Pragmatic validation of a test of academic literacy at tertiary level

Pragmatiese validasie van ‘n toets van akademiese geletterdheid op tersiêre vlak

Johann L. van der Walt & H.S. Steyn (Jnr.)

North-West University

Summary

*Validity is a fundamental consideration in language testing. Conceptions of validity have undergone a number of changes over the past decades, and validity is now closely connected with the interpretation of test scores. Validity remains an abstract concept, however, and can only be accessed though a process of validation. This article illustrates an approach to the validation of a test by postulating a number of claims regarding an administration of an academic literacy test (*Toets van Akademiese Geletterdheidsvlakke*) and presenting various kinds of evidence to investigate these claims.*

Introduction

It is generally accepted that validity is the central concept in language assessment. The AERA/APA/NCME (1999) test standards regard it as the most fundamental consideration in developing and evaluating tests. It is a complex concept, which has undergone a number of different interpretations. At present it is generally acknowledged that validity is contextual, local and specific, pertaining to a specific use of a test, i.e. one asks whether the test is valid for *this* situation. The validity of tests is determined though the process of validation, a process of test score interpretation, before the results can be used for a particular purpose. In order to determine the validity of a test, a validation argument has to be constructed, on the basis of which it can be suggested whether the interpretations and uses of the test results are valid.

The purpose of this article is to discuss current conceptions of validity and then illustrate the process of validation by constructing a validation argument for a widely used test of academic literacy. We will propose a number of claims and illustrate methods with which to investigate these, and use the test of academic literacy levels (*Toets van Akademise Geletterdheidsvlakke*) administered at the Potchefstroom campus of North-West University in 2007 to illustrate *a posteriori* validation procedures.

Validity

The concept of validity is not a fixed one, but has undergone different interpretations over the past 50 years. Two main perspectives can be distinguished.

The first is often called the 'traditional' view, which involves the question of whether one measures what one intends to measure. It considers validity to be an inherent attribute or characteristic of a test, i.e. a test is valid if it measures what it claims to be measuring (Kelley, 1927; Cattell, 1946; Lado, 1961). Three major types of validity are identified: criterion-related validity (including concurrent and predictive validity), content-related validity, and the one introduced by Cronbach and Meehl (1955), construct validity. The traditional approach reflects a positivistic paradigm, which assumes that a psychologically real construct or attribute exists in the minds of the test takers – this implies that if something does not exist, it cannot be measured. Variations in the attribute cause variation in test scores. Validity is thus based on a causal theory of measurement (Trout, 1999). Reliability, the index of measurement consistency, is regarded as distinct from validity, and is a necessary condition for validity.

The second view evolved in the 1980s, and replaced the definition of three validities with a single unified view of validity; one which portrays construct validity as central component (Chapelle, 1999: 256), and regards content and criterion validity as aspects of construct validity (Messick, 1989: 20). Messick's (1989) paper provided a seminal although somewhat opaque exposition of this view. This view – a more naturalistic, interpretative one – is the most influential current theory of validity. It shifted validity from a property of a test to that of test score interpretations. Validity is an integrated evaluative judgement based on theoretical and empirical evidence, which supports test score interpretation and use (Messick, 1989: 13). It is seen as a unitary but multifaceted concept. Messick (1989) introduced his much-quoted progressive matrix – the types of research associated with validity, which involve the definition and validation of the construct, decisions about the individual (involving fairness and relevance) [inferences], definition of social and cultural values, and real world decisions, e.g. admission or placement [uses]. Messick's innovation was the introduction of consequential validity, i.e. the social consequences and effects of a test on an individual

and society. Test consequences thus became a central part of validity. Admission or placement decisions can have a major impact on a test taker, and therefore aspects such as affect, washback and ethics are considered part of consequential validity (cf. also Fulcher & Davidson, 2007; Madsen, 1983; Hughes, 1989; Chapelle, 1999). In addition, the test context (including the environment, such as room temperature or seating), which may introduce construct-irrelevant variance, can have an impact on the test scores (cf. Fulcher & Davidson, 2007: 25). Brown (1996: 188) also mentions administration procedures and the environment of test administration as relevant factors. Weir (2005: 51) includes test taker characteristics (e.g. emotional state, concentration, familiarity with the test task types) as factors that may influence test performance, and ultimately affect the validity of the test scores.

This incorporation of a social dimension into validity, which can be interpreted fairly broadly, has not been without controversy, as many critics argue that validity does not involve decision-based interpretations, but only descriptive interpretations. But the psychometric tradition of language testing has obscured the role and effect of language testing in society, especially its sorting and gatekeeping roles, which ultimately depend on the policies and values that underlie any test. The practice of decision-based interpretations has now become part of validity, although as yet there is no coherent theory of the social context in current validity theory.

