Stellenbosch Papers in Linguistics, Vol. 38, 2008, 191-204

The validation of language tests

Johann L van der Walt and Faans Steyn (Jr.)

North-West University, Private bag X6001, Potchefstroom 2520
johann.vanderwalt@nwu.ac.za
faans.steyn@nwu.ac.za

1. Introduction

Validity and reliability have traditionally been regarded as the basic criteria which any language test should satisfy. The issue of reliability has received most of the attention, because it is easily determined through statistical analysis, especially in tests consisting of dichotomously scored items. However, validity has recently emerged as a most important consideration in developing and evaluating language tests.

The validity of a test can only be established through a process of validation, and this must ideally be done before the results can be used for any particular purpose. In order to carry out such validation, a validation study has to be undertaken, on the basis of which one can arrive at a conclusion as to whether the interpretations and uses of the test results are valid. The purpose of this paper is to discuss the process of validation of a language test. Relatively few validation exercises have been undertaken in the past (Davies and Elder 2005: 802-3), but validation is necessary because of the major impact which test results can have on the many stakeholders involved. Because tests are expensive and important exercises for all stakeholders, the validity of any test is an issue of major concern.

2. The concept of validity

Validity is often regarded as involving the question of whether a test measures what it intends to measure. This implies that validity is an inherent attribute or characteristic of a test, and assumes that a psychologically real construct or attribute exists in the minds of the test-takers – if something does not exist, it cannot be measured. Variations in the attribute cause variation in test scores. However, this is not the prevailing view of validity. The most influential conception of validity at present sees it not as a characteristic of a test, but as a property of the interpretation of test scores.

The current conception of validity derives from the work of the American psychologist Samuel Messick in the 1970s and 1980s at the American educational and measurement organisation, Educational Testing Services, culminating in his seminal 1989 chapter on validity in Linn's *Educational measurement*. Messick (1989) shifted perspectives on validity from a property of a test to that of test score interpretation, and validity is now closely associated with the interpretation of test scores. Messick (1989: 13) states that "(v)alidity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment". In contrast to the traditional definition of validity, Messick (1989: 20) presents a unified model, and portrays construct validity as a central underlying component, with content and criterion validity as aspects of construct validity. Messick (1989: 20) introduces his much-quoted progressive matrix (cf. table 1), which consists of a four-way classification of validity, described by the source of justification of the testing, which includes the consideration of either evidence or consequence, or both, and the function or outcome of the testing, which includes either test interpretation or use, or both. Messick (1989) does not regard these categories as watertight, but considers the boundaries to be "fuzzy". His framework has proven to be extremely influential in testing, but it is complex and demanding, and difficult to operationalise (cf. Shepard 1993; McNamara and Roever 2006).

Source of justification	Function of testing	
	Test interpretation	**Test use**
Evidential basis	Construct validity	Construct validity + Relevance/utility
Consequential basis	Construct validity + Value implications	Construct validity + Relevance/utility + Social consequences

Table 1. Messick's (1989) facets of validity

Messick (1989) particularly stresses the importance of the social consequences and effects of a test on an individual and society (now commonly known as "consequential validity"). Aspects such as washback (the effect of a test on teaching) and ethics are part of validity, as are administration procedures and the test environment. Weir (2005: 51) includes test-taker characteristics such as emotional state, concentration and familiarity with the test task as factors that may influence performance in a test, and thus affect the validity of test scores. The incorporation of a social dimension into validity has focused attention on the sorting and gate-keeping roles of a test, aspects which are often ignored.

Messick (1989) argues that all test constructs and score interpretation involve questions of values. All test result interpretation reflects the values to which the assessor adheres. It is assumed that scores reflect the "truth" about a test-taker. Validity thus represents the truth-value of a test. The task of the test developer or rater/interlocutor is to elicit the communicative ability of the test-taker, and then to arrive at an objective assessment of that ability – something that is ultimately elusive, because 'truth' remains a relative concept. This implies that there is no absolute answer to the validity question, and means that every important test-related issue is relevant to the validity of the test. Validity therefore involves a chain of inferences, based on the evidence provided to support each inference.

