Constitutive and regulative conditions for the assessment of academic literacy

Albert Weideman
University of Pretoria

Abstract

If we characterise language tests as applied linguistic instruments, we may argue that they therefore need to conform to the conditions that apply to the development of responsible applied linguistic designs. Conventionally, language tests are required to possess both validity and reliability; these are necessary conditions for such tests, and so is their theoretical defensibility ("construct validity"). The new orthodoxy, however, is that test designers must also seek consequential validity for their instruments, in that they have to consider test impact. Using an emerging framework for a theory of applied linguistics, this paper outlines how a number of constitutive or necessary conditions for test design (their instrumental power, their consistency and theoretical justification) relate to recently articulated notions. These notions are test acceptability, utility, and alignment with both the instruction that follows and with the real or perceived language needs of students, as well as their transparency, accountability and care for those taking it. The latter set of ideas may be defined as regulative or sufficient conditions for language tests. These concepts will be illustrated with reference to the design of a test of academic literacy levels that is widely used in South African universities.

Necessary and sufficient requirements for test design

One would be forgiven, upon taking a closer look at recent discussions in the field, if one concludes that many language test designers do not currently consider the notion of necessary and sufficient conditions for language tests to be topical. Perhaps, as in so many other respects, they take their cue from Messick, who remarked (1980: 1019) in an early comment:

> In all of this discussion I have tried to avoid the language of necessary and sufficient requirements, because such language seemed simplistic for a complex and holistic concept like test validity.

Like most else, Messick's objections relate in the first instance to his conception of test validation. If the validation of a test, defined by him (1980: 1023; cf. too American Educational Research Association 1999: 9) as the presentation of evidence of how adequate and appropriate the inferences are that we draw from test scores, is an action that is essentially ongoing, always in process, how can we then think that we shall ever have reached the point where we are able to say: we now have sufficient evidence (cf. too Messick 1988: 41)? And how shall we know, in

the rich variation or "density" of the evidence collected to validate the test, what is really necessary, and what is not (Messick 1980: 1019)? The first of these two points is wholly reasonable: if validation is an ongoing process, we may well be enduringly dissatisfied with both the process and with its result. There may be good reasons for such dissatisfaction: as has recently been emphasised in line with postmodernist perspectives on language testing (Shohamy 2001b, 2004; cf. too Davies & Elder 2005: 797; also 800f.), a validation process does not yield a one size fits all result, and one that is immutable for all times. What is a suitable test for one context (cf. Van der Walt & Steyn 2007), may not be for the next. But the last point, that we cannot know what is necessary, is a strange one, and in my opinion highly contestable, yet nowhere contested, as far as I have been able to determine. It is a contentious statement because arguments, through which evidence is organised, are built upon support, and if the necessary support is lacking, they do not hold. But even more interesting is that Messick does not here consider that, while evidence may in his definition never be *sufficient*, the production of evidence may indeed be a *necessary* foundation for the validation process. The latter conclusion seems to me to be fully compatible with his intentions, yet he does not draw it. Perhaps his reluctance to use the terms at all obscures the consideration of this possibility.

As is evident, Messick's remark is made fully within the context of the main tenet of his conceptualisation of test validity: the notion that language testing needs what has come to be called a "unitary" (Messick 1988: 35, 40f.; 1981: 9; 1989: 19) or overarching concept of test validity, through which all the other important requirements traditionally identified must be brought into focus. By traditionally identified conditions I mean requirements like the technical consistency or reliability of a test, its rational defensibility or construct validity, the relevance of its content to the domain which is being probed, the accurate and appropriate interpretation of its technically achieved measurements (scores), and the social consequences of the test results, as well as their wider impact. As McNamara and Roever (2006: 18) have remarked: "… through its concern for the rationality and consistency of the interpretations made on the basis of test scores, validity theory is addressing issues of fairness, which have social implications." Why the mantle has fallen (to use Messick's own imagery; 1980: 1015) on validity, and specifically construct validity, is perhaps less clear, though arguments are presented for the primacy of the kind of validity that relate to the theoretical defence of the construct (Messick 1980: 1014f.; 1988: 38f.). But primacy need not imply the need for further promotion to "unitary" or "the unified view" (Messick 1988: 43; 1981: 9). Following Messick (e.g. 1988: 35; 1989: 19), most current discussion assumes that we need some "overarching" notion, and generally accepts this on the – in my opinion insufficient – grounds that all other types of validity can either be subsumed under construct validity, or be provided with explanations that make them somehow less important.