The current interpretive paradigm thus allows a variety of data to inform test validity. In essence, validity is the truth-value of a test. But the question is: what is real or true? Truth remains a relative concept, a question of judgement, a matter of degree, subject to new or more relevant evidence. There is no such thing as an absolute answer to the validity question (Fulcher & Davidson, 2007: 18). This view allows every important test-related issue to be considered as relevant to the validity concept integrated under a single header. Validity therefore involves a chain of inferences. Any construct has to be empirically verifiable, and validity claims depend on the evidence provided to support it. Fulcher (1997) emphasises the fact that validity is a relative and local affair. The argument for local validity is not current – it was advanced by Lado (1961) and Tyler (1963) decades ago. Weideman (2006a: 83) also stresses that the social dimension is unique to each test administration. Tests are valid for a specific use, but determining

validity is an ongoing and continual process (Davies & Elder, 2005).

This second view of validity gained prominence in language testing in the 1990s, when Bachman (1990) introduced Messick's ideas to language testing research. The idea of validity as a unitary concept is now accepted by most researchers. Construct validity is generally regarded as the overarching validity concept, but there is still variation in the use of terminology and the sub-types of validity proposed. Bachman and Palmer (1996) introduced the overarching concept of test usefulness, but Bachman (2005) later returned to validity as metanarrative. Weir (2005: 14) also proposed the re-introduction of validity as superordinate category, and postulated the subcategories of context validity, theory-based validity, scoring validity and external validity. McNamara (2003: 470) points out that the social dimension of validity is now a "prime topic" in language testing debates (cf. also McNamara & Roever, 2006).

In the second view, reliability is no longer regarded as a separate quality of a test, but is part of overall validity. Weir (2005: 14) says: "... the traditional polarisation of reliability and validity … is unhelpful and reliability would be better regarded as one form of validity evidence". Most researchers still regard reliability as important, as in principle a test cannot be valid unless it is reliable (it can be reliable but invalid) (Alderson, Clapham & Wall, 2005: 187).

The second view of validity is, of course, not without its critics. One of the reasons is that it seems natural and instinctive to consider validity to be a feature of a test. Borsboom, Mellenburg and Van Heerden (2004: 3) say: "… we think that the argument that shifted the meaning of validity from tests to score interpretations is erroneous". They argue that there is no reason why one would conclude that the term validity can only be applied to test score interpretations. They propose a return to the traditional view (e.g. Kelley, 1927: 14), which states that a test is valid if it measures what it purports to measure, even though one can only validate interpretations. They argue that current accounts of validity only superficially address theories of measurement. Fulcher and Davidson (2007: 279) also ask: "Has this validity-as-interpretation mantra perhaps become over-used? If a test is typically used for the same inferential decisions, over and over again, and if there is no evidence that it is being used for the wrong decisions, could we not speak to the validity of that particular test – as a characteristic of it? Or must we

be on constant guard for misuse of all tests?"

The view of validity as interpretation is now widely accepted. But it is dependent on test results being used for the purpose for which the test is designed. Score interpretation must therefore be valid. Various factors can affect the interpretation, including external factors, as we have seen. Sufficient evidence allows a conclusion about overall test quality – its validity. It starts as local affair, with repeated use of a test for one purpose only, and ultimately one can argue that validity becomes a property of the test, i.e. that it tests what it purports to test; that it tests a property that exists and can be measured.

Validation

Validity can only be accessed through validation. Validity, in Messick's (1989) terms, remains an abstract concept, and validation is the process of operationalizing the construct of validity. It is an activity: the collection of all possible test-related activities from multiple sources. The validation process therefore involves the accumulation of evidence to support the proposed test score interpretations and uses (Lane, 1999: 1). The process is based on Kane's (1992) systematic approach to thinking through the process of validation. Kane sees validation as the construction of "an interpretative argument"; a number of inferences following each other, ending in decisions made about a candidate.

Davies and Elder (2005: 804) point out that it is not easy to operationalize Messick's (1989) and Bachman's (1990) intricate conception of validity. McNamara and Roever (2006: 33) also refer to a "decade and more of grappling" with this complex validity framework. Bachman (2005: 267), in an attempt to make test validation a manageable process, suggests the following procedure:

- Articulating a validation argument, which provides the logical framework linking test performance to an intended interpretation and use.

- Collecting relevant evidence in support of the intended interpretation and use.

Evidence collected may include what Davies and Elder (2005: 798) call the "usual

suspects" of content, construct and criterion validity, as well as reliability. But additional sources of validity evidence are also allowed (Davies & Elder, 2005: 801), mostly as part of the social dimension of testing (consequential validity interpreted broadly), such as student feedback, test consequences, ethics, social responsibility, washback, affect and impact of test scores. But there is not always a principled way of combining all the elements that can be regarded as validation evidence. Fulcher and Davidson (2007: 18) thus speak of a pragmatic approach to validity; an approach that "best explains the facts available".