Bachman (1990) adopted Messick's conceptualisation of validity in his influential *Fundamental considerations in language testing*. As a result, the idea of validity as a unitary concept is now widely accepted in language testing. Weir (2005: 14), for example, uses validity as a superordinate category, which includes the subcategories of context validity, theory-based validity, scoring validity and external validity. Shaw and Weir (2007: 7) refer to context validity, cognitive validity, and scoring validity as making up construct validity. Reliability is no longer a separate characteristic of a test; it is an aspect of validity: the term

"scoring validity" is commonly used as a superordinate term for all aspects relating to reliability (Weir 2005: 14). In addition, the social dimensions of validity have become a major concern in language testing as a result of Messick's insistence on taking into account the consequences of tests, and these are now generally regarded as forming part of the validity concept.

Messick has been criticised for shifting validity to test score inferences, mainly because it seems natural and instinctive to consider validity to be a feature of a test. Borsboom, Mellenburg and Van Heerden (2004: 3), for example, argue that there is no reason why one would conclude that the term "validity" can only be applied to test score interpretation, and argue that current accounts of validity only superficially address theories of measurement. Fulcher and Davidson (2007: 279) suggest that the "validity-as-interpretation mantra" may have become over-used, and ask "(i)f 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?" Despite these concerns, the view of validity as interpretation is now common. But it is important to remember that test results should only be used for the purpose for which the test is designed. Validity is contextual and specific, pertaining to a specific use of a test. With the repeated use of a test for its designated purpose, we feel that one can ultimately argue that validity becomes a property of the test.

3. The validation process

Validity is an abstract concept, and validation is the process of operationalising it. *A posteriori* validation involves two stages: (i) the test results and the evidence they produce of candidate ability, and (ii) the empirical validation of the procedures by which the test judgements were reached. The latter stage involves the collection of all possible test-related evidence from multiple sources, and is the one illustrated in the present paper. Kane (1992: 527) calls this the construction of "an interpretative argument". The evidence that may be collected includes construct validity, content validity and criterion validity and reliability indexes. In addition, social aspects such as test-taker feedback, test consequences, test ethics, social responsibility, washback and the impact of test scores may also be included.

4. Collecting validity evidence

We now discuss some of the types of evidence that can be collected in the test validation process. Before a validation study is carried out, however, the researcher has to formulate a number of hypotheses pertaining to the test results. Examples of hypotheses about the validity of the test are: that it has construct validity, that its scores are reliable, that it predicts future performance accurately, or that test-takers were familiar with the format of the test. The hypothesis is stated, and then evidence is collected to support or reject the hypothesis.

From the various sources of evidence available, we focus here on Classical Test Theory, Rasch measurement and factor analysis for the analysis of test scores, and test-taker feedback and test administration as social aspects that may influence the validity of the test scores. We illustrate our discussion by referring, where appropriate, to evidence we collected for a validation study of a test of academic literacy – the TAG test – administered at the Potchefstroom Campus of North-West University in January 2007 (cf. Van der Walt and Steyn 2007 for a detailed discussion). This test consisted of 63 multiple-choice items, testing various aspects of the academic literacy of first-year university students, and was used for placement purposes to determine which students should follow a course in academic literacy.

4.1 Classical test theory

Classical test theory stems from the psychometric tradition in testing and usually involves both item analysis and the overall reliability of the test. Classical item analysis determines item facility and item discrimination, and indicates the adequacy of the test items. Item facility is the proportion of candidates who answered the item correctly, while item discrimination is the extent to which the item discriminates between low and high performers. Item analysis is used to determine which items are not at the required difficulty level and do not discriminate well among test-takers. These items should be revised or omitted from the test. Item analysis is part of test development and evaluation and is particularly useful in multiple-choice tests.

If a test is reliable, its scores are free from error and can be depended upon for making decisions about the test-takers. Determining internal consistency coefficients to estimate the reliability of tests is the standard procedure. However, Weir (2005) warns that 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). If a test consists of a number of sections with dichotomously scored items, the internal consistency for each section and for the whole test can be determined by calculating the Cronbach alpha coefficients. The alpha coefficient is a function of the number of items on which it is based (i.e., it increases with the number of items). Weir (2005: 29) points out that the generally accepted norm in language testing is 0,8. The reliability coefficient for the test we administered was 0,88, while the coefficient for the various sections ranged from 0,62 to 0,89.

