Johann L. van der Walt

http://dx.doi.org/10.4314/jlt.v46i2.9

The meaning and uses of language test scores: An argument-based approach to validation

ABSTRACT In this article, I discuss how the quality of language tests can be determined by means of a validation process. In the past, the quality of language tests was often determined by examining their reliability, content validity and reflection of real-world tasks. There have also been attempts to define the language ability construct, but this has proved to be a divisive issue. Attempts at validation were often unsystematic and *ad hoc*, reflecting a "toolkit" approach. Recent work in validation suggests an argument-based approach, which focuses on both the interpretation and uses of test scores. One of the main proponents of this approach is Michael T. Kane. I outline and assess his approach to validation and discuss and evaluate the most common inferences in language testing, such as sampling, scoring, generalization, extrapolation, explanation and utilization/decision-making. This approach allows for a systematic approach to the evaluation of tests, but requires further refinement in language testing and assessment.

Keywords: validity, validation, argument-based approach, test interpretation, test usefulness, test inferences

1. Introduction

Albert Weideman has played a seminal role in language testing and assessment, in particular with the tests of academic literacy that he has pioneered and developed. He has always been concerned with the evaluation of the quality of tests, and it is clear from his scholarly work that test providers should be able to provide appropriate answers to any questions that stakeholders ask about tests. This implies that we should constantly evaluate our tests and ensure that they are valid and reliable, and thus have quality, acceptability and meaning. I would like to address these issues in this article.

Validity has been called the "central concept" in testing and assessment (Fulcher & Davidson, 2007: 3), and it can only be accessed via validation. O'Sullivan and Weir (2011) in fact state that language testing equals validation. We must therefore be able to articulate and justify our interpretation and uses of test scores. When we assess, the scores themselves are not the object of interest; we are rather interested in what the scores mean and how we use them. Weideman (2009: 242) reminds us that on their own scores are meaningless. Testing and assessment involves an interpretation of the scores (i.e. an explanation of their meaning and implications), which entails the process of making inferences about a learner's language ability on the basis of performance in a test. We use the inferences to make decisions about a learner. The issue, of course, is that the meanings and uses we attach to a score must be valid (cf. Chapelle & Brindley, 2010).

In this article I briefly discuss the evolution of language testing over the past five decades and the accompanying changes in the conceptions of validity and validation in the quest for demonstrable quality. I then focus in particular on Michael T. Kane's argument-based approach to validation, and assess how his approach balances both score interpretation and score uses in the validation of language tests. In so doing, I refer to and evaluate the common inferences in language testing.

2. Structuralist-psychometric testing

We can argue that the modern era in language testing dates from the 1950s and 1960s, with Lado (1961) as the first author in modern second language assessment. This was the era of discrete testing, statistical analysis and scientific justification. It was based on a specific view of the construct of language ability, a skills-and-elements model consisting of three elements of language knowledge, viz. phonology, structure and lexicon, which were tested in each of the four skills.

Lado (1961) referred to the most widely-used forms of validity evidence of the time, viz. criterion-related and content validity, to demonstrate test quality. A test was useful, i.e. valid, if it did the job it was employed to do (Cureton, 1951: 621). Lado (1961: 30) said that the validity of a test could be determined only indirectly, i.e. if the scores correlated highly with another valid criterion. The criterion model was used for prediction and placement, when there was agreement between performance on an assessment and performance on a task (Cureton, 1951: 623). The content model was based on using a sample of some type of performance to draw conclusions about the level of skill in that type of performance. Validity, framed in a realist philosophy of science, was regarded as a quality of a test. Reliability was a prerequisite for validity, and was regarded as the most important criterion for a language test.

3. Communicative language testing

As it was not possible to extrapolate the criteria of structural mastery in a discrete-point test to performance in using the language for communicative purposes, communicative language testing developed at the end of the 1970s and the early 1980s as a reaction against attaching too much importance to reliability. It advocated a direct approach to testing. The major criterion for language assessment was that it should be authentic.

Porter (1983:194) stated that "[i]t is the whole thrust of communicative testing that content validity should take first place in test design". A test also had to have face validity, and to some extent predictive validity (Brindley, 1986: 14; Fulcher, 2000: 484). Validity lay in the fact that test-tasks and real-world activities were aligned. A test was valid if it reflected a real-world task. The construct as such was not considered in any detail, other than an admission that there was no general notion of communicative proficiency, but only proficiency in particular activities (Porter, 1983: 192). In some cases, such as in Brendan Carroll's (1982) *Testing Communicative Performance*, validity was not referred to at all. Reliability was considered less important by those who showed a general antipathy to statistical analysis and language testing research.