The validation process involves the development of a coherent validity argument for and against proposed test score interpretations and uses. It takes the form of claims or hypotheses (with implied counterclaims) plus relevant evidence. But we must also examine potential threats to the validity of score interpretation. Kane, Crooks and Cohen (1999: 15) point out that "the most attention should be given to the weakest part of the interpretative argument because the overall argument is only as strong as its weakest link". Validation is therefore as much a process of raising doubts as of positive assertion.

What constitutes an adequate argument? Fulcher and Davidson (2007: 20) suggest the following basic criteria:

- Simplicity: explain the facts in as simple a manner as possible.
- Coherence: an argument must be in keeping with what we already know.
- Testability: the argument must allow us to make predictions about future actions or relationships between variables that we could test.
- Comprehensiveness: as little as possible must be left unexplained.

We now illustrate an approach to test validation by analysing the January 2007 TAG test administration and results. As the administration of the test, with the interpretation and use of the results, is an expensive and important exercise for its stakeholders (university management, students, parents and lecturers), the validity of the test is of major importance. Davies and Elder (2005: 802-3) report that relatively few comprehensive validation studies have been undertaken, and this article is an attempt to make a contribution in this regard.

TAG test: Validation claims and relevant evidence

The TAG test

The *Toets van Akademiese Geletterdheidsvlakke*[*] (Test of Academic Literacy Levels), or TAG, was administered to all first-year students at the Potchefstroom campus at the beginning of 2007. It was aimed at establishing whether these students possessed the necessary academic literacy skills to succeed in their content subjects. It was a medium to high stake test, as students who failed had to enrol for a course in Academic Literacy, for which parents must then pay an extra fee. TAG is a short test of 55 minutes, in multiple-choice format. The 2007 test, in Afrikaans, contained 63 items, and was administered to 2773 students.

The test content was based on a number of components that make up the construct of academic literacy. These are described in Van Dyk and Weideman (2004) and Weideman (2007). It is assumed that these components, taken together, constitute the construct 'academic literacy'. We accept this definition of the construct as valid for the purposes of this *a posteriori* analysis.

The test was divided into sections that tested the following: Placing five scrambled sentences into the correct sequence; interpreting a graph (a histogram); answering comprehension questions on a reading passage; deciding which phrase or sentence has been left out in a text; defining academic vocabulary items; identifying text types; and deciding where a word had been left out in a text, and which word had been left out.

Each inference in a validity argument is based on an assumption or claim that requires support (Bachman, 2005: 264; Lane, 1999: 1; Chapelle, 1999: 259). In our validation study of the TAG test, we constructed a number of claims and collected evidence to support these claims.

[*] The TAG test construct and results of numerous administrations of different versions of the test have been discussed in a number of publications, such as Van Dyk and Weideman (2004), Van der Slik and Weideman (2005), Weideman (2005), Weideman (2006a & b) and Weideman (2007).

Validation evidence

Claim 1: *The test is reliable and provides a consistent measure, with small variance the result of measurement error.*

A completely reliable test implies that tests scores are free from errors and can be depended on for making decisions about placement or admission. The use of internal consistency coefficients to estimate the extent of the reliability of objective-format tests is the industry standard. Reliability coefficients do not provide evidence of test quality as such: the estimated reliability is "not a feature of the test, but rather of a particular administration of the test to a given group of examinees" (Weir, 2005: 30). The internal consistency for each section and for the whole test was determined by calculating the Cronbach alpha coefficients. The results are displayed in Table 1.

	Alpha	No. of items
Section 1	0,84	5
Section 2	0,64	7
Section 3	0,70	22
Section 4	0,62	9
Section 5	0,67	5
Section 6	0,89	15
TAG	0,88	63

Table 1: Cronbach alpha coefficients

One should bear in mind that the alpha coefficient is a function of the number of items it is based on. The reliability coefficients for Sections 2, 3, 4, and 5 are below the generally accepted norm of 0,8 (Weir, 2005: 29), while the alpha for the test as a whole is very good at 0,88.

Claim 2: *The general ability of candidates matches the general level of difficulty of test items.*

An item-response analysis was performed, using the FACETS computer program (Facets for Windows Version No. 3.61.0). A two-facet Rasch model was fitted to data of the 63 items scores of the TAG for the 2773 students. In order to validate this model, the following diagnostic methods were applied (Hambleton, Swaminathan & Rogers, 1991), using the FACETS program (Facets for Windows Version No. 3.61.0) and STATISTICA (StatSoft, Inc., 2006):

- A plot of the student abilities of the more difficult items (upper half of items on the measurement scale) vs. those of the easier items (remainder of items) depicted most of the points to lie underneath the line of equality. This means that the model based on the more difficult items predicts a lower ability for a student than a model based on the easier items.

