Per Linguam 2012 28(2):1-14
http://dx.doi.org/10.5785/28-2-526

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

Per Linguam

A Journal for Language Learning
Tydskrif vir Taalaanleer

VALIDATION AND VALIDITY BEYOND MESSICK

Albert Weideman
University of the Free State

Following the texts of the Messick (1981: 10; 1980: 1023) formulations more closely, another representation of his well-known "Facets of validity" matrix is possible:

	adequacy of…	*appropriateness* of…
inferences made from test scores	depends on multiple sources of empirical evidence	relates to impact considerations/consequences of tests
the design decisions derived from the interpretation of empirical evidence	is reflected in the usefulness/utility or (domain) relevance of the test	will enhance and anticipate the social justification and political defensibility of using the test

This matrix can be read as four claims about language testing ("The technical adequacy of inferences made from test scores depends on multiple sources of empirical evidence; The appropriateness of inferences made from test scores relates to the detrimental or beneficial consequences..." and so forth). This representation still follows Messick's argument, but rather than validity, articulates the coherence of a number of assessment concepts. Such concepts as the technical adequacy of our assessment instruments, their appropriateness, the technical meaningfulness (interpretation) of their measurements, their utility, their social impact and public defensibility indicate that we can usefully reconceptualise not only validation and validity, but all of our efforts at designing assessments responsibly. The current debates in South Africa about standardisation and equivalence can be deepened if we examine ways of going beyond conventional notions of validation and validity, and take responsible design criteria to constitute the overriding condition(s) for the development of assessment instruments.

VALIDATION AND VALIDITY DEBATES TRACE THEIR LINEAGE TO MESSICK

Though multiple origins can be identified for conceptualising validation and validity, the contribution of Messick to the development of validity theory is customarily held up as its culmination (Xi, 2008: 179). The importance of Messick's work on this is often related to its proposal for a 'unitary' concept of construct validity, a characteristic that was taken further by several others, but with varying emphases: Kane's (1992) early association of validity with the interpretation of test scores is noteworthy (and enduring: cf. Kane, 2001, 2011), as are Bachman and Palmer's (1996: 17 *et passim*) subtle modification in promoting usefulness as the 'most important quality' of a test, and Kunnan's (2000: 1) assertion of the 'primacy of fairness'. All re-interpret the 'unitary' condition of validity in slightly different ways, and in so doing perhaps inadvertently introduce a potential disunity and variation within the conceptualisation of that overriding condition, despite their intention to clarify it.

One of the reasons often cited for such reinterpretation is that the operationalisation of the concept of validity, as formulated by Messick (1980, 1981, 1988, 1989), has neither been feasible (Davies & Elder, 2005: 789; Xi, 2008: 179) nor adequate to address social concerns in language test design and administration (McNamara & Roever, 2006: 249; Rambiritch, 2012).

Because Messick's original conceptualisations are the origins of subsequent and current re-conceptualisations, it is only fair to test the implied inadequacy of the original. The aim of this paper is to offer another reading of Messick, but from a conceptual reconsideration of his most influential work. Such an alternative reading will show that there is much more afoot, conceptually, than one would at first glance suspect, and that some of the concepts that were supposedly abandoned by Messick – an abandonment for which he is credited – arise, upon closer inspection, in new conceptual guises. The kind of conceptual analysis to be undertaken in this paper moves into an area where language assessment experts seldom venture: the foundations of the field of applied linguistics. Language testing belongs squarely within applied linguistics (McNamara & Roever, 2006: 255; cf. too McNamara, 2003; Weideman, 2006a, 2011), so the philosophical underpinnings of applied linguistics are equally relevant in this sub-field. What is more, Messick himself (1989:30f.) turns to the 'philosophical foundations of validity and validation' when he singles out the perspectives of Leibniz, Locke, Kant, Hegel and Singer in order to gain conceptual clarity. Achieving conceptual clarity is thus also the central concern of the arguments and interpretations that will be made below. The question to be answered is: How does one obtain conceptual clarity in applied linguistic concept formation, and, by extension, in the fundamental concepts and ideas of language testing?