Bachman and Palmer (1996: 23) have fewer qualms than Messick about the terms necessary and sufficient. To them, for example, reliability "is a necessary condition for construct validity … not a sufficient condition for either construct

validity or usefulness." But they have less of a quibble with the notion of an overarching idea that may combine all of the requirements for a language test: in the "most important quality of …test ... usefulness" (1996: 17 *et passim*) they find just such a comprehensive notion. That their notion of usefulness nonetheless attempts to bypass and challenge the primacy that Messick ascribes to validity is, from the conceptual arguments that will be presented in this paper, perhaps an even more significant development.

The terms "necessary and sufficient" are indeed not entirely adequate. However, though my reservations for employing the notion of necessary and sufficient requirements for language tests do not coincide with those of Messick, but derive rather from philosophical misgivings about cutting up the world in this way, my own continued use of the terms derives more from an attempt to make more intelligible two other concepts that I have used in the past. These are the notions of *constitutive* and *regulative* conditions for the responsible design of applied linguistic artefacts such as language tests and language courses. These are ideas that fit into an emerging framework for applied linguistics, a theory of applied linguistics, that I have used to articulate certain central current and historical ideas, such as consistency, validity, utility, transparency, accountability, and the like, that we encounter in applied linguistic conceptualisation (Weideman 2006, 2007b, 2008). I sometimes use the term constitutive interchangeably with the term "necessary", and regulative interchangeably with "sufficient", since they do seem to go some way towards explaining certain important components of "constitutive" and "regulative."

A foundational perspective

In this paper, I intend to broaden the argument referred to above, and the debate on having "overarching" or "unitary" ideas leading the field of applied linguistics, by articulating further the emerging framework referred to above, specifically in attempting to deal with such conceptualisation within the perspective of a foundational view of the field of applied linguistics.

From this, it follows that I consider language testing to be a part of applied linguistics. In McNamara and Roever's terms (2006: 255; cf. too McNamara 2003), language testing is in fact "a central area of applied linguistics," a claim that, for lack of space, I shall not produce further argument or evidence for (but cf. the discussion in Weideman 2006). If it is so, however, it also follows that tests, as instruments designed to measure language ability, have to meet the same criteria, or, phrased differently, would be subject to the same necessary and sufficient, or constitutive and regulative, conditions and requirements as are applicable to all applied linguistic artefacts, whether these be a designed language course, a language test, or a language policy or plan. If, as I have argued (Weideman 2007b), applied linguistics is indeed a discipline of design, it also means that such designs are led and guided by the technical dimension of our experience, and that it is this dimension that characterises concept-formation in all subfields of applied

linguistics. Furthermore, if that which we produce by way of tests, or courses, or language plans, are stamped and qualified, or primarily characterised by their technical design or blueprint, the first critical question for the current discussion would be: why, if that idea is desirable, should a unifying or overarching notion of an applied linguistic artefact, such as an instrument designed to measure language ability, not be located in the technical dimension? Why seek it in a derived, analogical concept such as (technical) validity or utility?

The second motivating factor in preparing the argument of this paper on constitutive and regulative conditions for language testing was an observation I made when dealing with certain difficult decisions in the design of one particular set of tests of academic literacy, the *Test of Academic Literacy Levels* (TALL) and its Afrikaans counterpart, the *Toets van Akademiese Geletterdheidsvlakke* (TAG). These are tests that are administered to more than 20000 students annually at three South African universities, who jointly develop and produce it: Northwest University, and the Universities of Pretoria and Stellenbosch. The tests have been described (Van Dyk & Weideman, 2004a, 2004b; Weideman 2003) and analysed in a number of other papers (Van der Slik & Weideman 2005, 2007, 2008; Weideman & Van der Slik 2007). The context of this paper is therefore not only the debates in the field of language testing that have been referred to above, particularly the debates on the technical validity (Messick) or usefulness (Bachman and Palmer) that have briefly been mentioned there, but, quite concretely, my current professional engagement with the design and development of a number of tests of academic literacy.