The construct validity of a test can be investigated by determining the correlation between its various sections (Alderson, Clapham and Wall 2005: 184). Each section measures something different and therefore contributes to the overall picture of a test-taker's ability. Alderson et al. (2005: 184) state that the correlations between sections should be fairly low – in the order of 0,3 to 0,5. If two sections correlate very highly with each other (e.g., 0,8 – 0,9), one might wonder whether they 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, should be higher (possibly around 0,7 or more) since the overall score is taken to be a more general measure of the attribute than is each individual section score. Alderson et al. (2005: 184) add that "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 coefficient can be identified, each with its own criterion:

- C1: The correlation coefficients between each pair of subtests. These correlations should be fairly low, from 0,3 to 0,5 (cf. Alderson et al. 2005: 184; Ito 2005).
- C2: The correlation coefficients between each subtest and whole test. These correlations should be 0,7 and more (cf. Alderson et al. 2005: 184; Ito 2005).

- C3: The correlation coefficients between each subtest and the whole test minus the subtest. These should be lower than those between each subtest and the whole test, i.e., C3 < C2 (cf. Ito 2005).

In our validation study, only eight of the fifteen correlations met criterion C1, with seven lower than 0,3. Only three of the six correlations met criterion C2, and all correlations met criterion C3.

4.2 Rasch measurement

A central concern in language testing is that of the equivalence of tests. Ideally, the various versions of a test (e.g., from one administration to the next) should be equal. This is, however, very difficult to achieve in practice. One way of doing this is to keep the number of passes constant. But this is an inherently unfair practice (e.g., if a particular group is strong, a good candidate may receive a lower score than he would have in a weaker group). The difficulty of a test depends on the ability of a particular group of candidates – an unstable value. The raw scores upon which classical test theory depends to indicate ability of candidates and test difficulty are therefore problematic – one has 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 measurement enables one to move beyond raw scores to underlying ability or difficulty, expressed not as scores but as measures. Rasch measurement 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. There is no linear relationship between raw scores and measures; in fact, the relationship between these is generally weak. Because Rasch measurement indicates underlying ability and difficulty, its function is inferential and not descriptive (Bachman 2005: 34), and this is why it is so useful. McNamara (1996) provides the following explanation of Rasch measurement:

> It enables estimates of a candidate's underlying ability to be made by analysing the candidate's performance on a set of items, after allowance has been made for the difficulty of the items and how well they were matched to

> the candidate's ability level. Thus the ability estimates (known as *measures*) are not simply dependent on the items that were taken; we have avoided the trap of assuming that ability is transparently visible from raw scores. Similarly, the underlying difficulty of items can be estimated from the responses of a set of candidates, by taking into account the ability of the candidates and the degree to which there was a match between the ability of the trial group and the difficult of the items. Central to this approach is the way in which candidate ability is related to item difficulty: this is done by estimating from the data the chances of a candidate of a given ability achieving a certain score on an item of given difficulty.
>
> (McNamara 1996: 152)

Rasch measurement provides information on how the abilities of test-takers and the difficulty level of test items match. The measurement can be done by using the FACETS program (Linacre 2006) to perform a multi-faceted Rasch analysis. A resultant item-ability map 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 units known as "logits". The logit-scale in figure 1 varies from +3 at the top to –3 at the bottom; the larger values indicating better test-taker abilities and more difficult items, whereas lower values indicate poorer test-taker abilities and easier items. Figure 1 indicates that no extreme difficulties occurred in our validation exercise (only a very few students had extreme abilities outside the limits +3 and –3). There was no significant mismatch; the estimated ability of the candidature was at the general level of difficulty of the items, and there was a good fit between test-taker ability and item difficulty.

```
Measure   Ability       Difficulty
+   3 + .           +          +
|     |             |          |
|     | .           |          |
|     | .           |          |
|     | *.          |          |
+   2 + .           +          +
|     | *.          | .        |
|     | **.         |          |
|     | ***.        | *        |
|     | ***.        |          |
+   1 + ***.        + *.       +
|     | ******.     | ***      |
|     | *****.      | ***.     |
|     | *********.  | **.      |
|     | *********.  | ***      |
*   0 * *********.  * *****.   *
|     | ******.     | *        |
|     | ********.   | ***      |
|     | *******.    | *.       |
|     | ***.        | *        |
+  -1 + ****.       + *.       +
|     | **.         | *.       |
|     | *.          |          |
|     | .           | .        |
|     | *.          |          |
+  -2 + .           +          +
|     | .           | *        |
|     | .           |          |
|     | .           |          |
|     | .           |          |
+  -3 +             +          +
```

