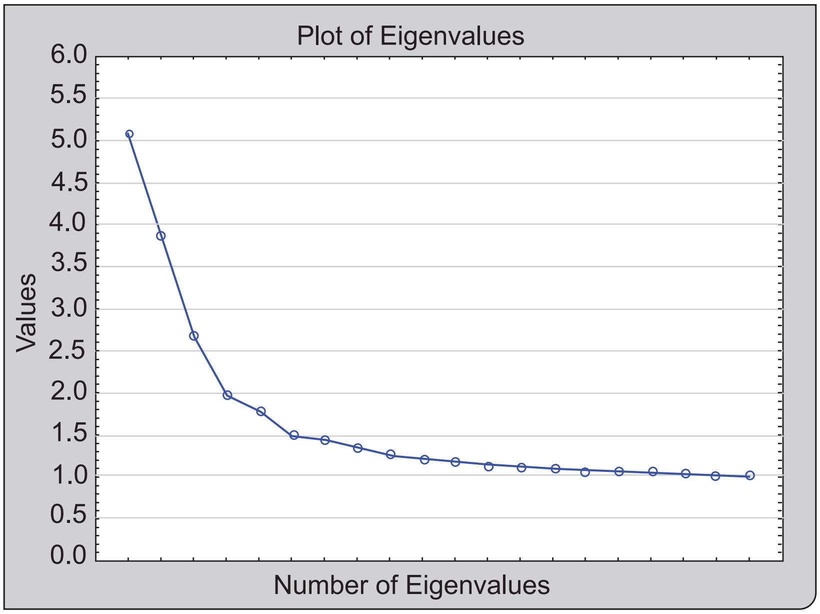


Figure 1: Scree plot test 1

The scree-plot in Figure 1 shows that for the items of Test 1 the variance of 5.1 was accounted for by the first principal component, which was only 8.9% of the total variance.

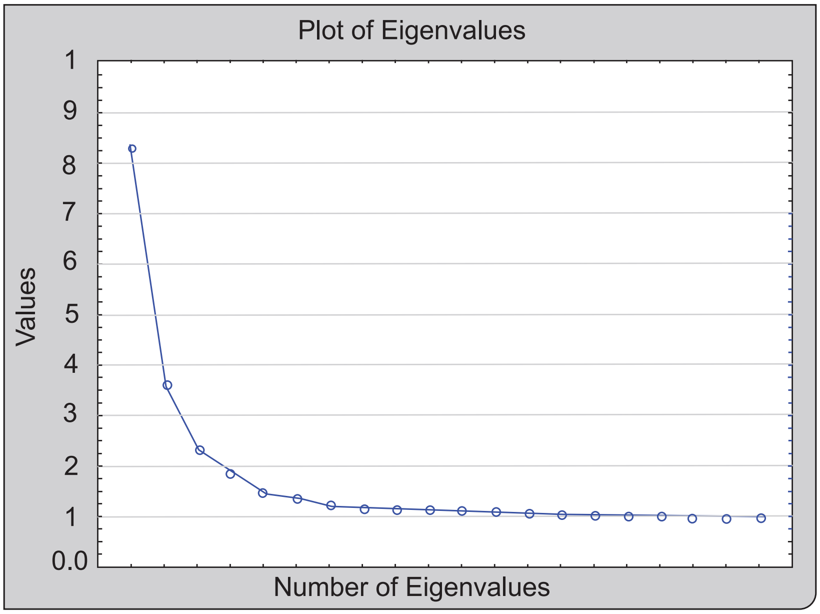


Figure 2: Scree plot Test 2

From the scree-plot (Figure 2) for Test 2, the variance of 8.3 was accounted for by the first principal component, which was only 13.2% of the total variance, slightly higher than that of Test 1. In both these plots the first component is not as dominant as one should ideally want it to be. This finding is in line with our analysis of the construct validity of another TAG test (cf. Van der Walt & Steyn, 2007).