The observation in question was that, as the tests were being developed and beginning to be administered, several trade-offs presented themselves. One such trade-off that I have noted before is discussed in more detail elsewhere (Weideman 2006: 78f.). It is the choice that, in the case of TALL and TAG, the test designers must make between the technical consistency and the appropriateness or relevance of these tests. Conventionally, tests are assumed to measure a single, homogenous ability. If they do not, but instead measure more than one (i.e. a heterogeneous) ability, this shows up particularly well in one technical measure of consistency, a factor analysis. For the latest (2008) version of TALL, for example, it is clear from this kind of analysis that the technical consistency of the test, while satisfactory, does show up some “outlying” items. Items 1-5 and 50-67 lie further away from, and are therefore less closely associated with the measurement of a single factor (being further away from the zero line on the scattergraph, Figure 1, below):

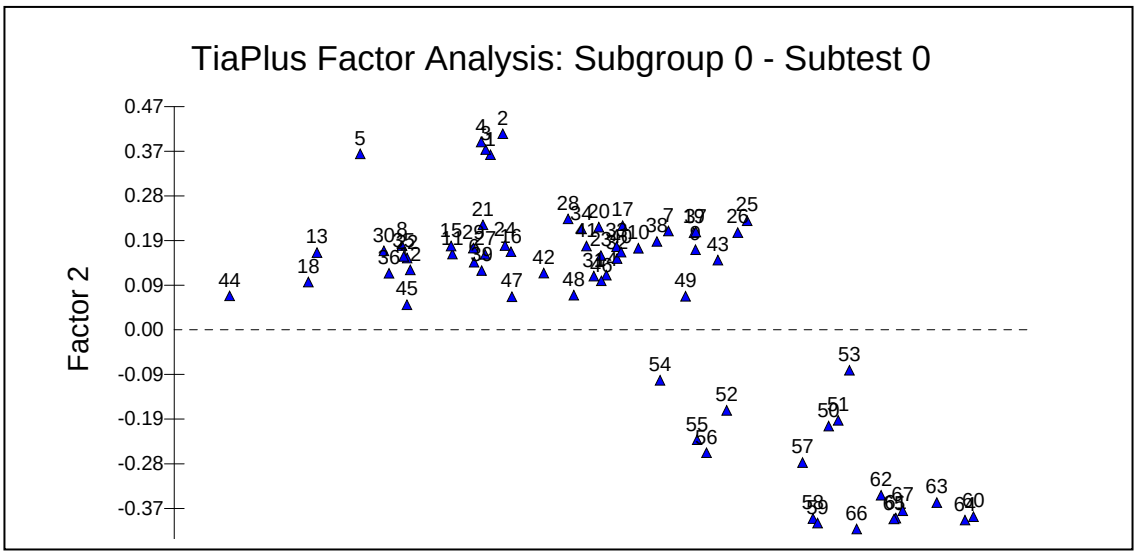

Figure 1: Measures of homogeneity and heterogeneity in TALL 2008

The test designers have a choice: either to do away with these items, representing two subsections of the test, viz. *Scrambled text* (items 1-5) and a modified cloze procedure called *Text editing* (50-67) in order to enhance the technical consistency (reliability) for the instrument, or to accept the heterogeneity of what is being measured. In the end, the test designers chose to include them, mainly on the basis of their appropriateness (relevance) to the construct being measured, arguing that for an ability as richly varied and potentially complex as academic language ability, one would expect, and therefore have to tolerate, a more heterogeneous construct.

The thesis of this paper, to be more closely examined below in the last two sections, is that the theoretical account of such trade-offs is made easier if done within a robust conceptual framework for the whole of applied linguistics.

The aim of this paper is thus to use TALL and TAG as examples of how an alternative conceptualisation of certain foundational dimensions of language testing, and, by extension, applied linguistics, may be achieved. It presents an alternative to two other fundamental, and fundamentally divergent, perspectives on the issues: those within Messick's sphere of influence, and that introduced by Bachman and Palmer (1996). I therefore turn to these first, before returning to the articulation of the alternative.

The notion of the "most fundamental" consideration in language testing

Consider two different conceptualisations of the answer to the same question: what is the ultimate foundation of language testing? The first, one of the most oft-quoted (Pitoniak 2008; cf. too Messick 1980, 1988), comes from the American *Standards for educational and psychological testing* (American Educational Research Association 1999: 9), and reads:

(1) Validity is … the most fundamental consideration in developing and evaluating tests.