Figure 1. Example of an item-ability map (Van der Walt and Steyn 2007: 146)

In addition to the summary map in figure 1, Rasch measurement also provides detailed information for each item and test-taker. Table 2, for example, provides information on estimated item difficulty in terms of the likely challenge items present to test-takers, as well as what is known as "fit statistics". The extract from our validation data in table 2 gives the difficulty levels of items as measured on the logit-scale, ordered from the most difficult item (no. 21 with difficulty measure 1,77) to the easiest one (no. 1 with difficulty measure –2,27), together with the fit statistic "infit mean square". Fit statistics indicate 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 appropriate "fit". If not, they are "misfitting" or "overfitting" items, and should be inspected and reconsidered (cf. McNamara 1996: 169-175).

Test item	Underlying difficulty measure	Infit mean square
21	1,77	1,02
46	1,39	0,99
45	1,08	1,02
61	1	1
50	0,99	1,01
….	….	….
39	-1,15	1,02
6	-1,18	1,02
36	-1,69	1
38	-2,18	0,99
1	-2,27	1

Table 2. Extract from a table indicating item difficulty measures (Van der Walt and Steyn 2007: 148)

McNamara (1996: 172) points out that infit statistics indicate the degree of fit in the most typical observations in the model. 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 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, i.e. less variation than might normally be expected. In our test results, the infit mean square values varied between 0,97 and 1,04. Thus, all items were in accordance with the fitted Rasch model, and we could therefore conclude that the test was a fair one, and there was no need to adjust the difficulty level of the test.

4.3 Factor analysis

Factor analysis is often employed in construct validation studies. The aim is to establish whether a test measures the postulated factors. Factor analysis combines two or more variables into a single factor, and is done to see whether the various sections of, for example, an academic literacy test, which tests vocabulary, cohesion, comprehension, etc. in different sections, can be shown to test a single underlying factor, namely academic literacy. It isolates the underlying factors or components that explain the data. If all attributes are related, reduction is possible.

There are two approaches to factor analysis: principal component analysis and common factor analysis. The merits of each of these methods are hotly debated, but as Velicer and Jackson

(1990) point out, there is little basis to prefer either component analysis or factor analysis in practice, as the choice of method is "not a decision that will greatly affect empirical results or substantive conclusions" (Velicer and Jackson 1990: 21).

We performed a principal component analysis on the six sections of our test, using the STATISTICA program (StatSoft Inc. 2006). Both Sections 1 and 5 had only one factor or construct that explained a relatively high percentage of variation (cf. Van der Walt and Steyn 2007: 150). The communalities of these sections lay between 0,18 and 0,81. Two of the other sections had two factors, which could have been caused by underlying sub-constructs, especially as the construct of academic literacy seems to be a multifaceted one. This can be investigated further by means of an oblique rotation. The remaining two sections were not construct valid.

4.4 Feedback from test-takers

Questionnaires and individual and group interviews can be employed to obtain feedback from test-takers on their perceptions of the test and their test experience. In recent years, qualitative methods (e.g., verbal protocol analysis) have also been increasingly used in addition to quantitative methods in order to examine aspects of test process as well as product.

We investigated the degree to which our test process was transparent. Weir (2005: 54) points out that candidates should be familiar with the task type before sitting for the test proper, as the degree of the candidate's familiarity with the demands of a test may affect the way in which the task is dealt with. Weir (2005) states that specimen past papers and clear specifications should minimise difficulties in this respect, and that an exemplification of tasks and procedures should be readily available. One can therefore ask test-takers questions relating to these issues. The feedback we received from a post-test questionnaire revealed that our test was not very transparent; only a handful of test-takers looked at a specimen test, and many were misinformed about the purpose of the test. Almost half of the test-takers reported that they found the test to be too long. Students who failed the test had to enrol for a course in academic literacy, and about two-thirds of them reported that they were unhappy about having to do the course (Van der Walt and Steyn 2007: 152).