Morrow (1979: 150) distinguished between communicative language testing and performance testing, and admitted the link between the two. A communicative test was criterion-referenced and showed whether a test-taker could perform a set of specified activities. Asking the question: "What can the candidate do?", however, implies a performance-based test.

4. Performance testing

The Communicative Language Testing campaign was matched by a greater emphasis on performance tests, which also emphasised the assessment of the practical demonstration of language skills (Brindley, 1986: 6) (in part also because of a reaction against the previous over-emphasis on reliability). Performance tests are not new; they have been in use for more than 50 years, especially in the assessment of language for specific purposes, but have become increasingly prominent over the past two decades. Performance-based testing can be characterised in terms of tasks. Wigglesworth (2010: 111) defines performance testing as follows: "… tasks are designed to measure learners' productive language skills through performances which allow candidates to demonstrate the kinds of language skills that may be required in a real world context". Wigglesworth (2010: 119) adds: "Task-based testing is attractive as an assessment option because its goal is to elicit language samples which measure the breadth of linguistic ability in candidates, and because it aims to elicit samples of communicative language (language in use) through tasks which replicate the kind of activities which candidates are likely to encounter in the real world". Douglas (2000: 19) points out that a specific language test is one "in which test content and methods are derived from an analysis of a specific purposes target language use situation, so that test tasks and content are authentically representative of tasks in the target situation", particularly those found in university study and the workplace. However, task-based testing is also used in general educational contexts to evaluate language learning, and include compositions tasks, oral interview tasks, listening tasks and so on.

A key question in performance assessment is to decide on what to assess. In the first instance, it can be tested whether the task itself has been completed, and then the language used is only a means for achieving the task requirements. Answers are judged as appropriate or not appropriate (Chapelle, Enright & Jamieson, 2008: 4). A second approach is to test language use and merely use tasks to serve as the medium to elicit the language (Wigglesworth, 2010: 113). These two forms of assessment are reflected in McNamara's (1996: 43) distinction between strong and weak forms of second language performance testing. These forms amount to the basic question: can a test-taker perform a task, and how does he perform it (with regard to fluency, accuracy and complexity)?

The performance testing approach has been a significant development in language assessment, and is influential and widespread. It is, however, complex and expensive. It is also vulnerable to attack on grounds of poor reliability. A performance-based test still requires validation, i.e. justifying inferences from test performance, but its focus is on content validity, in particular on sampling from the domain of interest (to ensure content relevance and content coverage) and the development of scoring criteria. McNamara (1996: 16) admits that construct validation remains a requirement for a performance test, but seems to find the answer to this problem in the conflation of content and construct validity.

5. Construct validity

While there was little detailed discussion of validity in language testing in the 1980s (Chapelle, 1999: 256), the notion of construct validity was "rediscovered" in educational measurement. This concept developed out of personality testing and was introduced by Cronbach and Meehl (1955) in 1955 after an investigation by the American Psychological Association Committee on Psychological Tests of the qualities of a test. A concept such as *ego strength*, for example, has no criterion or content measure, but only a theory that sketches the presumed nature of the trait (cf. Cronbach, 1971: 463). Originally, Cronbach and Meehl (1955) proposed construct validity as an addition or alternative to criterion and content validity, but by the late 1970s, it seemed that the choice of validity type was often arbitrary and *ad hoc*, depending mainly on the availability of data (Kane, 2012: 7). The various kinds if validity were used as a kind of "toolkit", with only loose criteria for the selection of tools (Kane, 2001: 331). There was a tendency to look at the available evidence, and only then decide on the claims to be formulated. This was a rather unsystematic way of validating a test.

In order to counter this tendency, the construct model was proposed as a general model for a unified framework for validity. The 1985 *Standards for Educational and Psychological Testing* replaced the three validities of criterion, content and construct with a single unified view of validity, with construct validity as central (Chapelle, 1999: 256). It was argued that, from a scientific point of view, construct validity was the whole of validity (Kane, 2012: 7). In Messick's (1989) seminal paper he underscored the position that validity should be regarded as a unitary concept. Kane (2001: 324) points out that the "construct-validity model came to be seen, not as one kind of validity, but as a general approach to validity that includes all evidence for validity, including content and criterion evidence, reliability, and the wide range of methods associated with theory testing". (Reliability was by now regarded as part of validity, and remains an important although somewhat independent form of validity evidence.)