- A plot of the student abilities of the odd-numbered items vs. those of the even-numbered displays points that are evenly distributed along the equality line, with high correlation (0,83), which means that a student's abilities are similar for two sets of items selected from the test.

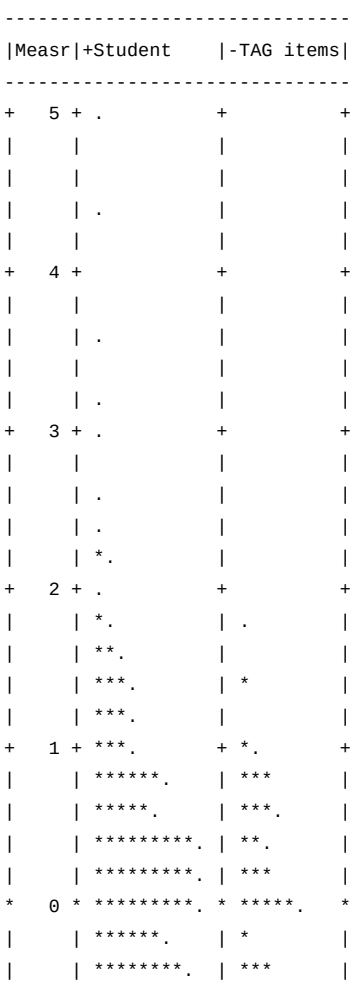
- A plot of item difficulties for higher ability students vs. those for lower ability students: with a high correlation (0,88) and even distribution of the points along the equality line, it seems that the difficulties of items are predicted similarly for the model fitted on the low and high ability groups of students.

- Comparison of the distributions of the standardised residuals for the data with that of data simulated from the fitted model: since the histograms for the two distributions are very similar, the model fitted simulated data in the same manner as that of the original data.

We therefore conclude that, on the whole, there was an appropriate fit of the model on the data.

The item-ability map in Figure 1 displays the distributions of the students' abilities and the TAG item difficulties, both relative to a logit-scale measure. This measure varies from +5 at the top to –5 at the bottom, the larger values indicating better student abilities and more difficult items, while lower values indicate poorer student abilities and easier items. The map provides estimates of person ability and item difficulty. These are expressed in terms of the relation between the ability of individual candidates and their relative chances of giving a correct response to items of given difficulty; the chances being expressed as logits (McNamara, 1996: 200). This map allows comparison of candidate ability and item difficulty.

From this display it is clear that no extreme difficulties occurred, while only a very few students had extreme abilities outside the limits -3 and 3. There was no significant mismatch; the ability of the candidature was at the general level of difficulty of the items, and there was a good fit between student ability and item difficulty.



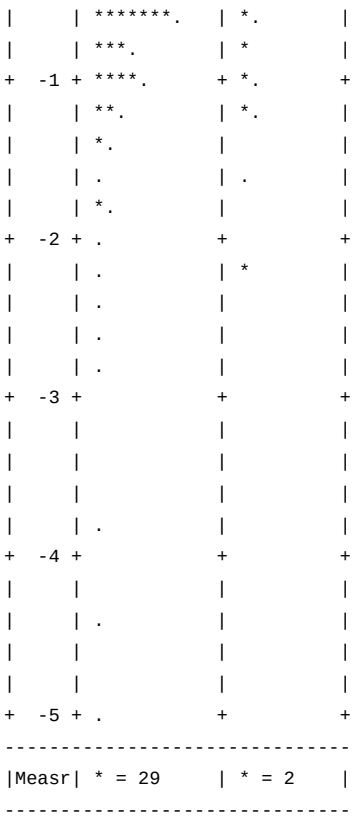


Figure 1: Item-ability map

Claim 3: *Infit mean square values of test items fall within an acceptable range.*

By means of item-response analysis, a Rasch model that summarises the observed patterning throughout the set of relations between candidates and test items was fitted. Here we wish to consider the extent to which the pattern of responses observed for individual items conforms to and reinforces the general pattern in the model, or goes against it (McNamara, 1996: 169). If the pattern for the individual items, allowing for normal variability, fits the overall pattern, the items show appropriate 'fit'. If not, they are 'misfitting' or 'overfitting' items (McNamara, 1996: 169-175). The fit statistics in Table 2 give the difficulty levels of items as measured on the logit-scale, ordered from the most difficult item (no. 21 with measure 1,77) to the easiest one (item 1 with measure –2,27), together with the fit statistic 'infit mean square'.