I shall begin by reconsidering three different readings of Messick's own summary (1980, 1981, 1989) of his unitary concept of validity below, before turning to a broader consideration of language testing principles and, finally, their meaning for large scale testing on a national basis, such as in the final Grade 12 examinations in South Africa.

TWO READINGS AND THEIR CONCEPTUAL PROBLEM

We should observe, first, that Messick's contribution to the notions of validation and validity has often been reinterpreted, in order to give various emphases to what the reinterpreting scholar or scholars wished to highlight. So, for example, we have already noted that Kane (1992: 527) claims that

> (1) Validity is associated with the interpretation assigned to test scores rather than with the scores or the test.

This definition is widely accepted – it seems to echo closely Messick's own claim (1980: 1023; cf. too 1981: 18):

> (2) Test validity is … an overall evaluative judgment of the adequacy and appropriateness of inferences drawn from test scores.

The subtle reinterpretation of definition (2) that we find in definition (1) lies in the emphasis placed in (2) on the judgement of the adequacy and appropriateness of the inferences drawn from test scores (Messick,1980), as against validity not being associated with either scores or the test (Kane, 1992). Kane's redefinition of validity in (1) speaks in the first instance only of an 'interpretation' that is assigned, not about its adequacy and appropriateness. That fine point is often missed, but is crucial in Messick's formulations, as will become clearer below.

Similarly, Bachman and Palmer (1996: 17; also Bachman, 2001: 110), having first shifted the emphasis from validity to test usefulness in their declaration of the latter, and not validity, as the "most important consideration in designing and developing a language test", are subsequently happy to return to Messick's underlying fundamental: the concept of construct validity. They state that

> (3) Construct validity pertains to the meaningfulness and appropriateness of the *interpretations* that we make on the basis of test scores (Bachman & Palmer, 1996: 21; emphasis in the original).

Once, more, Messick's own formulation is given a subtle twist: not the adequacy of the interpretation (actually 'judgement' in Messick, 1980), but its meaningfulness now is placed in the spotlight. To many, who do not see the shift, Bachman and Palmer here are simply following Messick: Xi (2008: 179) even declares that their notion of test usefulness makes 'Messick's work more accessible'. This view is not widely shared, though: Fulcher and Davidson (2007: 15) observe, for example, that Bachman and Palmer's 'notion of test "usefulness" provides an alternative way of looking at validity, but has not been extensively used in the language testing literature'. Language testing experts may unfortunately often not be disinterested but so imbued with the currently orthodox notions of validation and validity that they sometimes cannot see this point, but to any disinterested observer, it should be clear that usefulness can never conceptually be the same as validity. The best illustration of this remains Bachman and Palmer's own model of test usefulness (1996: 18) that *incorporates* validity:

Figure 1: Bachman & Palmer's model of test usefulness.

Usefulness = Reliability + Construct validity + Authenticity + Interactiveness + Impact + Practicality

The incorporation masks their divergence of opinion with Messick, but surely no one would disagree that part (construct validity) of a whole (usefulness) that is made up of several other parts cannot conceptually be the same as that whole. The norms of logic do not allow that, so through such distinctions we are nowhere closer to conceptual clarity or to accessible technical ideas of what conditions language tests must satisfy. What is more, to make validity dependent on interpretation (cf. definitions (3) and (1) especially), and not a characteristic of a test, runs the risk of downplaying the quality of the instrument. No amount of interpretation can improve the measurement result (score) obtained from an inadequate instrument that gives a faulty and untrustworthy reading. This is simply confusing the meaningful, legitimate technical *interpretation* of the effects of measurement with those effects (the test scores) themselves. Moreover, as I have pointed out elsewhere (Weideman, 2009), these are subjective and objective components of the process of dealing with the results of taking measurements.

If these kinds of reinterpretations appear not to give us conceptual clarity, then perhaps it is worthwhile to return to Messick's own formulation (1980: 1023, 1989: 20) of his position, perhaps best summarised in the well-known 'Facets of validity' matrix:

Figure 2: Messick's 'Facets of test validity'.