Messick's influence is more than evident in this first definition. The second comes from a less frequently quoted, but themselves a not uninfluential pair of testing experts, Bachman and Palmer (1996: 17; also Bachman 2001: 110):

(2) The most important consideration in designing and developing a language test … the most important quality … is its usefulness.

There is no doubt that these two perspectives are essentially divergent. But it is remarkable that one of Bachman's most influential books on the subject of language testing carries a title – *Fundamental considerations in language testing* (1990) - that derives almost word for word from the first definition above, though Bachman (1990: ix) himself justifies his choice of title with reference to two works by other authors, Carroll and Lado. The divergence between (1) and (2), however, is very difficult to interpret, and not seriously or prominently discussed. As Fulcher and Davidson (2007: 15) observe: "The notion of test 'usefulness' provides an alternative way of looking at validity, but has not been extensively used in the language testing literature". Yet it is an important difference of opinion, not the least because influential test making enterprises, such as the Princeton-based Educational Testing Services, continue to make use of the first, as happened recently in the case of Pitoniak (2008). But to understand it easily takes one into the realm of speculation.

One line of speculation is that the divergence is never foregrounded because Bachman and Palmer's notion of test usefulness incorporates validity, as in the following model they propose (1996: 18):

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

Figure 2: Bachman & Palmer's model of test usefulness

The incorporation thus masks the divergence. Why do Bachman and Palmer (1996) themselves make nothing of the difference, however? One reason may be that the nature of their book is explanatory rather than polemical; they may wish to save their (uninitiated) reading audience the agony of weighing up arguments for and against certain opposing views. Another plausible explanation for me is that the lack of prominence given to the divergence of their views with the orthodox one lies in the massive influence of the views of Messick and the institutional base that he represented. This is certainly what Fulcher and Davidson (2007: 15) imply, and McNamara and Roever (2006: 248) also do not mince words about just how

powerful Messick's work was: "hugely influential" is the quality they ascribe to this.

Because there is both divergence and congruence between definitions (1) and (2), we need to consider not only where they differ, but also what they share. The degree of congruence of Bachman and Palmer's (1996) views with Messick and those influenced by him (such as the authors of definition [1]) lies in three main conceptual notions.

First, their understanding of validity coincides with that of Messick in that they give precedence to construct validity. At the same time, their definition of validity (4, below) is essentially the same as that of Messick (1980: 1023; cf. too 1981: 18) in the following pronouncement (3):

> (3) Test validity is … an overall evaluative judgment of the adequacy and appropriateness of inferences drawn from test scores.

They say:

> (4) 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).

In this definition, they echo many others working from the foundations laid by Messick: cf. the following statement by Kane (1992: 527):

> (5) Validity is associated with the interpretation assigned to test scores rather than with the scores or the test.

Strictly speaking, Kane's definition (5, above) is not wholly aligned with Messick's. Messick also held (1989: 14) that "the properties that signify adequate assessment are properties of scores, not tests", so that what is interpreted or judged are scores, while Kane's definition appears to exclude scores from possessing the property of adequacy. But, given a little latitude, they no doubt constitute a fair degree of congruence.

The second convergence between (1) and (2) is more subtle. In the concept of test usefulness, Bachman and Palmer pick up on a point that is less well developed by Messick, but nonetheless goes back to his views. Messick speaks, for example, of predictive and diagnostic "utility" (1980: 1015) or predictive "efficiency" (1980: 1017; 1021; 1989: 20), and at least implies an awareness of utility in his notion of the "instrumental value of the test … accomplishing its intended purpose" (1980: 1025; also 1981: 10, 11, 12).

Despite this degree of congruence, one is tempted to think that in promoting test usefulness to the most important consideration in language testing, Bachman and Palmer (1996), instead of engaging head on with all too influential a notion,

have bypassed it by introducing a plausible new idea. The critical question to ask in such a case would be: what then prevents the development of equally plausible arguments that promote a third idea, other than validity and usefulness, to prime position?

This brings me to the final issue to be considered in this section, a point that depends on the third line of convergence between Bachman and Palmer's (1996) views and those of Messick. Yet in this instance the convergence operates at an even deeper level: though the one calls the most fundamental concept "usefulness" and the other terms it "validity", both parties entertain the notion that an overarching or unified view of language testing is both required and desirable.