The need for an extended analysis, the need for a clear statement of the proposed interpretation, and the need to evaluate the interpretation became the basic methodological principles in the evaluation of tests and assessments (Kane, 2001: 324). This approach emphasises the ubiquitous role of inferences and assumptions in our interpretations of test scores (Kane, 2001: 325).

The unitary concept of validity, propagated by Messick (1989), has been very influential and informs many professional standards and codes for assessment, but has not provided clear guidelines for the validation of tests and is not easy to implement in practice – hence the work undertaken by scholars such as Kane (e.g. 2001, 2006, 2012) and Weideman (e.g. 2009, 2011) to achieve conceptual clarity.

6. Interpretation and usefulness of test scores

In their discussion of construct validity, Cronbach and Meehl (1955) focus on what assessment scores meant and implied. They argued that we do not validate a test, but "a principle for making inferences" (Cronbach & Meehl, 1955: 297). Their focus was on the interpretation and meaning of scores. This approach gradually became the dominant one in educational measurement towards the end of the 20th century, and the previous focus on test usefulness (that a test measures what it says it measures and achieves its goals) was to an extent sidelined. In a sense, the balance between score interpretation and score usefulness was disturbed by this development.

Messick (1989, 1994) redressed this imbalance by emphasising both interpretation and consequences resulting from uses. He included social values and consequences in his conception of construct validity, and the notion of consequential validity was formulated. So, in addition to the concept of validity being broadened beyond that of a test-internal measure, it now included consideration of the social, educational and political contexts and consequences of tests (Hall, Smith & Wicaksono, 2011: 211). Validity, defined in terms of usefulness, necessarily includes some consideration of consequences. Authors, however, differ on the relevance of consequences in validity. Borsboom, Mellenbergh and Van Heerden (2004), for example, insist on the meaning-based interpretation of scores only. Eventually a strong consensus emerged that the question: "Does this test measure what it is intended to measure?" (a question on test score usefulness, e.g. that a score predicts some criterion) does not have one single answer, and that it should rather be rephrased as "What is the evidence that supports particular interpretations and uses of scores on this test?" (Alderson & Banerjee, 2002: 79). We thus see a gradual shift from talking about the validity of a test (as a measure of an ability) to talking about the development and validation of the proposed interpretation and use of the scores (Kane, 2012: 7). Kane (2001: 324) states: "It is not the test or the test score that is validated, but a proposed interpretation of the score".

7. The language ability construct

Messick's (1989) reconceptualisation of validity put the construct of a test – what we are trying to measure – in the centre of focus (Alderson & Banerjee, 2002: 80). This required that test designers consider anew what language ability was. It required a construct theory, i.e. a theory of language ability. There was (and still is) strong support in the literature that a language proficiency construct should underlie all language tests (cf. Alderson, Clapham & Wall, 1995: 17; Weir 2005: 18). Various theoretical models of a language proficiency construct have been proposed over the past five decades. These include those of Canale and Swain (1980), Bachman (1990) and Bachman and Palmer (1996).

The issue of a language ability construct has been a divisive one in language assessment. While there is agreement that language ability is multi-componential for both classroom and large-scale testing (i.e. it cannot be limited to a single trait such as grammatical or lexical knowledge), and that any definition must take into account its context of use, there is no consensus on the specific components that constitute language ability (cf. Alderson & Banerjee, 2002: 80; Chapelle, Enright & Jameson, 2010: 4; Purpura, 2010: 55). In their search for a construct for the new TOEFL, Chapelle et al. (2010: 4) found it difficult to base their validity argument on a single theory of language proficiency, as “no agreement exists concerning a single best way to define constructs of language proficiency to serve as a defensible basis for score interpretation”. We have the same problem when we consider individual skills. Knoch (2009: 73), for example, points out that there is no theory currently available that can by itself serve as a basis for the design of a rating scale for writing. Chapelle et al. (2010: 4) point out that this problem is not unique to construct definition in language testing. Psychological theories are also often too vague to motivate psychometric models (cf. Borsboom, 2006: 437). Kane (2001: 325) also points out that educational and social sciences have little solid theory, which can make it difficult to apply construct validity.

8. Two validity frameworks

In summary, we have two validity frameworks:

- Performance testing has led to a view of assessment as task-based, where the domain of interest is the point of departure (cf. Messick, 1994).
- The construct-based approach has led to a view of assessment as competency-centred, in which a theoretical construct underlies score interpretation (cf. Messick, 1994; Weideman 2009, 2011).

In view of this, the question is: How can a test best be evaluated and validated, and its meaningfulness and defensibility described?