We then performed a factor analysis, which is widely used to determine the construct validity of tests. As each test section measures specific aspects of academic literacy (cf. Weideman, Patterson & Pot 2014: 8), the construct validity of each of the different sections is of interest. If a minimum number of factors present high communalities and account for a large percentage of the variance, then the construct validity of a test section is proved (Van der Walt & Steyn, 2007:148). Principal component analysis was performed by means of the Factor procedure in SAS (SAS Institute Inc., 2011). Tables 5 and 6 (Kok, 2012) display the results obtained for the two tests. A single construct did not arise from any section of Test 1, although Sections 1 and 3 each formed two constructs. Sections 2 and 4 to 6 did not achieve construct validity, as a large number of factors were extracted, and these factors account for only a small percentage of variance. Section 1 of Test 2 has one construct that explains 61% of the variance, and thus this section is construct valid. Both Sections 2 and 5 can be divided into two constructs. Sections 3, 4, and 6 did not achieve construct validity.

Table 5: Construct validity of the sections of Test 1

TAG Placement Test			
	No of Components	Percentage Variance Explained	Communalities
Section 1	2	61	0.23 - 0.84
Section 2	3	58	0.26 - 0.54
Section 3	2	62	0.31 - 0.78
Section 4	3	35	0.16 - 0.46
Section 5	7	48	0.26 - 0.73
Section 6	3	56	0.26 - 0.93

Table 6: Construct validity of the sections of Test 2

TAG Semester Test			
	No of Components	Percentage Variance Explained	Communalities
Section 1	1	61	0.18 - 0.81
Section 2	2	45	0.21 - 0.72
Section 3	6	38	0.24 - 0.56
Section 4	2	37	0.26 - 0.49
Section 5	2	66	0.38 - 0.79
Section 6	4	61	0.45 - 0.78

Both the TAG tests had a minimal number of sections of which the construct validity was proved. We also found this in our previous analysis of a TAG test (Van der Walt & Steyn, 2007:151). It seems that some aspects that are tested are not part of a clear construct. This may be ascribed to the fact that academic literacy is a multi-faceted and multidimensional construct and very difficult to reduce to one underlying ability. In this regard, Van der Slik and Weideman (2005: 32) argue that a degree of heterogeneity in a test of academic literacy has to be tolerated as it assesses such a rich and varied construct. In addition, it is inevitable that sections of the test will overlap to some extent. Weideman, Patterson and Pot (2014:8) point out that aspects of the academic literacy construct can be assessed in more than one of the subtests.

An obvious example is the task of comprehending a text. A detailed specification of task types makes this overlap clear (cf. Weideman, Patterson & Pot, 2014:8-9).

4.4 Conclusion regarding the validity of the tests

Based on the evidence presented above, we argue that the two tests can be regarded as valid. The reliability of both is acceptable, there are similar percentages of misclassifications as in previous studies, and construct validity is as good as can be expected with regard to academic literacy and in line with our previous study (cf. Van der Walt & Steyn, 2007).

5 Comparison of the two tests

Before we could compare the results of the two tests, we first had to consider whether the academic literacy classes that the students attended influenced the results of Test 2. We suspected that this was not the case, as the initial sections of the two modules were aimed at aspects such as study methods and planning an academic essay – aspects not assessed in the tests. In order to verify this, the regression discontinuity method (Lee & Munk, 2008) was applied.

Students (n=522) who scored below the cut-off score of 35% on Test 1 were assigned to a compulsory academic literacy module AGLA111 (*Introduction to Academic Literacy*), whilst those (n=1060) who achieved a score above this proceeded to enrol in AGLA121 (*Academic Literacy*). The aim of the AGLA111 module is to develop basic academic skills, such as vocabulary and the reading and writing of academic texts. The AGLA121 module is intended for students who are not regarded as at-risk in their studies and aims to develop academic skills at a slightly more advanced level than the AGLA111 module. We did not expect either of these modules to have any marked effect on the results of Test 2.

To test if the module AGLA111 in comparison with AGLA121 had the intended treatment effect, we investigated whether the students enrolled in it (i.e. students who scored less than 35% in Test 1) performed poorer in Test 2 than those enrolled in AGLA121. As pointed out above, the allocation of students to each module is not done randomly, but via Test 1 as a placement test, which functions as a selection variable. The outcome variable is the score achieved in Test 2, and it is assumed that this variable is a continuous function of the score achieved for Test 1.