Messick's "unified view" (1981: 9) is normally described in the form of a by now over-familiar matrix. In Messick's own formulation (1980: 1023; 1989: 20), his perspective can be described in terms of various aspects of test validity, as follows:

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

Figure 3: Messick's "Facets of test validity"

The most intelligible interpretation of this widely quoted representation is probably that of McNamara and Roever (2006: 14; Figure 4, below; for another, cf. Davies & Elder 2005: 800):

	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?

Figure 4: McNamara & Roever's interpretation of Messick's validity matrix

However, following the texts of the Messick (1981: 10; 1980: 1023) formulations more closely, and turning some of the terms around so that we make a

small adjustment to the matrix, another representation may be possible. Here is one such a possible reinterpretation (Figure 5):

	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

Figure 5: The relationship of a selection of fundamental considerations in language testing

This matrix (Figure 5) can be read as a number of claims about or requirements for language testing, as follows (left to right, top to bottom):

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

(7) 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.

(8) 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.

(9) 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.

What is important to note, however, is that, while the representation still follows Messick's argument, the matrix in Figure 5 is by no means a "validity matrix," as the original claims to be. Nor are the statements derived from it ([6] to [9] above) solely about validity; while obliquely related to the technical power of a test, they rather articulate the coherence or systematic fit of a number of concepts relating to language testing. To some, given the tidiness of the conceptual representations in Figures 3 and 4 above, comments by observers like McNamara and Roever (2006: 249) that "validity theory has remained an inadequate conceptual source for understanding the social function of tests" may therefore come as a surprise. To any disinterested observer, however, it should be evident that the more appropriate labelling of Figure 5 must be that it is a representation of

the relationship between a select number of fundamental concepts in language testing. Once again, we therefore have to question why everything has been subsumed under "validity." Surely concepts like technical adequacy, appropriateness, the technical meaningfulness (interpretation) of measurements (test scores), utility, relevance, public defensibility and the like must be conceptually distinguishable to make sense?

This is also why the swipe that McNamara and Roever (2006: 250f.) take at Borsboom, Mellenbergh and Van Heerden (2004) – that their intention is "to take the field back 80 years" – rests upon a misunderstanding of their critique of validity theory, specifically the form in which it has been outlined by Messick. The point is that if one does not deliberately distinguish what is conceptually distinct, the distinction so avoided subsequently obtrudes itself upon the conceptual analysis. A good illustration of this is how, when under the influence of Messick, and before him of Cronbach (cf. McNamara & Rover 2006: 10), language testing experts steer clear of saying a test "has" validity (since the orthodox view of validity is to define it as "a judgment of the adequacy and appropriateness of inferences drawn from test scores" – Messick 1980: 1023 – and not to ascribe it to the instrument itself) the term either surfaces in an unguarded moment, or reappears as a synonym or synonymous concept. So, for example, McNamara and Roever themselves continue to speak about the validity of a test (2006: 17), or to assume that a "test is … a valid measure of the construct" (2006: 109), and to speak about "items measuring only the skill or the ability under investigation" (2006: 81) – and not about the interpretation of the scores derived from these items. The avoidance of ascribing validity to a test 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). 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, i.e. yields the intended measurements? But causes and effects, and the relationship between causes and effects in the field of testing, are analogical technical concepts, i.e. 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 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. Therefore even those sympathetic to the orthodox view will continue to speak of the "effectiveness" of the use to which a test can be put (Lee 2005: 2), or of a test being "valid in a specific setting" (2005: 3), or that we may investigate through verbal protocols the consequences of a test since these "should be considered valid and useful data in their own right", or provide contorted formulations so as not to use the term 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" (2005: 4).

The way that the other traditionally distinguished "types" of validity return – though now in the fashionably orthodox guise of "sources of evidence" – when hypotheses are formulated to build an argument justifying the interpretation of tests scores is a final point in question. As Davies and Elder (2005: 798) observe,

> in spite of the unitary view now taken of validity, it has to be operationalized through the usual suspects of content and construct validity, concurrent and predictive validity … To the usual suspects we need to add reliability; some will want to include face validity. However, the joker is the validity … called consequential validity.