	Test interpretation	Test use
Evidential basis	Construct validity	Construct validity + Relevance/Utility
Consequential basis	Value implications	Social consequences

This matrix is possibly the most quoted, and certainly the most well-known, summary of his position. Yet it is interesting to note that this summary has itself been subject to reinterpretation, ostensibly in order to make it clearer: McNamara and Roever (2006: 14; for another, cf. Davies & Elder, 2005: 800) have offered the following reinterpretation, a second reading of what Messick says:

Figure 3: McNamara & Roever's interpretation of Messick's validity matrix

	What test scores are assumed to mean	When tests are actually used
Using evidence in support of claims: test fairness	What reasoning and empirical evidence support the claims we wish to make about candidates based on their test performance?	Are these interpretations meaningful, useful and fair in particular contexts?
The overt social context of testing	What social and cultural values and assumptions underlie test constructs and the sense we make of test scores?	What happens in our education systems and the larger social context when we use tests?

As in the case of the slight modifications of the definitions of validity, in (1) and (3), in relation to Messick's own definition (2), that I referred to above, we might again note a subtle shift: 'fairness' (Kunnan) and 'meaningfulness' (Bachman & Palmer), to name but two additional emphases, now present themselves in the second reading. A charitable analysis will no doubt find that, though perhaps not strictly attributable to Messick, they contribute towards our overall understanding of validity, thus explaining the necessity for the slight shift.

A THIRD READING AND A POSSIBLY USEFUL REINTERPRETATION

If one is interested in what Messick himself has said, however, a third reading is possible. This reading, following closely the texts of the Messick (1981: 10, 1980: 1023) formulations, soon makes it evident that, at least in conceptual terms, things are more complicated. When one re-reads the original texts, it is soon apparent that in Messick's conceptualisation the concepts of 'adequacy' and 'appropriateness' are key terms. What is more, they refer to two distinct and distinguishable concepts.

Appropriateness is, in analytical terms, actually not merely a concept, but a concept-transcending idea (Strauss, 2009: 195) that captures the analogical social dimension of the technically qualified design of language tests. *Adequacy*, on the other hand, is a concept that is linkable directly to the effects (or effectiveness) of applying a technical instrument such as a language test. The technical adequacy of a test refers to its force to measure what it claims to be measuring, its effectiveness, which is the classic definition of validity. A measuring instrument is adequate if its result (definable as the effect of the measurement which is caused by the application of the instrument) has the desired force. In its original physical sense, the concept of force is expressed in terms of cause and effect. In the analogical technical sense, it is used when we are dealing with technically qualified instruments such as tests; the measurement, when applied, acts as technical cause to achieve a certain technical effect, that is, to obtain a result. This should not be surprising: *adequacy* is used by Messick, I believe, simply as a concept synonymous with validity. Having discarded the notion of validity as one that cannot be a characteristic of a test – since the validity now resides in the interpretation of the scores rather than in the instrument – a substitute concept is needed, a role that the concept of adequacy soon steps in to fulfil.

Despite, in line with the current orthodoxy, having foresworn the practice of using validity as a characteristic of a test, as we saw is the case especially in definitions (1) and (3) above, using validity as descriptive of a test therefore merely returns in another guise, that of adequacy, or in similarly synonymous terms for the concept of effectiveness. The discarded concept arises from its conceptual ashes to reassert itself. When ascribing validity to a test is no longer tolerable, that often leads to circumlocutions such as a 'test … accomplishing its intended purpose' (Messick, 1980: 1025), or of tests 'purported to tap aspects' of a trait (Messick, 1989: 48; 50, 51, 73). In utilising synonymous concepts, Messick is in no way alone, however. Many who subscribe to the current orthodoxy, and may even decry, for example, Popham's early (1997) and Borsboom, Mellenbergh and Van Heerden's later (2004) views that are critical of Messick, themselves use synonymous concepts as substitutes for validity. In that case, they may find themselves referring to the 'effectiveness' of the use to which a test can be put (Lee, 2005: 2), of a test being 'valid in a specific setting' (Lee, 2005: 3), or that we can investigate through verbal protocols the consequences of a test, since these 'should be considered valid and useful data in their own right'. They may similarly employ some circumlocution to avoid referring to validity as a quality of a test: '… if we ensure that a given test measures the construct … we say that the resulting scores provide an empirically informed basis for decision-making' (Lee, 2005: 4). McNamara and Roever (2006: 17) themselves continue to speak about the validity of a test, or to assume that a 'test is … a valid measure of the construct' (McNamara & Roever, 2006: 109), and to speak about 'items measuring only the skill or the ability under investigation' (McNamara & Roever, 2006: 81) – and not about the interpretation of the scores derived from these items.