The regression discontinuity method (cf. Lee & Munk, 2008) was implemented to test the effect that attendance of AGLA111 (in comparison with AGLA121) had on the students' performance in Test 2, by fitting a linear multiple regression model, using the REG procedure of SAS (SAS Institute Inc., 2011). As predictors we used the placement test score (Test 1) (S) and the dichotomous AGLA111 vs. AGLA121 variable T. The average Grade 12 Afrikaans and English scores (Z) were used as a control variable. Also included as predictors were the interactions between T and S (T*S) and T and Z (T*Z). In order to interpret the interaction effects, S and Z had to be centred (i.e. by subtracting their average values). Different combinations of predictors resulted in a series of models. The adjusted coefficient of determination, R^2*,* was also calculated for each model. The adjusted R^2 values may be interpreted as the squared sample correlation coefficient between the outcome variable and its predicted value, adjusted for the number of variables in the model.

The desired model is one that has a high adjusted R^2 and a small number of variables with practical significance. Hence, to determine whether the variables entered into the model have practical significance, their respective squared semi-partial correlations as measures to determine the unique contributions of each predictor were evaluated. Each model with its predictors, adjusted R^2 value and squared semi-partial correlation coefficient for each predictor is displayed in Table 7 (Kok, 2012).

Table 7: Multiple linear regression model results for predicting Test 2 Score

Model	Predictors	Adj R-sq	Semi-partial sq
1	S	0.3926	0.19
	T		0.00008
2	S	0.3924	0.171
	T		0.00027
	T*S		0.00018
3	S	0.5023	0.07
	T		0.0003
	T*S		0.00002
	Z		0.11
4	S	0.5024	0.07
	T		0.0006
	Z		0.07
	T*Z		0.00012
5	S	0.5026	0.07
	T		0.0005
	Z		0.11

S: Centred Placement test score; T=1: AGLA111, T=0: AGLA121; Z: centred average Gr12 language mark; T*S: interaction T with S; T*Z: interaction T with Z.

Since the squared semi-partial correlations are very small throughout for T or its interaction with S or Z (effect size smaller than 0.01 – cf. Cohen, 1988), it can be concluded that the AGLA111 course in comparison with the AGLA121 course had no effect on the Test 2 score, when controlling for Test 1 and Grade 12 language scores. This indicates that students did not learn more from AGLA111 than from AGLA121 between the administrations of the two tests.