9. An argument-based approach to validation

The validation of tests must be a rational process that sets out the justification of interpretations and uses of scores. It is an empirical process, as evidence must be collected to support specific and local score interpretations (cf. Van der Walt & Steyn, 2007: 142). Validity can therefore be seen as a process of argument (Fulcher & Davidson, 2007: 159). This approach was first mooted by Cronbach (1988: 4), who stated that “validation of a test or test use is evaluation ... and I invite you to think of ‘validity argument’ rather than ‘validation research’”.

Kane’s (e.g. 2001, 2006, 2012) argument-based approach offers a possible solution to the problem of validity frameworks and practical validation. He has operationalised the abstract Messick model and proposed what he regards as a transparent and usable argument-based process. Kane focuses on both the meaning of assessment scores and the consequences of their use, and attempts to make validation more accessible than it has been before.

Kane’s approach does not require a language ability theory or construct *per se*, although it does not disregard applied linguistic discussion of such constructs completely. There is no strict requirement of a formal theory (Kane, 2001: 327). His approach requires an explicit statement of the proposed interpretation and uses of scores – an *interpretive argument* – followed by a *validity argument* that evaluates the interpretive argument (Kane, 2006: 23). The interpretive argument, instead of the construct, forms the basis of score interpretation. Kane (2006: 23) states that an interpretive argument lays out “the network of inferences and assumptions leading from the observed performances to the conclusions and decisions based on the performance”. It involves the collection of evidence in support of the proposed interpretations. The validity argument entails a critical evaluation of the proposed interpretations.

10. Designing a language test

When we design a test, we have certain interpretations and uses in mind. We thus have a specific purpose in mind – for example, to place students in a course. We decide on the approach we shall adopt and then develop the measurement procedure. Kane (2006: 25) suggests that we start by developing the proposed interpretation, i.e. the interpretive argument. If the interpretation and test process correspond from the outset, it will contribute greatly to the validity of the testing procedure and thus its quality. In terms of Kane's (2001: 330; 2006: 25-26) approach, the design of a test would involve the following steps:

- Outline an interpretive argument.
- Develop a test plan.
- Develop the test.
- Evaluate the inferences and assumptions you make throughout the test development process. This is an iterative process, and you continue until there is a proper fit between the test and the interpretive argument.

Following these steps can ensure the quality of a test. They will involve piloting the test to ensure that time limits are appropriate and items are reliable and function properly, for example (cf. Weideman, 2009: 247).

Kane (2006: 24) provides an example of an initial interpretive argument for a placement test (Table 1). Each inference is based on a number of assumptions.

Table 1: An interpretive argument for a Placement Test (Kane, 2006: 24)

Inference	Assumptions
Scoring: From observed performance to an observed score.	The scoring rule is appropriate. The scoring rule is applied accurately and consistently.
Generalization: From observed score to universe score.	The observations made in the test are representative of the universe of observations defining the testing procedure. The sample of observations is large enough to control sampling error.
Extrapolation: From the universe score to the level of skill.	The test tasks require the competencies developed in the courses and required in subsequent courses. There are no skill-irrelevant sources of variability that would seriously bias the interpretation of scores as measures of level of skill in the competencies.
Decision: From conclusion about level of skill to placement in a specific course.	Performance in courses, beyond the initial course, depends on level of skill in the competencies developed in earlier courses in the sequence. Students with a low level of skill in the prerequisites for a course are not likely to succeed in the course. Students with a high level of skill in the competencies taught in a particular course would not benefit much from taking the course.

At some stage, the development process is complete. Now we need to adapt a more neutral and critical stance. Cronbach (1980: 103) states: “a proposition deserves some degree of trust only when it has survived serious attempts to falsify it”. Once we have what we regard as the finished product, we need to examine the plausibility of the interpretive argument. The validity argument provides an evaluation of the interpretive argument as a whole to determine if it makes sense. We can look at evidence collected at the development stage, and for a low-stakes test, this exercise may be sufficient. But we can also collect new evidence if necessary. For high-stakes tests, a more extensive evaluation may be called for, especially if the interpretation makes very ambitious claims and has to convince a large and varied audience. If the test or examination procedure survives a significant challenge, it can be regarded as a valid one; one of quality.