Observed % correct	Measure	Infit Mean square	TAG item
19.8	1.77	1.02	21
25.5	1.39	0.99	46
30.9	1.08	1.02	45
32.3	1	1	61
32.6	0.99	1.01	50
34.9	0.87	1.01	9
37.3	0.74	0.99	20
37.3	0.74	0.99	32
37.5	0.73	1	63
38.2	0.7	0.97	17
38.2	0.7	0.99	41
38.7	0.68	0.99	60
39.9	0.62	1	16
40.3	0.6	1.01	37
41.7	0.53	1	23
41.8	0.53	1	26
41.8	0.52	1.02	56
41.9	0.52	0.97	49
42.8	0.47	1	57
43.4	0.45	1	62
43.5	0.44	0.98	42
43.5	0.44	0.99	15
45.1	0.37	0.98	35
46.7	0.29	0.97	34
47.9	0.23	1.02	8
50.3	0.12	1.04	3
50.5	0.11	1.01	40
50.6	0.11	1	59
50.6	0.11	1.01	27
51.8	0.05	0.99	18
52.3	0.03	0.98	19
52.7	0.01	1	2
53.0	-0.01	1	11
53.0	-0.01	0.99	24
53.3	-0.02	1	47
53.5	-0.03	0.99	4
53.6	-0.03	1	58
54.4	-0.07	0.99	31
54.6	-0.08	1.02	44
54.9	-0.09	1.01	10
57.2	-0.2	1	22
58.6	-0.27	1.02	14
59.5	-0.31	1.01	43
61.0	-0.39	1	25
62.1	-0.44	1	55
62.1	-0.44	0.99	12
63.0	-0.48	0.99	54
63.1	-0.49	1	30
63.4	-0.51	1.02	51
65.6	-0.61	1.01	48
65.8	-0.62	0.99	29
70.0	-0.85	1	53
70.6	-0.88	1.01	5
71.2	-0.91	1.01	28
71.5	-0.93	0.98	13
72.7	-1	0.99	52
74.6	-1.11	1	7
75.3	-1.15	1.02	39
75.7	-1.18	1.02	6
83.1	-1.69	1	36
88.5	-2.18	0.99	38
89.4	-2.27	1	1

Table 2: Fit statistics

McNamara (1996: 172) points out that infit statistics are informative as they focus on the degree of fit in the most typical observations in the model. He states that infit mean square values 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 that values in the range of 0,75 to 1,3 are acceptable. Values greater than 1,3 show significant misfit, i.e. lack of predictability,

while values below 0,75 show significant overfit.

Since the infit mean squares have values that vary between 0,97 and 1,04, all items seem to be in accordance with the fitted Rasch model.

Claim 4: *The internal correlations of the different test sections satisfy specific criteria.*
Bachman (1990: 258) states that patterns of correlations among item scores and overall test scores provide evidence of construct validity. Alderson et al. (2005: 184) indicate that an internal correlation study can be used in order to examine this. They point out that the reason for having different test sections is that they all measure something different and therefore contribute to the overall picture of the attribute. They expect these correlations to be fairly low – possibly in the order 0,3 to 0,5. If two sections (components) correlate very highly with each other (e.g. 0,9), one might wonder whether the two sections are testing different attributes, or whether they are testing essentially the same thing. The correlations between each section and the whole test, on the other hand, might be expected to be higher – possibly around 0,7 or more – since the overall score is taken to be a more general measure of the attribute than each individual section score. Alderson et al. (2005: 184) add: "Obviously if the individual component score is included in the total score for the test, then the correlation will be partly between the test component and itself, which will artificially inflate the correlation. For this reason it is common in internal correlation studies to correlate the test components with the test total *minus* the component in question". Three different types of correlation coefficients can be identified, each with its own criterion:

- The correlation coefficients between each pair of subtests (C1). These correlations should be fairly low, from 0,3 to 0,5 (cf. Hughes, 1989: 160; Alderson et al., 2005: 184; Ito, 2005).
- The correlation coefficients between each subtest and whole test (C2). These correlations should be 0,7 and more (cf. Alderson et al., 2005: 184; Ito, 2005).
- The correlation coefficients between each subtest and the whole test minus the subtest (C3). These should be lower than those between each subtest and the whole test, i.e. C2 > C3 (cf. Ito, 2005).

The correlational pattern that results is indicated in Table 3.

Section	1	2	3	4	5	6	Total
1							0,294
2	0,207						0,387
3	0,297	0,378					0,514
4	0,214	0,312	0,495				0,573
5	0,141	0,161	0,304	0,358			0,441
6	0,146	0,195	0,282	0,388	0,391		0,406
Total	0,447	0,533	0,763	0,697	0,557	0,737	

Table 3: Correlations

The table indicates the following:

- C1 (shaded areas): only eight of the fifteen correlations meet the criterion, with seven lower than 0,3.
- C2 (last row): only three of the six correlations meet the criterion.
- C3 (last column): all correlations meet the criterion.