The point of this again echoes the observation above that, for the sake of conceptual clarity, one should distinguish between the objective effect of the measurement that derives from a valid, effective instrument and the subjective interpretation of that effect. If clear conceptual distinctions are not made, the distinction that has been avoided is subsequently likely to re-assert itself. The distinction between technical causes and effects remains relevant:

It seems to me that some of the critique of validity theory merely wants to say: If a test does what it is supposed to do, why would it not be valid? Surely a test that accomplishes its intended purpose has the desired effect, that is, yields the intended measurements? However, causes and effects, and the relationship between causes and effects in the field of testing, are analogical technical concepts, that is, concepts formed by probing the relationship of the leading technical function of a designed measurement instrument to the physical sphere of energy-effect, the domain in which these concepts are originally encountered. To say that a test is valid is therefore merely identical to saying that it has a certain technical or instrumental power or force, that its results could become the evidence or causes for certain desired (intended or purported) effects (Weideman, 2009: 241).

By taking the employment of the terms *adequacy* and *appropriateness* in the original Messick texts, and recasting them, we might make possible another representation that might in its turn enable us to come up with a reinterpretation that will show a way out of the conceptual impasse referred to above (Weideman, 2009: 240):

Figure 4: The relationship of a selection of fundamental considerations in language testing.

	adequacy of…	*appropriateness* of…
inferences made from test scores	depends on multiple sources of empirical evidence	relates to impact considerations / consequences of tests
the design decisions derived from the interpretation of empirical evidence	is reflected in the usefulness/utility or (domain) relevance of the test	will enhance and anticipate the social justification and political defensibility of using the test

As I have remarked elsewhere (Weideman, 2009), the statements generated by this matrix (Figure 4) can be read as a number of claims about or requirements for language testing, as follows (left to right, top to bottom):

(4) The technical adequacy of inferences made from test scores depends on multiple sources of empirical evidence.

(5) The appropriateness of inferences made from test scores relates to the detrimental or beneficial impact or consequences that the use of a test will have.

(6) The adequacy of the design decisions derived from the interpretation of empirical evidence about the test is reflected in the usefulness, utility, or relevance to actual language use in the domain being tested.

(7) The appropriateness of the design decisions derived from the interpretation of empirical evidence about the test will either undermine or enhance the social justification for using the test, and its public or political defensibility.

These claims have been numbered here to facilitate comparison and contrast with the validity definitions (1) to (3) above. They can indeed be simplified, reinterpreted, and made more accessible and useful, as guidelines for those who design language tests. Some (not all) of such possibly blander versions might include the following:

(4a) Use multiple sources of empirical evidence to make adequate inferences about test scores.

(5a) The more appropriate the inferences made from test results, the more likely they are to be beneficial to everyone.

(6a) The test design and its relevance will improve if one heeds empirical evidence about actual language use in the domain being tested.

(7a) A good test will use empirical evidence to defend its social appropriateness.

I shall return below to the use of such guidelines. To return to the current argument, however, note that the unsimplified claims ([4] to [7]) still follow Messick's formulation, and yet the matrix in Figure 4 is by no means a 'validity matrix', as the original claims to be. Nor are the statements derived from it ([4] to [7]) strictly about validity. Statements (4) to (7), or their subset (4a) to (7a) may be obliquely related to the technical force of a test, it is true, but conceptually we would gain much if we note that they articulate the coherence or systematic fit of a number of concepts relating to testing. What is more, they also articulate some social dimensions of language testing (McNamara & Roever, 2006; Rambiritch, 2012), particularly the social appropriateness, impact, benefits of and public accountability for tests. Thus, if these statements are not only about validity, validity itself may perhaps not be the overriding, unifying condition to which tests should be subjected. This conclusion might explain why, for example, McNamara and Roever (2006: 249) observed that 'validity theory has remained an inadequate conceptual source for understanding the social function of tests'. There is no doubt, however, that a more appropriate label for Figure 4 would be that it deals with the relationships among a select number of fundamental concepts in language testing. It is in that direction, therefore, that one needs to seek to deploy a new understanding of Messick's original work, and it is to that which I turn below.