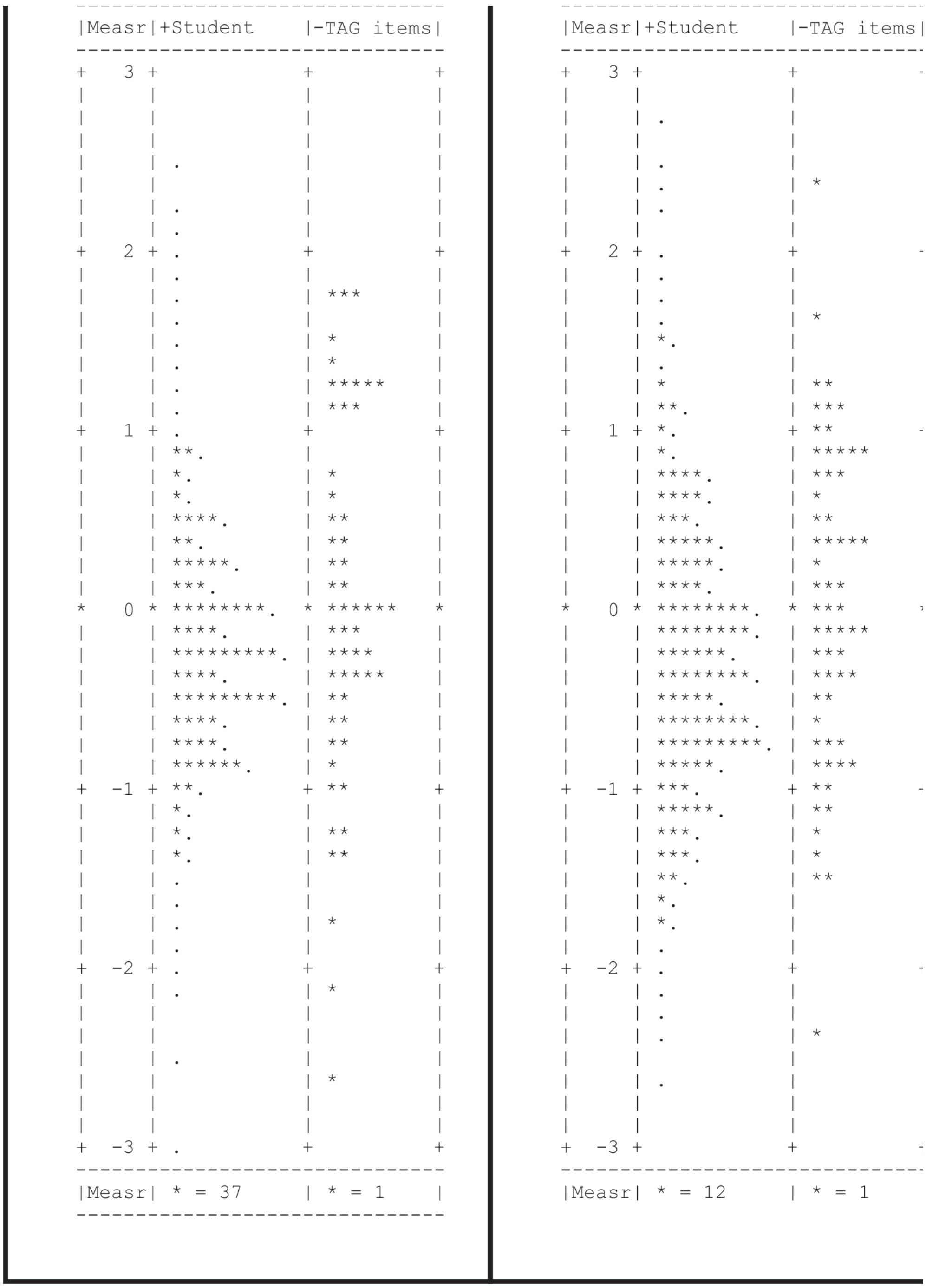
5.1 Infit mean square analysis

The raw scores on which Classical Test Theory depends to indicate the ability of candidates and test difficulty are problematic – we have no way of knowing whether the characteristics of candidate ability and item difficulty would be maintained for the candidate over different items and for items if administered to different candidates (McNamara 1996:153). Rasch analysis (cf. McNamara, 1996) enables one to move beyond raw scores to underlying ability or difficulty, expressed not as scores but as *measures*. It is more sophisticated and more complex than classical analysis. It takes the raw scores of all the candidates' responses on all the items into account in forming estimates of item difficulty, and estimates how difficult items would be for other, similar candidates. Rasch analysis thus provides information on how the abilities of test-takers and the difficulty level of test items match. There is no linear relationship between raw scores and measures; in fact, the relationship between these is generally weak. As Rasch analysis indicates underlying ability and difficulty, its function is inferential and not descriptive (Bachman, 2005:34).

A multi-faceted Rasch analysis can be done by using the WINSTEPS program (Linacre, 2008). The resultant item-ability map, as mentioned, provides estimates or predictions of test-taker ability and item difficulty, and reports estimates of probabilities of test-taker responses under the condition of item difficulty. These are expressed in terms of the relation between the ability of individual candidates and their relative chances of giving correct responses to items of given difficulty (McNamara, 1996:200). These chances are expressed in logits. The logit-scale in Table 8 ranges from +3 at the top to –3 at the bottom; the larger values indicating better test-taker abilities and more difficult items, while lower values indicate poorer test-taker abilities and easier items. Table 8 indicates that no extreme difficulties occurred in both Test 1 and Test 2 (only a very few students had extreme abilities outside the limits +3 and –3). For both tests there was no major mismatch; the estimated ability of the candidature was at the general level of difficulty of the items.