It must be noted that one weakness inherent in the correlational approach to construct validation is that it only evaluates the relevance of those performance criteria that are already included, and that it cannot identify others that are relevant to the construct but which have been omitted (Moritoshi, 2002: 11)

Claim 5: *Each section of the TAG test displays construct validity.*

Bachman (1990: 259) indicates that factor analysis is extensively employed in construct validation studies. The construct validity of each section of the test can be verified by means of principal component analysis, a factor analytic model that reduces data and extracts the principal components or factors that underlie the construct being assessed. The results obtained by means of STATISTICA (Statsoft Inc., 2006) are displayed in Table 4.

Section	No. of components	Percentage variance explained	Communalities
1	1	61	0,18 – 0,81
2	2	71	0,25 – 0,68
3	6	39	0,18 – 0,70
4	2	37	0,20 – 0,59
5	1	43	0,37 – 0,48
6	2	50	0,15 – 0,66

Table 4: Factor analysis results

Only Sections 1 and 5 formed one construct, while sections 2 and 6 can be split up into two constructs. In this regard, a principal factor analysis with an oblique (OBLIMIN) rotation was performed. For section 2 the resultant factor pattern gives items 6 - 9, 12 as first sub-factor, while items 10 and 11 belong to the second sub-factor, with a correlation of -0,57 between the sub-factors. In the case of section 6, the sub-factors are formed by items 56 to 63 and 49 to 55 respectively, with a correlation of -0,64 between sub-factors. Sections 3 and 4 are not construct valid. To be construct valid, as few as possible factors that explain the maximum percentage of variance are required, with communalities as high as possible. (In Section 4, only 37% of the variation is explained by the two factors.)

Claim 6: *The first principal component dominates the whole test.*

In Figure 2, the eigenvalues for each principal component are plotted in a simple scree plot line (cf. Cattell, 1966) in their sizes as a diminishing series. The percentage variance explained by the first component was 13,5% relative to 6,2%, 3,9%, 3,0% of the second, third and fourth components respectively. The first component is therefore not as dominant as one should ideally wish.