The formulation of a validity argument is based on Toulmin’s (2003) practical argumentation model (typical of fields such as law, literary analysis or sociology) (cf. also Weideman, 2009: 242). An argument consists of claims, which are the interpretations that we want to make about what a test taker knows or can do (Bachman, 2005a: 9). Any claim is made on the basis of data and warrants (generally held principles). The data consist of information on which the claim is based (Toulmin, 2003: 90), i.e. the test taker’s responses in the test, and the warrant is a proposition used to justify the inference from data to claim (Bachman, 2005a: 10). Warrants rest on assumptions, and these must have backing, which is provided by empirical data. The model makes provision for rebuttals and counterevidence, which are possible alternative explanations or counterclaims. The validity argument boils down to a summary that considers and weighs all the validation evidence.

11. Inferences in language testing

The argument-based approach sets out the reasoning involved in the interpretation of scores and the process involved in drawing conclusions and making decisions. Kane (2006: 24) refers to four major inferences that are commonly found in test-score interpretations, viz. scoring, generalization, extrapolation, and utilization/decision. In another version, he includes an explanation inference (Kane, 2001: 330). Chapelle et al. (2008, 2010) and Chapelle (2012) also include a sampling inference.

Chapelle et al. (2008: 14) argue that we should begin with a description of the targeted domain, as it is crucial in language assessment to specify the domain of interest. Bachman (2002:15) also stresses the importance of making an analytic analysis of the tasks in the target domain. This specification links performance in the target domain to observations of performance in the test domain (Chapelle et al., 2008: 14). A sample has to be drawn from the tasks in the domain. This selection has to be justified as being representative of the tasks in the test domain. Chapelle (2012: 22) suggests that a *sampling* inference should be the first inference in the interpretive argument. A domain may consist of tasks such as responding to a written assignment at university. This requires the ability to write on assigned topics and write assignments, reports, summaries, research articles, research proposals, dissertations, theses, submission letters, conference abstracts, funding applications, and so on. A sample specification may include an argumentative essay task and a report in a test. At school level, where we are mainly concerned with achievement testing, the domain is specified in the curriculum document, which often also contains the assessment standards. The testing of writing often includes essays, letters, faxes, agendas, memos and so on, and a sample may include an essay and a letter in a test. Such content-based claims have been criticised as being subjective and having a confirmatory basis. However, Kane (2012: 6) states that these aspects "don't bother [him] too much", as much evidence in the evaluation of proposed interpretations and uses of assessment scores includes evidence that is not objective or quantitative and tend to have a confirmatory basis. Weideman (2009: 242) also refers to a subjective component of validity.

The nature and the boundaries of the domain are crucial in the interpretation of language test scores. Bachman (2002: 15) stresses the importance of a principled specification of tasks in the domain, and points out that tasks in real life are both extremely complex and diverse and that they are subject to great variation depending on a range of factors, and argues that it is very difficult to use tasks in a test situation to predict performance in real life. Others, such as Wu and Stansfield (2001), are less sceptical, but nevertheless argue that the authenticity of tasks needs to be verified. It is thus important that any sampling should be clearly stated and carefully supported, as it impacts on the relevance of the content and coverage of the test.

The *scoring* or *evaluation* inference involves connecting a test performance to a score. It is based on the assumptions about the appropriateness and consistency of the scoring procedures (e.g. that a rating scale is valid and reliable) and the conditions under which the performance was obtained. It starts out with an observation, for example, that a candidate makes a number of grammatical errors in an academic essay, does not include and introduction or conclusion, and does not refer to any sources. Scoring is an inference from an observation of this performance to a score: It leads to an observed score of, say, 3 out of 10, together with scores on the other writing tasks in the test, which give a total of, say, 35%.

Kane regards the scoring inference as a simple and straightforward one, but I am not so sure that this is the case in language testing, especially in the assessment of performances where a rating scale or scoring rubric is used. The assessment relies on the subjective judgement of a rater. Raters are notoriously unreliable and mark allocation often shows great variation, not only among individual markers but at different times with the same individual. Valid rating scales – such as the one designed by Hattingh (2009) for the assessment of Grade 12 ESL essays – and intensive training are essential if evaluation is to be reliable. Content validity and reliability remain important criteria in the scoring inference. There should be no construct-irrelevant factors (i.e. performance should not be influenced by irrelevant factors) and no task- or construct under-representation.

Generalization has received little attention in language assessment. It relates the observed score on a specific measure to a universe score, or a score that might be obtained from performances in tasks similar to those in the assessment, assigned by other raters, in other test versions and on other test occasions. Observed scores are estimates of expected scores over parallel tests and across raters. This kind of inference is based on the assumptions of generalizability theory, i.e. that a student is likely to obtain the same score on similar tasks. In our example, generalization may lead to an expected score of, say, 35% on other versions of the academic writing test.