4.5 Evidence from test administration

The administration of the test can be scrutinised and evaluated, as this can have a direct bearing on the validity of the scores obtained by test-takers. For example, after our test many test-takers reported that they had been tired and sleepy during the test, and two-thirds felt that they could not deliver their best performance.

5. Arriving at a validity argument

From all of the evidence accumulated, a validity argument must now be constructed. In this regard, Fulcher (1997: 113) reminds us that that validity is not only a relative but also a local affair, as each test administration is unique (cf. also Weideman 2006: 83).

One of the problems one faces in constructing a validity argument is that it is difficult to combine all the elements that can be regarded as validation evidence in a principled manner – this can sometimes amount to an *ad hoc* collection of evidence. One of the reasons for this is that frameworks for validity tend to be complex, and it is also difficult to integrate social issues into validity theory in a coherent manner (cf. McNamara and Roever 2006). In addition, validity issues often involve local or national politics and policies (McNamara 2006: 36). As a result, validity arguments often become a pragmatic affair in practice, and amount to an approach that "best explains the facts available" (Fulcher and Davidson 2007: 18).

Test developers will agree that all tests or test situations have flaws, and one must therefore consider both strengths and weaknesses of any test. Our validation results (cf. Van der Walt and Steyn 2007) indicated that the test we investigated could confidently be used to decide which first-year students had to follow a course in academic literacy. It was used for a specific, clearly defined purpose (ability in academic literacy), its reliability was good (cf. section 4.1), and there was a good fit between student ability and item difficulty (cf. section 4.2). The internal correlations were probably as good as could be expected (cf. section 4.1). More than one underlying component or factor was extracted, which did not explain a high percentage of the total variance (cf. section 4.3). As Van der Slik and Weideman (2005) point out, academic literacy is a rich and multidimensional construct, and this may be the reason why we did not find one underlying trait. It was also clear to us that much more could be done to improve the administration of the test, such as making the process more transparent, and

ensuring that students were well-rested before the test (cf. sections 4.4 and 4.5).

6. Conclusion

Validity is a multifaceted concept, and many facets together play a role in the validity of a test. One of the problems in validation studies is the difficulty of achieving a balance between theoretical rigour and manageability. As a result, a pragmatic stance is often adopted, as was the case in our test validation procedure (cf. Van der Walt and Steyn 2007: 153). The framework we illustrate here includes statistical procedures, based on both classical and item-response theory, as well as social aspects, in the form of student feedback. The conclusions that the researcher arrives at depend on his/her judgement of the evidence collected, and provide an interpretation that is relative and local. The validation of any given test can therefore not be a one-off exercise, but should be a continual one.

References

Alderson, J.C., C. Clapham and D. Wall. 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. 2005. *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

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

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

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

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

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

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

Linacre, J.M. 2006. Facets for Windows Version No. 3.61.0. Copyright © 1987-2006. www.winsteps.com.

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

McNamara, T.F. 2006. Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly* 3(1): 31-51.

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

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

Shaw, S. and C.J. Weir. 2007. *Examining writing: Research and practice in assessing second language writing. Studies in language testing 26*. Cambridge: Cambridge University Press.

Shepard, L. 1993. Evaluating test validity. *Review of Research in Education* 19(1): 405-450.

StatSoft, Inc. 2006. STATISTICA, version 7.1. www.statsoft.com.

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

Van der Walt, J.L. and H.S. Steyn. 2007. Pragmatic validation of a test of academic literacy at tertiary level. *Ensovoort* 11(2): 140-155.

Velicer, W.F. and D.N. Jackson. 1990. Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate behavioural research* 25(1): 1-28.

Weideman, A.J. 2006. Transparency and accountability in applied linguistics. *Southern African Linguistics and Applied Language Studies* 24(1): 71-86.

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