Table 8 indicates the degree of match between the model and the data. If the pattern for the individual items, allowing for normal variability, fits the overall pattern, the items show an appropriate 'fit'. If not, they are 'misfitting' or 'overfitting' items, and should be inspected or reconsidered (cf. McNamara 1996:169-175). Table 8 also indicates this match between the abilities of the students and the difficulty level of the test items.

Table 8: Item-ability maps of Test 1 (left) and Test 2 (right)



```
--------------------------------
|Measr|+Student     |-TAG items|
--------------------------------
+   3 +             +          +
|     |             |          |
|     |             |          |
|     |             |          |
|     | .           |          |
|     |             |          |
|     | .           |          |
|     | .           |          |
+   2 + .           +          +
|     | .           |          |
|     | .           | ***      |
|     | .           |          |
|     | .           | *        |
|     | .           | *        |
|     | .           | *****    |
|     | .           | ***      |
+   1 + .           +          +
|     | **.         |          |
|     | *.          | *        |
|     | *.          | *        |
|     | ****.       | **       |
|     | **.         | **       |
|     | *****.      | **       |
|     | ***.        | **       |
*   0 * ********.   * ******   *
|     | ****.       | ***      |
|     | *********.  | ****     |
|     | ****.       | *****    |
|     | *********.  | **       |
|     | ****.       | **       |
|     | ****.       | **       |
|     | ******.     | *        |
+  -1 + **.         + **       +
|     | *.          |          |
|     | *.          | **       |
|     | *.          | **       |
|     | .           |          |
|     | .           |          |
|     | .           | *        |
|     | .           |          |
+  -2 + .           +          +
|     | .           | *        |
|     |             |          |
|     |             |          |
|     | .           |          |
|     |             | *        |
|     |             |          |
|     |             |          |
+  -3 + .           +          +
--------------------------------
|Measr| * = 37      | * = 1    |
--------------------------------
```

```
--------------------------------
|Measr|+Student     |-TAG items|
--------------------------------
+   3 +             +
|     |             |
|     | .           |
|     |             |
|     | .           |
|     | .           | *
|     | .           |
|     |             |
+   2 + .           +
|     | .           |
|     | .           |
|     | .           | *
|     | *.          |
|     | .           |
|     | *           | **
|     | **.         | ***
+   1 + *.          + **
|     | *.          | *****
|     | ****.       | ***
|     | ****.       | *
|     | ***.        | **
|     | *****.      | *****
|     | *****.      | *
|     | ****.       | ***
*   0 * ********.   * ***
|     | ********.   | *****
|     | ******.     | ***
|     | ********.   | ****
|     | *****.      | **
|     | ********.   | *
|     | *********.  | ***
|     | *****.      | ****
+  -1 + ***.        + **
|     | *****.      | **
|     | ***.        | *
|     | ***.        | *
|     | **.         | **
|     | *.          |
|     | *.          |
|     | .           |
+  -2 + .           +
|     | .           |
|     | .           |
|     | .           | *
|     |             |
|     | .           |
|     |             |
|     |             |
+  -3 +             +
--------------------------------
|Measr| * = 12      | * = 1
```

Fit statistics as infit mean square values can be used to evaluate each item. They have an expected value of 1; individual values will be above or below this according to whether the observed values show greater variation (resulting in values greater than 1) or less variation (resulting in values less than 1) (McNamara 1996:172). McNamara (1996:173) suggests criterion values in the range of 0.75 to 1.3. Values greater than 1.3 show significant misfit, i.e. lack of predictability, while values below 0.75 show significant overfit. In our test results, the infit mean square values for Test 1 range from 0.91 to 1.08, while for Test 2 they were between 0.87 and 1.10. Thus, all items were in accordance with the fitted Rasch model, and we could therefore conclude that both tests were fair and their difficulty levels acceptable.

5.2 Equipercentile measure

An equipercentile measure can be used when both forms of a test are valid and reliable and the two sets of scores are regarded to be equal (Mohandas, 2007:3). This enables one to equate the two test forms directly.