CONDITIONS FOR RESPONSIBLE TEST DESIGN

If validity is not the overriding condition for test design, but rather a systematic set of principles (of which [4] to [7] above are probably a subset, related to the technical adequacy and appropriateness of a test), the subsequent critical question then has to be: Why should conditions for responsible test design continue to be subsumed under 'validity'? As has been demonstrated, we achieve no greater conceptual clarity when we conflate the various design conditions that apply to tests. Far from helping us to reinterpret validity in order to clarify it, such reinterpretation may instead confuse. Statements (4) to (7) above confront us with concepts such as technical adequacy, appropriateness, the technical meaningfulness (interpretation) of measurements (test scores), utility, relevance, public defensibility and the like, and to make sense of them, they must be conceptually distinguishable as constitutive technical concepts or regulative, technical ideas that transcend concepts. Therefore, if they are distinguishable, that means that they are conceptually distinct.

There is not enough space here to trace, exhaustively, the generation of constitutive technical concepts and regulative ideas guiding the design of language tests in recent conceptualisations (Weideman, 2009). These two sets of conditions for language testing have, however, recently been elaborated in two studies by Van Dyk (2010) and Rambiritch (2012). Van Dyk's (2010) study was initially conceived as a validation study for the ICELDA-designed (ICELDA 2012) Toets van Akademiese Geletterdheidsvlakke (TAG, the Afrikaans counterpart of TALL, the widely used Test of Academic Literacy Levels; cf. too Van der Walt & Steyn 2007), but was reconceptualised, in light of the foregoing argument, to focus more comprehensively on the constitutive conditions for language test design such as systematicity, reliability, validity and validation, construct defensibility, and meaningfulness of results. The second study, by Rambiritch (2012), though again paying attention to constitutive conditions such as the technical consistency and validation of another ICELDA-developed test, the Test of Academic Literacy for Postgraduate Students (TALPS), deals more comprehensively with the regulative requirements of accessibility, transparency, and accountability, and does so in a systematic way that has not been attempted before.

The emerging framework for test design (Weideman, 2009) in particular, and for applied linguistic designs in general (Weideman, 2007) from which this more comprehensive set of conditions derives, makes it clear that we may usefully consider as either founding or constitutive, or as leading, disclosing and regulative requirements for our test designs those concepts and ideas listed below. These requirements have been formulated in the style of statements (4a) to (7a) above. Again, they are blander than may be desirable, but conceptually they can all be traced to the framework being developed in the studies that have been referred to above:

- Systematically integrate multiple sets of evidence in arguing for the validity of a test.
- Specify clearly and to the public the appropriately limited scope of the test, and exercise humility in doing so.
- Ensure that the measurements obtained are adequately consistent, also across time.

- Ensure effective measurement by using a defensibly adequate instrument.
- Have an appropriately and adequately differentiated test.
- Make the test intuitively appealing and acceptable.
- Mount a theoretical defence of what is tested in the most current terms.
- Make sure that the test yields interpretable and meaningful results.
- Make not only the test, but information about it, accessible to everyone.
- Obtain the test results efficiently and ensure that they are useful.
- Align the test with the instruction that will either follow or precede it, and as closely as possible with the learning.
- Be prepared to give account to the public of how the test has been used.
- Value the integrity of the test; make no compromises of quality that will undermine its status as an instrument that is fair to everyone.
- Spare no effort to make the test appropriately trustworthy.