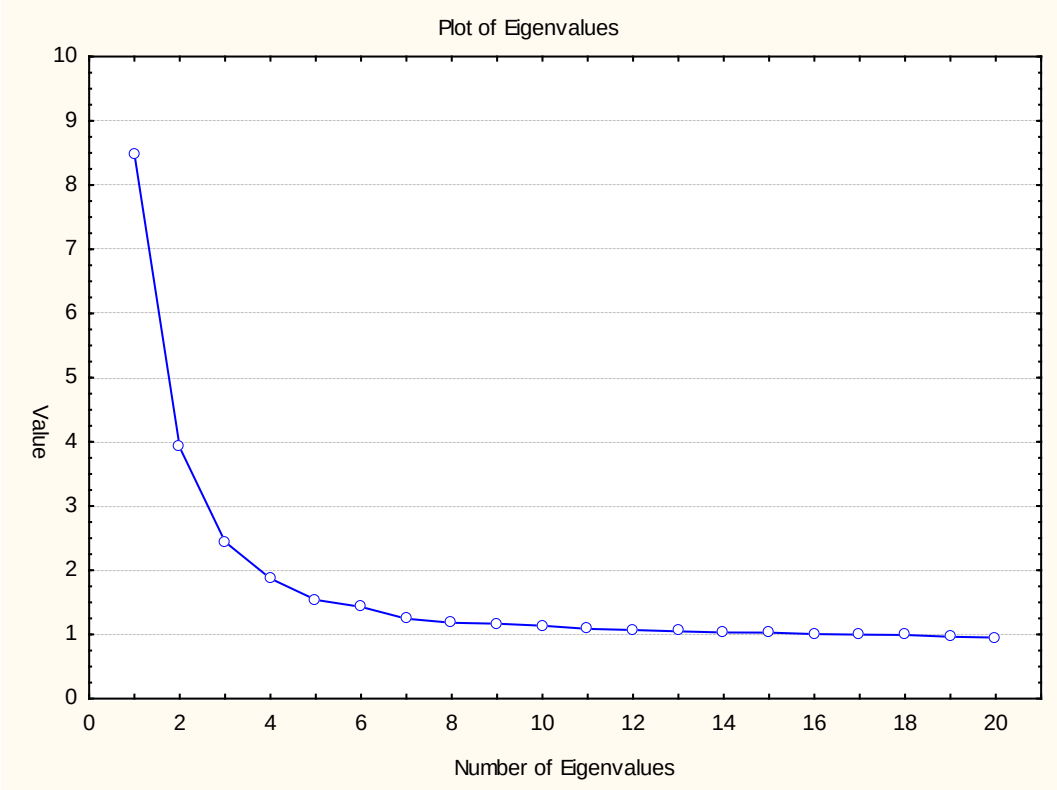


Figure 2: Scree plot

Claim 7: *Test takers were familiar with the demands of the test.*

A questionnaire aimed at establishing how first-year students experienced the test was completed by a group of 754 students from eight faculties during the first week of classes. Because of practical constraints, this group was not randomly selected, but formed an availability sample. The findings cannot be generalised to the whole population, but provide an indication of how students felt about the whole test procedure. Claims 7 to 10 are based on the findings of this questionnaire.

Weir (2005: 54) points out that candidates should be familiar with the task type before sitting the test proper, as the degree of the candidate's familiarity with the demands of a test may affect the way the task is dealt with. He states that specimen past papers and clear specifications should help difficulties in this respect, and that an exemplification of tasks and procedures should be readily available.

Students were informed about the TAG test when they enrolled at the university, and full information was provided in a first-year guide that was sent to them in the November preceding their arrival towards the end of January. No information was provided on the university website. Table 5 indicates where students first heard about the test.

	Percentage
When I enrolled	20
In the first-year guide	42
From the residence committee	13
From friends	16
Other	7
No response	2

Table 5: Where did you first hear about the test?

Most of these students read about the test in the first-year guide. The guide informed students that they could look at a specimen test on the website of the University of Pretoria. Only 11 percent looked at this example test. Of these, only 7 percent did the practice test. Seventy-five percent who did not look at the specimen test indicated that they would have liked to do so. It therefore seems as if the test is not as transparent as it could be at the Potchefstroom campus, and that much more could be done to inform students about the test and its format.

Claim 8: *The circumstances under which the test is administered are conducive to valid scores being obtained.*

The test was written on the fourth day after students' arrival on campus. From the second day, they are subjected to an Orientation and Information week under the supervision of the residence committees. This involves a full programme, and students are kept busy the whole day (and part of the night!).

We asked students whether they felt they could deliver their best performance in

the test: 65 percent indicated that they could not. The reasons they indicated were as follows: tired (40%), sleepy (21%), stressed (25%), ill (7%) and other reasons (7%). Sixty-three percent of the students went to bed after midnight, while 57 percent reported that they had to get up before 6:00 on the morning of the test. It seems as if the circumstances in which the test is administered are not ideal, and these could affect the validity of the test results.

Claim 9: *Students experience the test as relevant to their studies.*

Forty-five percent of the students had a good idea of the purpose of the test; thirty-four percent had a vague idea, while 21 percent did not know what the purpose was. The test results were indicated in terms of codes ranging from 1 to 5, indicating to what extent they were at risk in their studies. Students whose results fell within the codes 1 to 3 must enrol for the course in Academic Literacy. Twenty-nine percent of the respondents had to take the course. Sixty-nine percent of those who had to take the course declared that they would do so under protest.

Claim 10: *Students found the test experience agreeable.*

Eleven percent of the students said they found *all* the questions clear, 68% knew what they had to do at *most* questions, 19% at *some* of the questions, while 1% did not understand *any* question (1% gave no reply). We also asked the students what they thought about the length of the test. Fourteen percent could finish the test, 39% could finish but had to work fast, while 47% reported that they could not finish the test. This is somewhat disconcerting, as most responses should ideally be in middle category. Fifty-eight percent felt that the test was too long to finish within the allocated time.

Conclusion

A problem in the current conception of validation studies is the balance between theoretical rigour and manageability – this remains a challenge for validation research. As a result, a pragmatic stance is often adopted, as was the case here. The framework employed in this validation study includes statistical procedures, based on both classical and item-response theory, as well as a social dimension, in the form of student feedback.

A variety of evidence was collected, providing a profile of the test results and its administration. It is obvious that validity is a multifaceted concept, and that many factors together play a role in the validation of a test. Each claim that is formulated contributes to an aspect of the validity of the test. The aim of the article was to illustrate a method of doing a validation study, and it is clear that the conclusions made depend on a judgement and interpretation of the results obtained.

It must be stressed that there is no such thing as a perfect test, as is no test situation. This is why provision is usually made for the misclassification of candidates when tests results are analysed. But the TAG test investigated here performs very well. It is used for a specific, clearly defined purpose. Its reliability is good, and there is a good fit between student ability and item difficulty. The internal correlations are probably as good as can be expected. More than one underlying trait (or factor) was extracted, and the first was not as dominant as expected. This may be due to the fact that academic literacy is a rich and multidimensional construct (cf. Van der Slik & Weideman, 2005). It is also clear that much more can be done to improve the administration of the test, such as informing students better, explaining the relevance of the test, and ensuring that students can deliver their best performance in the test. The problems students had with the length of the test also warrant further investigation.