In Table 9 every 5th percentile for the distributions is displayed for the total test scores of the two tests. The distribution of total scores of Test 2 was shifted substantially to the right of that of Test 1, indicating better performance in Test 2 than Test 1. The first quartiles were 34 and 42, medians were 42 and 53, and the third quartiles were 51 and 64 respectively. This indicates that Test 2 was easier than Test 1 for the study population.

Table 9: Equipercentile equating of Test 1 and Test 2

Percentile	Test 1 Total	Test 2 Total
0	0	10
5	24	28
10	28	34
15	31	37
20	33	39
25	34	42
30	36	44
35	37	46
40	39	49
45	40	51
50	42	53
55	43	55
60	45	58
65	47	60
70	49	62
75	51	64
80	53	67

Percentile	Test 1 Total	Test 2 Total
85	55	71
90	59	74
95	65	79
100	84	96

5.3 Bland-Altman plot

Finally, a Band-Altman plot (Bland & Altman, 1999) was drawn (Figure 3). This plot displays the agreement between the two tests on individual student level. This is done by plotting the differences between the test scores per student against their mean scores.

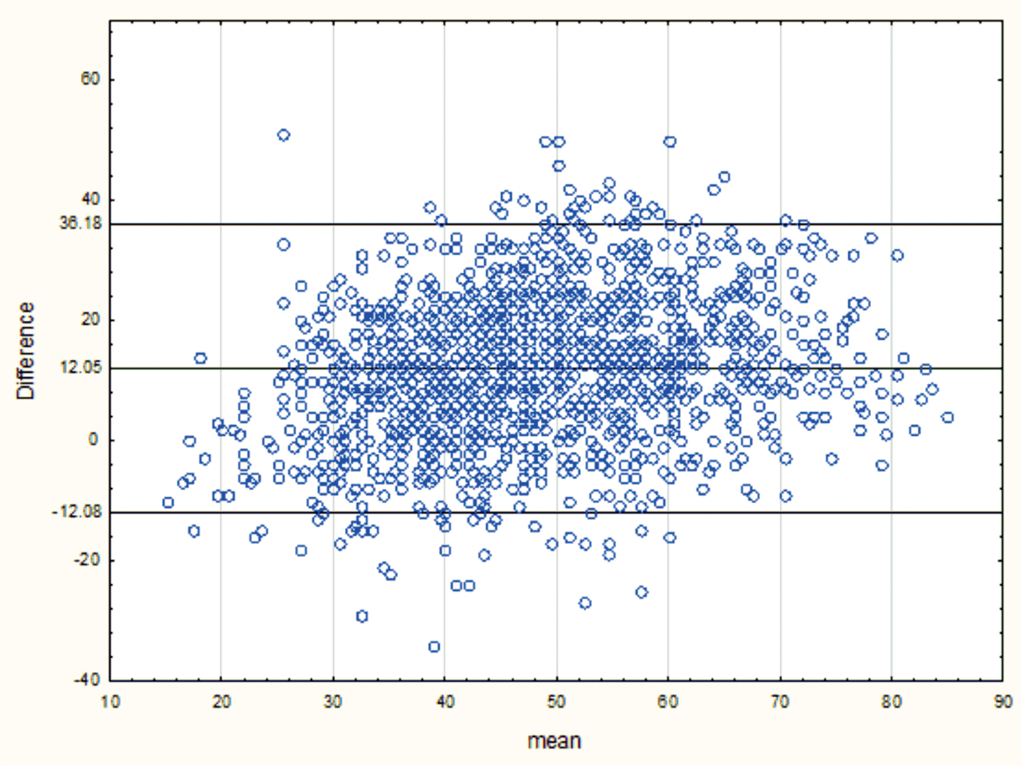


Figure 3: Bland-Altman plot

The plot shows the lack of agreement in the level of the two tests, since the mean difference line is well above zero (12.05) (the average scores for Test 1 and 2 were 42.04 and 54.09 per cent respectively). This supports the analysis of the equipercentile measure that the study population found Test 2 easier than Test 1. Also, it seems that there is no relationship between the differences and the means, where the mean scores are used as a proxy for the true score per student. This implies that differences do not increase with students' aptitude. The two outer horizontal lines on the plot give the 95% limits of agreement within which most differences between test scores will lie and also indicate relatively few extreme differences.