It should be clear that many of the conditions are simply reformulations of well-known concepts. So, for example, the requirement that a test must be consistent is a reference to its technical reliability, usually expressed in an index that measures this, such as Cronbach's alpha or Greatest Lower Bound (GLB; cf. Jackson & Agunwamba, 1977). Similarly, the requirement that a test should be effective is a reformulation of its being valid. Its being intuitively appealing is a reference to the notion of face validity, and obtaining useful results efficiently refers in turn to the notion of the technical usefulness or utility of the test. Aligning it with instruction and learning requires utilising the positive effects of washback, the harmonisation of teaching, testing and learning that is so difficult to achieve, but remains the essence of purpose of all responsible pedagogy.

These formulations, it should be noted, have the benefit not only of relating the test to its intrinsic conventional conditions (reliability, construct and other forms of validity, and so on), but also of specifying the so-called 'social' dimensions of tests (accessibility, accountability, fairness) as *inherent* requirements for responsible test design, not as add-ons – a critique often levelled against Messick (Popham, 1997).

The simplification of the technical design criteria as articulated here has the further benefit of facilitating their application to the large-scale testing undertaken annually in the Grade 12 exit examinations, to which I finally turn.

APPLICATION TO THE SOUTH AFRICAN FINAL SCHOOL EXAMINATIONS

Space will not allow me to discuss the relevance of all of the above criteria to what is probably the most important set of tests in the South African context, our Grade 12 exit-level examinations, so I shall select only five on the basis of their being potentially less well attended to than some of the others, in order to make the application. The selection is justified, in addition, by the conditions deriving not from a haphazard conceptualisation but from the same framework. That means, as will be apparent below, that they are integrated and interlinked; a reference to one calls up attention to another.

The condition that carries the heaviest public and political weight is the one that asks of us, Umalusi (the Council for Quality Assurance in General and Further Education and Training), and the larger education system that generates these examinations to value their integrity, and to brook no compromise of quality. In this respect, we are clearly failing. The public perception, whether accurate or not, is that we have experienced 'category creep'; put bluntly: that an A in today's examinations is more or less equivalent to a D or at most a C of 50 years ago. Strategies to address this perceived devaluation must restore the integrity of these tests, otherwise substitutes of all sorts will arise, even though the latter may follow the politically safe route of claiming that they are not substitutes. Fortunately, the results of the examinations in question still remain the best (most valid or effective) predictor of performance in the year and, in some contexts, even several years beyond their origin.

One of the best strategies to ensure the integrity of these tests is to attend to another of the conditions above: the preparedness to give a public account of how the results were obtained. I refer here to reports in the press, for example, on the variation in home language marks and averages across different languages, that are a clear indication of testing that is unfair to some. In my own dealings with Umalusi, I have been impressed by their concern for more than just this one issue. However, there is no doubt that public accountability can be improved. Applying another condition clarifies one way of how this can be done: Ensure that the measurements obtained are adequately consistent, also across time. For the latter to happen, one needs some measure of standardisation. The kind of standardisation currently employed, while rationally defensible, is not adequate to be credible to the public at large. In saying so I do not wish to defend the clearly ignorant barrage of criticism that Umalusi has to endure annually. There is clearly scope for large-scale education of the public; for example, that in the absence of standardised measurements across the years, it is unreasonable to expect averages not to vary from one year to the next. Having seen how effective public communication can be done by making available enough information about tests to prospective students at the multilingual universities joined in ICELDA, I can recommend better, utterly honest, and clearer communication.

This brings me to the final two criteria. There should be a concern about whether the tests in question are defensibly adequate instruments, which in turn relates strongly, in my opinion, to whether we have articulated not only what we are measuring (the construct) but also a suitable measure of differentiation in what we test. Again, my own experience in the testing of academic literacy indicates that monotone designs are inadequate, and that the richer the construct that is measured, and the more differentiated task and item design is, the more likely one is also to have a stable and consistent instrument that can outperform any alternative (Van der Slik, 2008; Van der Slik & Weideman, 2005, 2007, 2008, 2009, 2010; Van Dyk & Weideman, 2004a, 2004b; Weideman, 2003, 2006b; Weideman & Van der Slik, 2008). We look forward to more consistent, effective and differentiated measurements, and should pledge our own commitment to obtaining those. There is much work to be done in the responsible design of tests, and Umalusi needs the support of its associated communities of experts to achieve its goals in this respect.