A good example of "the usual suspects" rejoining the parade is to be found in the validation procedures and methods discussed by Alderson and Banerjee (2001) as being appropriate for studying test impact and washback.

The discussion above should alert us to be critical of notions that subsume others. Conceptual clarity is essential, indeed a necessary condition for test design, as I hope to illustrate below (and cf. Davies & Elder 2005: 799 for what they term a less than charitable view of Messick's failure in this respect). The requirement for conceptual acuity for the sake of an improved designed instrument is not served if we conflate concepts.

Help from the neighbours: the notion of validity in another field

A look at how questions of force, as well as of cause and effect, are treated in other fields may give us some insight into the difficulties associated with the conceptualisation of validity in language testing, by considering specifically how analogical concept-formation is theoretically accomplished elsewhere. For the technical validity of a test is indeed an analogical technical concept, linking the technical aspect of experience to the physical sphere of energy-effect. I selected the field of jurisprudence, since the problem of causality has been of specific interest there. Moreover, there was a surge of interest in issues of juridical causality during more or less the same period (1950 to 1990) within legal philosophy, as was the case with testing, if the discussion in Hommes (1972: chapter IX) is anything to go by. My choice of Hommes's work for comparison is deliberate for another reason: following in the conceptual footsteps of his predecessor, Dooyeweerd (1953), his treatment of juridical causality is done within the same theoretical and philosophical framework that I am using here.

The domain of law, moreover, as McNamara (2003: 467) has pointed out, is similarly characterised by the production and weighing of evidence, the defensibility of inferences, and the like.

What one finds in Hommes's (1972) discussion of force and causality in the field of jurisprudence is indeed highly instructive. The concepts that legal philosophers at the time appear to be grappling with are notions such as the juridical power of a multiplicity of legal norms to effectively regulate legal

relations by connecting juridical consequences to valid, objective juridical facts (Hommes 1972: 151f.). In the same way that test theory may connect the validity of a test to its technical consistency without subsuming the one under the other, Hommes (1972: 151) observes that legal power rests upon juridical constancy, though cannot be equated with it. This is not much different from the way that Davies and Elder (2005: 796) phrase the relationship between the technical consistency of a test and its validity:

> Reliability, we say, is necessary but not sufficient… reliability appears as a separate, parallel (if junior) quality of a test, supporting validity but somehow independent of it.

In addition, in the same way that test theory claims that inferences must be made from the objective scores obtained during the measurement, Hommes (1972: 160) speaks of the actions of competent legal organs that give effect to certain legal conditions. We read of dynamic legal processes in which valid legal facts are interpreted in order to relate legal causes to legal effects, as well as of the ability to foresee, within the bounds of reason, the legal consequences of certain juridical actions, and about the adequacy (1972: 174) of the evidence that is mustered. We are warned against entertaining a view of legal action that denies the typicality of human freedom (1972:177), and, by extension, the way that actions need to be ascribed to persons, and accounted for by them.

There is no doubt that the debates and the conceptualisations have much in common across the two fields. If we substitute “legal” or “juridical” with “design” or “technical”, the resemblances are almost uncanny. How such a discussion may relate further to the problems faced by test theory can perhaps best be illustrated by means of an example. Suppose that a burglary occurs at the premises of owners who have indemnified themselves against just such an occurrence by taking out insurance. The criminal act constitutes the cause of the objective loss of property, which is the effect; the objective loss of property in turn becomes the legal cause for the legal effect of restitution becoming appropriate. However, should the burglary not have happened at all, but merely have been fraudulently constructed by the insured to make it appear as if there had been a loss, their claim becomes invalid. One can in such a case no longer connect the objective state of affairs to any legally required restitution by the insurer. If, on the other hand, there is sufficient evidence that the burglary did in fact occur, and that they had therefore experienced an objective loss of property, they would have a valid claim. Of course, the validity (the factual legal force) of the claim needs to be subjectively ascribed to the condition of their property after the insured occurrence has taken place, but that merely states the obvious: that subjective juridical action of necessity involves interpretation of certain objective juridical states of affairs. It does not take away the objective validity of those affairs. As one can see from the alternative (the burglary was conceived for the purpose of instituting a fraudulent claim), there are other objective states that are invalid and inadequate grounds for legally receiving restitution.

Could it be that one might make the same kind of argument for the validity of a technical instrument such as a test? In the next section, I turn to a possible way out of the dilemma.