We investigated the validity of only one administration of the TAG test here. If it proves to be valid in most respects over number of administrations, and remains to be used for its specific purpose, the test itself can come to be regarded as a valid instrument, in terms of the traditional interpretation of validity. Validation thus remains an ongoing process.

Bibliography

AERA/APA/NCME. See American Educational Research Association.

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

American Educational Research Association, American Psychological Association,

National Council for Measurement in Education. 1999. *Standards for educational and psychological testing*. Washington, DC: Author.

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

Bachman, L.F. 2005. *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L.F. & Palmer, A. 1996. *Language testing in practice*. Oxford: Oxford University Press.

Borsboom, D., Mellenbergh, G.J. & Van Heerden J. 2004. The concept of validity. *Psychological review* 111 (4): 1061-1071.

Brown, J.D. 1996. *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Cattell, R.B. 1946. *Description and measurement of personality*. New York: World Book Company.

Cattell, R.B. 1966. The scree test for the number of factors. *Multivariate behavioural research* 1: 245-276.

Chapelle, C.A. 1999. Validity in language assessment. *Annual review of applied linguistics* 19: 254-272.

Cronbach, C.J. & Meehl, P.E. 1955. Construct validity in psychological tests. *Psychological bulletin* 52: 281-302.

Davies, A. & Elder, C. 2005. Validity and validation in language testing. In Hinkel, E. (ed.) 2005: 795-813.

Facets for Windows Version No. 3.61.0. Copyright © 1987-2006, John M. Linacre.

Fulcher, G. 1997. An English language placement test: Issues in reliability and validity. *Language testing* 14(2): 113-138.

Fulcher, G. & Davidson, F. 2007. *Language testing and assessment: an advanced resource book.* Abingdon, Oxon: Routledge.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. 1991. *Fundamentals of item response theory, volume 2*. Newbury Park: Sage Publications.

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

Hughes, A. 1989. *Testing for language teachers*. Cambridge: Cambridge University Press.

Ito, A. 2005. A validation study on the English language test in a Japanese nationwide university entrance examination. *Asian EFL journal* 7(2), Article 6. [Online]. Available http://www.asian-efl-journal.com/June_05_ai.pdf . Accessed 20 July 2007.

Kane, M.T., Crooks, T. & Cohen, A. 1999. Validating measures of performance. *Educational measurement: issues and practice* 18(2): 5-17.

Kane, M.T. 1992. An argument-based approach to validity. *Psychological bulletin* 112: 527-535.

Kelley. T.L. 1927. *Interpretation of educational measurements*. New York: Macmillan.

Lado, R. 1961. *Language testing: the construction and use of foreign language tests*. New York: McGraw-Hill.

Lane, S. 1999. *Validity evidence for assessments*. [Online]. Available http://nciea.org/publications/ValidityEvidence_Lane99.pdf. Accessed 21 May 2007.

Lepota, B. & Geldenhuys, J. 2005. *25 years of applied linguistics in Southern Africa: Themes and trends in Southern African linguistics.* Pretoria: University of Pretoria.

Linn, R.L. (ed.) 1989. *Educational measurement*. New York: Macmillan.

Madsen, H.S. 1983. *Techniques in testing*. Oxford: Oxford University Press.

McNamara. T. F. 1996. *Measuring second language performance*. London: Longman.

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

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

Messick, S. 1989. Validity. In Linn, R.L. (ed.) 1989: 13-103.

Moritoshi, T.P. 2002. *Validation of the test of English conversation proficiency*. MA dissertation. University of Birmingham. [Online]
Available http://www.cels.bham.ac.uk/resources/essays/MoritoshiDiss.pdf. Accessed 12 May 2007.

Newton-Smith, W.H. (ed.) 1999. *A companion to the philosophy of science*. Oxford: Blackwell.

StatSoft, Inc. (2006). STATISTICA (data analysis software system), version 7.1. www.statsoft.com.

Trout, J.D. 1999. Measurement. In Newton-Smith (ed.) 1999: 265-276.

Tyler, L. 1963. *Tests and measurement*. Englewood Cliffs, NJ: Prentice-Hall.

Van Dyk, T. & 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.

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

Weideman, A.J. 2005. Integrity and accountability in applied linguistics. In Lepota, B. & Geldenhuys, J. (eds.) 2005: 174-197.

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

Weideman, A.J. 2006b. Assessing academic literacy in a task-based approach. *Language matters* 37(1): 81-101.

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

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