A FINAL WORD

This contribution began by referring to the inordinate importance of the notions of validity and validation in testing, surveying several attempts that supposedly clarify them. The consideration of these clarifications yields the somewhat unhappy conclusion that perhaps, especially if we look at the origins of these concepts, we should not try to subsume everything under validity. The question then is: Do we need to abandon attempts at making reinterpretations, taking further the original concepts and ideas of earlier experts? The answer in that case is *no*, but we should seriously consider abandoning the notion of an overarching validity in favour of referring instead to an idea of responsible test design. In the two preceding sections, I have attempted to articulate both the possible conditions for test design derived from an emerging framework that a team of doctoral students and I have been working on, and how they may be applied to the South African context. It is a pity that we see too little of such foundational discussion within applied linguistics and in language test design. This contribution has tried, in reconceptualising not only validity, but also a host of other, systematically generated technical concepts and ideas, to do exactly that.

I believe we should not be afraid to go beyond Messick. Once we are prepared to do that, a whole new world of responsible test design beckons and awaits.

REFERENCES

BACHMAN, LF. 2001. Designing and developing useful language tests. In Elder, C., A., A., Brown, E Grove, K Hill, N Iwashita, T Lumley, T McNamara & K O'Loughlin (Eds.). *Experimenting with uncertainty: Essays in honour of Alan Davies*. Cambridge, UK: Cambridge University Press. 109-116.

BACHMAN, LF & AS PALMER, 1996. *Language testing in practice: Designing and developing useful language tests*. Oxford, UK: Oxford University Press.

BORSBOOM, D, GJ MELLENBERGH & J VAN HEERDEN. 2004. The concept of validity. *Psychological Review,* 111(4):1061-1071.

DAVIES, A & C ELDER. 2005. Validity and validation in language testing. In Hinkel, E. (Ed.). *Handbook of research in second language teaching and learning*. Mahwah, NJ: Lawrence Erlbaum. 795-813.

FULCHER G & DAVIDSON F 2007. *Language testing and assessment: An advanced resource book.* New York: Routledge.

INTER-INSTITUTIONAL CENTRE FOR LANGUAGE DEVELOPMENT AND ASSESSMENT (ICELDA). 2012. ICELDA … a partnership of four multilingual universities. [Online]. Available http://icelda.sun.ac.za/.

JACKSON, PW & CC AGUNWAMBA..1977. Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I. Algebraic lower bounds. *Psychometrica,* 42: 567-578.

KANE, MT. 1992. An argument-based approach to validity. *Psychological Bulletin* 112(3):527-535.

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

KANE, MT. 2011. Validity score interpretations and uses: Messick lecture, Language Testing Research Colloquium, Cambridge, April 2010. *Language Testing* 29(1):3-17.

KUNNAN, AJ. 2000. Fairness and justice for all. In Kunnan, AJ (Ed.). *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida*. Cambridge, UK: University of Cambridge Local Examinations Syndicate. 1-14.

LEE Y-J. 2005. Demystifying validity issues in language assessment. Applied Linguistics Association of Korea Newsletter. October. [Online]. Available http://www.alak.or.kr/2_public/2005-oct/article3.asp.

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

MCNAMARA, T & C ROEVER. 2006. *Language testing: The social dimension*. Oxford, UK: Blackwell.

MESSICK, S. 1980. Test validity and the ethics of assessment. *American Psychologist* 35(11):1012-1027.

MESSICK, S. 1981. Evidence and ethics in the evaluation of tests. *Educational Researcher* 10(9):9-20.

MESSICK, S. 1988. The once and future issues of validity: Assessing the meaning and consequences of measurement. In Wainer, H & IH Braun. (Eds.). *Test validity.* Hillsdale, NeJ: Lawrence Erlbaum. 33-45.

MESSICK ,S. 1989. Validity. In Linn, RL (Ed.). 1989. *Educational measurement*. (3rd ed.) New York, NY: American Council on Education/Collier Macmillan. 13-103.

POPHAM, WJ. 1997. Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and Practice*. Summer 1997: 9-13.