6 Conclusion

Many definitions of a standardised test refer to it as a test that is administered and scored in a consistent manner, with reliability and validity the essential elements that determine the quality of any test. A standardised test, however, cannot have versions at significantly different difficulty levels. TAG tests are administered under prescribed conditions and, as they consist of multiple-choice items, scoring is consistent. The question, however, is whether they consistently assess at a similar level. This is always difficult to achieve in practice, as it is virtually impossible to ensure that all items of all tests are equally difficult. But it remains a basic requirement that all versions of a standardised test should be more or less at the same difficulty level. This is hard, and equating is therefore necessary to provide information on which tests are relatively easier or more difficult than others (cf. Petersen et al., 1989).

We were able to make use of a single group of students taking two forms of the TAG test. This is the most efficient design, as student ability is directly controlled (Albano, 2011:3). The data show that both tests can be regarded as valid, based on the sampling, scoring and construct evidence obtained. Academic literacy, however, remains a multifaceted construct. The test also contains a few very short sections, which are not conducive to construct analysis. In terms of the standard of the two tests, Test 2 proved to be easier than Test 1. There may extraneous factors that contributed to this. In our 2007 study (Van der Walt & Steyn, 2007), we found that some students were unfamiliar with the format, could not cope with the demands of the test, and felt that they could not deliver their best performance. These factors may have influenced the results of Test 1. The ten-week course could also have influenced the results, and students could also have been more familiar with the test format in the second test.

Our data indicate that both tests are good and fair ones. However, there is clear evidence that the levels of the two tests were not equivalent. We do not believe that the differences in the results could be ascribed to the short period instruction the students received. We think that further research on the difficulty levels of this test series is necessary, so that they can be brought in line. However, it seems as if this will have to be achieved by means other than test design. Patterson and Weideman (2013 a & b) have suggested that the original blueprint for the test be expanded, so that the primacy of logical and analytical modes in academic discourse can be assessed more productively, but it remains to be seen if this will have an influence on the equivalence of the tests. A reliable model for the adjustment of scores can also prove to be useful, and can compensate for the difficulty to achieve equivalence at the test design stage (but only if a norm for a cut-off point can be established). At present, each set of test results continues to be treated on its own merits when the tests are administered, and specific results, conditions and requirement are taken into account when students are placed and cut-off points are established. This remains a good practice until standardisation or near-standardisation can be achieved.

Acknowledgements

We would like to thank Nandi Kok for assistance with some of the statistical analyses and Albert Weideman for comments on a draft.

References

Albano, A. 2011. Statistical equating methods. http://www.edmeasurement.net/Equating/Albano%202010%20Equate.pdf. Date of access: 15 June 2013.

Alderson, J.C., Clapham, C. & Wall, D. 2005. *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Bachman, L.F. 1990. *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L.F. 2002. Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. *Educational Measurement: Issues and Practice* 21(3):5-18.

Bland, J. M. & Altman, D. G. 1999. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* (8):135–160.

Chapelle, C.A. 2012.Validity argument for language assessment: The framework is simple ... *Language Testing* 29(1):19-27.

Chapelle, C.A., Enright, M.K. & Jamieson, J.M. 2008. *Building a validity argument for the Test of English as a Foreign Language*. New York: Routledge.

Chapelle, C.A., Enright, M.K. & Jamieson, J.M. 2010. Does an argument-based approach to validity make a difference? *Education Measurement: Issues and Practice* 29(1):3-13.

Cohen, Jacob. 1988. *Statistical Power Analysis*. 2nd ed. Academic Press: New York.

Hunter, J.E. & Schmidt, F.L. 2004. *Methods of Meta-analysis: Correcting error and bias in research findings.* 2nd ed. Sage Publications: Thousand Oaks.