The generalization inference can be problematical, as it depends on the sampling that has been done in the first inference – whether the sample is representative and adequate, so that generalization is in fact possible. There is always the potential for reduced generalizability (the "low-generalizability problem") since tasks tend to be context-specific, which means that inferences which are based on them may not always extend to the domains they are intended to represent. In a performance test in particular, a "substantial number of tasks may be needed" to estimate a student's achievement (Shavelson, Baxter & Pine, 1992: 26). If the test tasks are not adequately representative of the target domain, adequate generalization will not be possible. Performance tests often have low generalizability when only overall scores are recorded and the mixes of different skills and knowledge – the individual differences in student profiles – are not reflected in the overall score. We also have to be sure that scores are in fact reliable before we can generalize them, and that there are no or few misclassifications. In addition, it is crucial to consider the level of generalization required, e.g. certification is high-stakes, and requires a high level of generalizability, whereas student-level results require lower levels of generalizability.

An *extrapolation* inference is the next link and advances the argument about the meaning of the score. It links the universe score to a target score. This interpretation argues that the test score reflects (predicts) the same quality of performance as a performance in the real world. For example, in a placement test, it makes a claim about how well the student is likely to do in various courses. It reflects what a test taker knows or can do, based on the universe score. It indicates the score that would be obtained in a non-assessment task in the target domain (extrapolated to a broader, different domain), and leads to the target score. In our example, the student is likely to encounter difficulty in writing academic assignments at university and will obtain low scores in them. The extrapolation inference depends on both content and criterion (predictive and/or concurrent) validity: test tasks must be representative samples of the domain, and scores must correlate with criterion measures.

Kane (2001: 330) allows for the inclusion of an *explanation* inference, which is also referred to as a *theory-based* inference. He thus accommodates construct-based validity, which is an important consideration in language testing. This inference serves to explain the performances, making the interpretation richer, and attributes to the target score the meaning of a theory-defined construct. The warrant that underlies the explanation is that scores can be attributed to a construct of language ability, which accounts for the performance.

Much work remains to be done with regard to the inclusion of a construct in the interpretive argument. References to a construct can be very abstract and have little practical meaning. What are in fact tasks can be paraded as constructs. What needs to be better understood is how the role of the construct can and should influence the way it is defined. Chapelle (2012: 24) argues that theoretical constructs for language tests are not *a priori* existing entities, but are constructed at "the interface of prior work, conceptual possibilities and pragmatic needs". As we have seen, Kane suggests that we should not or need not begin with a construct. But I think it is quite feasible to begin by asking what skills or knowledge should be tested, and then to ask what behaviours would represent these skills and knowledge, and then decide on tasks that can elicit them (cf. Messick, 1994: 17). Despite the difficulties in defining a language proficiency construct precisely, we often have a specification of the abilities we want to test in mind, and these can take the form of a construct of sorts. We must remember, however, that a construct remains an abstract concept, and that it needs to be operationalised. If we spell out the purpose of an assessment, we can determine the aspects of proficiency that will serve the purpose in terms of either competencies or behavioural aspects.

The *utilization* or *decision* inference focuses on the implications and uses of test scores. A utilization inference warrants the appropriateness of certain implications suggested by the scores. This involves the ethics of language testing, and considers any implications, in order to ensure a positive effect on teaching. A decision inference connects the score-based interpretation to an intended decision. The estimates of the quality of performances can be used to make decisions about placement, certification or curricula. For example, the student may be placed in a remedial or supplementary course in academic writing.

This inference can be complex as well as controversial in language testing. Shohamy (2012: xv) points out that validity "requires the protection and guarding of the personal rights of candidates as well as positive washback on learning by addressing the diverse communities in which the tests are used". Bachman and Palmer (2010), building on their previous work, suggest that the uses and possible consequences should in fact be the starting point for both test design and evaluation[1]. There is often no direct link between a test score and a decision. A decision about a candidate may have to be based on multiple sources (e.g. various examination papers in a subject), and all these sources would have to be included in the interpretive argument. In addition, Davies (2012: 41) points out that there are limits to test developers' control over test users, and thus limits to their responsibility in how tests results are used. The utilization/decision inference involves consequential validity, and depends on the soundness of the interpretations preceding it. Reliability, construct validity, and content and criterion validity will determine and influence this inference.