RAMBIRITCH, A. 2012. *Accessibility, transparency and accountability as regulative conditions for a post-graduate test of academic literacy*. Unpublished doctoral thesis. Bloemfontein: University of the Free State.

STRAUSS, DFM. 2009. *Philosophy: Discipline of the disciplines*. Grand Rapids, MI: Paideia Press.

VAN DER SLIK, F. 2008. Gender bias and gender differences in two tests of academic literacy. *Southern African Linguistics and Applied Language Studies* Special issue: Assessing and developing academic literacy (ed.: J. Geldenhuys) 27(3): 277-290.

VAN DER SLIK, F. & WEIDEMAN, A. 2005. The refinement of a test of academic literacy. *Per linguam* 21(1):23-35.

VAN DER SLIK, F. & WEIDEMAN, A. 2007. Testing academic literacy over time: Is the academic literacy of first year students deteriorating? *Ensovoort* 11(2): 126-137.

VAN DER SLIK, F. & WEIDEMAN, A. 2008. Measures of improvement in academic literacy. *Southern African linguistics and applied language studies* 26(3):363-378.

VAN DER SLIK, F. & WEIDEMAN, A. 2009. Revisiting test stability: further evidence relating to the measurement of difference in performance on a test of academic literacy. *Southern African Linguistics and Applied Language Studies* Special issue: Assessing and developing academic literacy (ed.: J. Geldenhuys), 27(3):253-263.

VAN DER SLIK, F. & WEIDEMAN, A. 2010. Examining bias in a test of academic literacy: Does the *Test of Academic Literacy Levels* (*TALL*) treat students from English and African language backgrounds differently? *SAALT Journal for language teaching* 44(2):106-118.

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.

VAN DYK, T. 2010. *Konstitutiewe voorwaardes vir die ontwerp en ontwikkeling van 'n toets vir akademiese geletterdheid*. Unpublished Ph. D. thesis. Bloemfontein: University of the Free State.

VAN DYK, T. & WEIDEMAN, A. 2004a. Switching constructs: on the selection of an appropriate blueprint for academic literacy assessment. SAALT *Journal for language teaching* 38 (1): 1-13.

VAN DYK, T. & WEIDEMAN, A. 2004b. Finding the right measure: from blueprint to specification to item type. SAALT *Journal for language teaching*. 38 (1): 15-24.

WEIDEMAN, A. 2003. Assessing and developing academic literacy. *Per linguam* 19 (1 & 2): 55-65.

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

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

WEIDEMAN, A. 2007. Towards a responsible agenda for applied linguistics: Confessions of a philosopher. *Per linguam* 23(2): 29-53.

WEIDEMAN, A. 2009. Constitutive and regulative conditions for the assessment of academic literacy. *Southern African linguistics and applied language studies* Special issue: Assessing and developing academic literacy (ed.: J. Geldenhuys) 27(3): 235-251.

WEIDEMAN, A. 2011. Straddling three disciplines: Foundational questions for a language department. 30th DF Malherbe Memorial Lecture. *Acta varia*. Bloemfontein: University of the Free State.

WEIDEMAN, A. & Van der Slik, F. 2008. The stability of test design: Measuring difference in performance across several administrations of a test of academic literacy. *Acta academica* 40(1): 161-182.

WEIDEMAN, A. & Van Rensburg, C. 2002. Language proficiency: current strategies, future remedies. *SAALT Journal for language teaching* 36 (1 & 2): 152-164.

XI, X. 2008. Methods of test validation. In Shohamy, E & Hornberger, N. (eds.). *Language testing and assessment. Encyclopedia of language and education 7.* New York: Springer Science + Business Media, pp. 177-196.

BIOGRAPHICAL NOTE:

Albert Weideman is professor of applied language studies in the Department of Linguistics and Language Practice, University of the Free State. He is also CEO of ICELDA, a formal partnership of four multi-lingual South African universities (UP, NWU, Free State and Stellenbosch), that designs language tests and courses. His research focuses on developing a foundational framework for working responsibly within applied linguistics. Email address: albert.weideman@ufs.ac.za