Kane, M.T. 2001. Current concerns in validity theory. *Journal of Educational Measurement* 38(4):319-342.

Kane, M.T. 2006. Validation. In: Brennan, R.L. (Ed.) 2006. *Educational Measurement,* 4th ed. Praeger: American Council on Education. pp. 17-64.

Kok, Nandi. 2012. Test and Item Analysis for the TAG Test of Academic Literacy. Research paper in partial fulfilment of the requirements for the STTN611 module of Hons. B.Sc. (Statistics). Potchefstroom: NWU.

Lee, H. & Munk, T. 2008. Using regression discontinuity design for program evaluation. http://www.Amstat.org/sections/SRMS/proceedings/y2008/Files/301149.pdf. Date of access: 30 April 2013.

Linacre, John M. 2008. Winsteps Ministep. Rasch-Model Computer Programs, Version 3.66.0. Chicago, USA. http://www.winsteps.com. Date of access 15 April 2013.

Mohandas, R. 2007. Test equating. http://info.worldbank.org/etools/docs/library/117785/handout-equating.pdf. Date of access: 4 April 2013.

Petersen, N.S., Kolen, M.J. & Hoover, H.D. 1989. Scaling, norming, and equating. In: Linn, R.L. (Ed.). 1989. *Educational measurement. 3rd ed*. Washington, DC: American Council on Education.

Patterson, R. & Weideman, A. 2013a. The typicality of academic discourse and its relevance for constructs of academic literacy. *Journal for Language Teaching* 47(1): 107-123.

Patterson, R. & Weideman, A. 2013b. The refinement of a construct for tests of academic literacy. *Journal for Language Teachin*g 47(1): 125-151.

Purpura, J.E. 2010. Assessing communicative language ability: Models and their components. In: Shohamy, E. & Hornberger, N.H. (Eds.). 2010. *Language testing and assessment. Encyclopedia of language and education, Volume 7.* New York: Springer. pp. 53-68.

SAS Institute Inc. 2011. *The SAS System for Windows Release 9.3 TS Level 1M0*. SAS Institute: Cary, NC.

TiaPlus, 2008. *TiaPlus: Classical Test of Item Analysis.* Cito M & R Department: Arnheim.

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

Van der Walt, J.L. 2012. The meaning and uses of language test scores: An argument-based approach to validation. *Journal for Language Teaching* (46)2:141-156.

Van der Walt J.L. & Steyn, H.S. (jnr.) 2007. Pragmatic validation of a test of academic literacy at tertiary level. *Ensovoort* 11(2): 137-153.

Van Dyk, T.J. & Weideman, A.J. 2004. Switching constructs: On the selection of an appropriate blueprint for academic literacy assessment. *Journal for Language Teaching* 38(1):1-13.

Weideman, A.J. 2007. *Academic literacy: Prepare to learn.* Pretoria: Van Schaik.

Weideman, A., Patterson, R. & Pot, A. 2014. Construct refinement in tests of academic literacy. Paper presented at AAAL 2014 colloquium on *Exploring post-admission language assessments in universities internationally,* Portland, Oregon.

Weideman, A.J. & Van der Slik, F. 2008. The stability of test design: Measuring differences in performance across several administration of an academic literacy test. *Acta Academica* 40(1): 161-182.

Weir, C.J. 2005. *Language testing and validation.* Houndmills, Basingstoke: Palgrave Macmillan.

ABOUT THE AUTHORS

Johann L van der Walt

School of Languages
North-West University, Potchefstroom 2520

Email: Johann.VanDerWalt@nwu.ac.za

The author is emeritus professor of English at North-West University and currently manages short courses in languages at the Potchefstroom campus. His research interests include language testing, second language acquisition and language teaching.

Faans Steyn

Faculty of Natural Sciences, Statistical Consultation Services
Private Bag X6001, North-West University, Potchefstroom, 2520

Email: faans.steyn@nwu.ac.za

Prof Steyn is a statistical consultant at the North-West University on a part time basis after his retirement. Throughout his career he helped researchers at the university to plan their research and to do sound statistical analyses. He also undertook research in practical statistics to develop new methods to analyse data.