The inferences form the logical links in a chain, expressed by a "bridges" metaphor (Kane, Crooks & Cohen, 1999: 9; Chapelle et al., 2008: 9), and can be represented as follows:

> TARGET DOMAIN – Sampling – SAMPLE OF OBSERVATIONS – Scoring – OBSERVED SCORE – Generalization – UNIVERSE SCORE – Extrapolation – TARGET SCORE – Theoretical interpretation/explanation – THEORY-BASED CONSTRUCT – Implications – UTILIZATION/DECISION

Each inference within the interpretive argument requires a set of supporting evidence. Writers such as Bachman (2005b), Weir (2005) and Xi (2010) describe various methods for the collection of evidence that can be collected for each inference. These can include analysis of variance, Rasch analysis[2], verbal protocols, score reliability analyses and so on (cf. Xi, 2010). In this regard, Albert Weideman's contribution in providing empirical evidence for the interpretation and uses of TAG and TALL scores has been enormous, with a large number of publications that indicate how these can be supported (cf. http://icelda.sun.ac.za for a list of his publications in this regard).

[1] Bachman and Palmer (2010) argue that all assessments should be evaluated in terms of how well they work in practice. They suggest an Assessment Use Argument (AUA) for the development and evaluation of an assessment. They subsume reliability and validity notions under the AUA, and state that one can justify assessments by formulating an AUA and collecting evidence to support it (Bachman & Palmer, 2010:30)

12. Conclusion

The validity of a test ultimately determines its quality. Validation is always a matter of degree and never an all-or-none judgement. In practice, this often means that validation evidence contains many lines of evidence. The claims or propositions formulated can in fact be almost unlimited. They can also be *ad hoc* and unsystematic. Kane's approach provides a structured process and makes provision for the systematic development of evidence rather than "discovering" it. Each assumption is tied to a particular inference, and this avoids the *ad hoc* stipulation of claims and empirical evidence in order to prove validity.

A concern is that there may be a tendency to regard the basic design of a test as less important. This can surely not be the case. We must be able to say if a test is a good one or not (Davies, 2012: 38). Validity in some sense must still reside inside a test (cf. Davies & Elder, 2005; Borsboom et al., 2004). There must be a reason why we would use one test rather than another one, or as Davies (2012: 38) puts it, a "principled basis" for such a decision.

Davies (2012:40) stresses the importance of concrete validation and evidence collection procedures. Kane (2012:15) states that validation is simple in principle, but difficult in practice. In many respects it still remains a somewhat pragmatic affair. The argument-based approach needs to be further refined, and in this regard, Xi (2010:189) points out that much more effort is required to integrate validity evidence into a coherent argument to support a particular test use. Research should be extended to second-language achievement tests (e.g. the South African matriculation examination) in particular. With the context of interest as the point of departure, Kane's proposal is in line with the current task-based language teaching and learning approach; a method which finds support in second language acquisition studies.

A number of frameworks for validation have recently been proposed (e.g. Hattingh, 2009; Bachman & Palmer, 2010; O'Sullivan & Weir, 2011). Kane's argument-based approach is a major contribution to the discussion of the validation of language tests. It provides a systematic approach for the evaluation of a test at the development, trialling and implementation stages, and provides guidelines for a well-articulated validation methodology.

REFERENCES

Alderson, J.C. & Banerjee, J. 2002. Language testing and assessment (Part 2). *Language Teaching* 35: 79-113.

Alderson, J.C., Clapham, C. & Wall, D. 1995. Language testing 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.

[2] Hattingh (2009) provides a good example of how Rasch analysis can be used in the development and validation of a rating scale for ESL essays.

Bachman, L.F. 2005a. Building and supporting a case for test use. *Language Assessment Quarterly* 2(1): 1-34.

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

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

Bachman, L.F. & Palmer, A.S. 2010. Language assessment in practice: Developing language assessments and justifying their use in the real world. Oxford: Oxford University Press.

Borsboom, D. 2006. The attack of the psychometricians. *Psychometrika* 71(3): 425-440.

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

Brindley. G. 1986. The assessment of second language proficiency: Issues and approaches. Adelaide: National Curriculum Resource Centre.

Canale, M. & Swain, M. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics* 1(1): 1-47.

Carroll, B.J. 1982. Testing communicative performance: An interim study. Oxford: Pergamon.

Chapelle, C.A. 1999. Validity in language assessment. *Annual Review of Applied Linguistics* 19: 254-272.

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

Chapelle, C.A. & Brindley, G. 2010. Assessment. In Schmitt, N. (ed.). An introduction to Applied Linguistics. Abington, Oxon.: Hodder & Stoughton.

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.

Cronbach, L.J. 1971. Test validation. In Thorndike, R.L. (ed.). Educational measurement. 2nd ed. Washington, DC: American Council on Education.

Cronbach, L.J. 1980. Validity on parole: How can we go straight? *New Directions for Testing and Measurement* 5: 99-108.

Cronbach, L.J. 1988. Five perspectives on validity argument. In Wainer, H. & Braun, H. (eds.). Test validity. Hillsdale, N.J.: Lawrence Erlbaum.

Cronbach, L.J. & Meehl, P.E. 1955. Construct validity in psychological tests. *Psychological Bulletin* 52(4): 281-302.

Cureton, E.E. 1951. Validity. In Lindquist, E.F. (ed.). Educational Measurement. 1st ed. Washington, DC: American Council on Education.

Davies, A. 2012. Kane, validity and soundness. *Language Testing* 29(1): 37-42.

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

Douglas, D. 2000. Assessing language for specific purposes. Cambridge: Cambridge University Press.

Fulcher, G. 2000. The ‘communicative’ legacy in language testing. *System* 28: 483-497.

Fulcher, G. & Davidson, F. 2007. Language testing and assessment: An advanced resource book. London: Routledge.

Hall, C.J., Smith, P.H. & Wickaksomo, R. 2011. Mapping Applied Linguistics. London: Routledge.

Hattingh, K. 2009. The validation of a rating scale for the assessment of compositions in ESL. Unpublished Ph.D. thesis. Potchefstroom: North-West University.

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.). Educational measurement. 4th ed. Praeger: American Council on Education.

Kane, M.T. 2012. Validating score interpretations and uses. Messick lecture, Language Research Colloquium, Cambridge, April 2010. *Language Testing* 29(1): 3-17.

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

Knoch, U. 2009. Diagnostic writing assessment: The development and validation of a rating scale. Frankfurt am Main: Peter Lang.

Lado. R. 1961. Language testing: The construction and use of foreign language tests. London: Longmans.

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

Messick, S. 1989. Validity. In Linn, R.L. (ed.). *Educational measurement*. 3rd ed. New York: Macmillan.

Messick, S. 1994. The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23(2): 13-23.

Morrow, K. 1979. Communicative language testing: Revolution or evolution? In Brumfit, C.J. & Johnson, K. (eds.). The communicative approach to language teaching. Oxford: Oxford University Press.

O'Sullivan, B. & Weir, C.J. 2011. Language testing = validation. In O'Sullivan, B. (ed.). Language testing: Theories and practice. Basingstoke: Palgrave Macmillan.

Porter, D. 1983. Assessing communicative proficiency: The search for validity. In Johnson, K. & Porter, D. (eds.). Perspectives in communicative language teaching. London: Academic Press.

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

Shavelson, R.J., Baxter, G.P. & Pine, J. 1992. Performance assessments: Political rhetoric and measurement reality. *Educational Researcher* 21(4): 22-27.

Shohamy, E. 2010. Introduction to volume 7: Language testing and assessment. In Shohamy, E. & Hornberger, N.H. (eds.). Language testing and assessment. Encyclopedia of language and education: Volume 7. New York: Springer.

Toulmin, S.E. 2003. The use of argument. Updated edition. Cambridge: Cambridge University Press.

Van der Walt, J.L. & Steyn, H.S. jnr. 2007. Pragmatic validation of a test of academic literacy at tertiary level. *Ensovoort* 11(2): 138-153.

Wigglesworth, G. 2010. Task and assessment based assessment. In Shohamy, E. & Hornberger, N.H. (eds.). Language testing and assessment. Encyclopedia of language and education: Volume 7. New York: Springer.

Weideman, A. 2009. Constitutive and regulative conditions for the assessment of academic literacy. *Southern African Linguistics and Applied Language Studies* 27(3): 235-251.

Weideman, A. 2011. Academic literacy tests: Design, development, piloting and refinement. *Journal for Language Teaching* 45(2): 100-114.

Weir, C.J. 2005. Language testing and validation: An evidence-based approach. Houndmills, Basingstoke: Palgrave Macmillan.

Wu, W-P. & Stansfield, C. 2001. Towards authenticity of task in test development. *Language Testing* 18(2): 187-206.

Xi, X. 2010. Methods of test validation. In Shohamy, E. & Hornberger, N.H. (eds.). Language testing and assessment. Encyclopedia of language and education: Volume 7. New York: Springer.

ABOUT THE AUTHOR

Johann L. van der Walt
School of Languages
North-West University
Email: Johann.VanDerWalt@nwu.ac.za