Subjective and objective components of validity

If validation is a process, what would a reasonable result of that process be? Is it inconceivable that the process of producing evidence will confirm that, to the best of the test designer's knowledge, the test has the desired effect, i.e. it yields certain objective scores or measurements? As Davies and Elder (2005: 797) observe, through acquiring over time, and through repeated validation arguments, an adequate reputation, any test must eventually present a principled choice to those wishing to use it, and that choice can be attributed to little else than its known validity. Since one test may provide more useful inferences than another, they conclude (2005: 798), "it is not just a trick of semantics, therefore, to say that one test is more valid than the other for a specific purpose." This is perhaps also why more recent comments made by Bachman and Palmer (1996: 19f.), for example, continue to speak, contrary to Messick, of reliability and validity as qualities of tests; in their view the validity of test scores become the (objective) basis for subsequent inferences about the results.

The scores of tests are indeed (technically qualified, and theoretically grounded) objects. *On their own*, and that is the point where Borsboom *et al.* (2004) may misunderstand the proponents of orthodox validity theory, they are, however, meaningless.

This is so for all of human endeavour. Objects do not have meaning separate from human interpretation: all objects find themselves in an unbreakable subject-object relation. Even terminally discarded objects that we find on rubbish dumps attest to their former attachment to and use by human subjects, and in their discarded state once again come to stand in subject-object relations, though this time lumped together with others as objects of waste that have to be subjectively treated as such. So objective evidence in the sphere of jurisprudence needs interpretation by juridical subjects (competent legal organs) in order for them to become meaningful, useful, accessible, transparent, and the like. Similarly, in the field of language testing, the technical measurements that are made with the aid of a specifically designed instrument need to be interpreted, and the way that the interpretation is accomplished needs to refer to evidence that relates this interpretation of the scores to the theoretical rationale for the design of the instrument (Messick 1980: 1014). The subjective interpretation is made on the basis of objective measurements; there is a technical subject-object relation at play here. In the relative section below, I shall return to this distinction.

Though this is nowhere stated in the literature under discussion here, the real threat that orthodox validity theory sees lies in the modernist belief in "scientific facts." Modernism views measurements arrived at through rational means as self-sufficient, universally and always "true", and therefore valid *per se*. Language testing has indeed moved beyond a belief in the self-sufficiency, the autonomy, of

science (cf. Hamp-Lyons 2001). Yet, as technical applied linguistic instruments, tests are indeed theoretically grounded. Hence the importance attached, and correctly, in the orthodox view, to construct validity, the "rational basis" (Messick 1980: 1015, 1013) of such a measurement tool. I return below to how this relates to, and can be theoretically clarified in terms of a foundational framework.

To conclude this section: the distinction between the subjective process of validation and the objective validity of a test is an essential one, which seems to be forgotten in some of the discussions referred to above. Viewed subjectively, validity is the achievement of validation. Viewed objectively, it is a function of test scores. If the latter were not the case, we would not have been able to ascribe or impute an adequate interpretation to such scores, for those scores would have lacked not only validity, but also interpretability.

Constitutive concepts in language testing

This section will deal with the first of two sets of analogical concepts in applied linguistics and its subfield, language testing, which together constitute an emerging theoretical framework for the discipline. By analogical concepts I understand the theoretical conceptualisation of the relationship between a unique mode of reality, in the current case the technical dimension of our experience, and another unique function, such as the physical sphere of energy-effect. As we have remarked above, this connection is captured in the concept of technical validity. In the same way, the relationship between the technical aspect of, say, a test design, and the kinematic dimension of reality is echoed conceptually in the notion of technical consistency, since the kinematic is characterised by uniform or constant movement. All of these connections between uniquely different, but related modes of reality, echo the idea that, while there are unique dimensions to our experience, none of them is absolute, and each of them is necessarily related to all the others. In theoretical concept-formation, we analytically approximate these relationships.

Among the many modes of experience in which they function, applied linguistic instruments have two critically important terminal functions. In addition to their leading technical function, applied linguistic instruments also have a founding function, the dimension in which their technical design finds its base. Like other applied linguistic objects, tests have as their qualifying or leading function a deliberate technical design, as well as a founding function, their basis, in the analytical aspect of our experience. Phrased differently, they find their foundation in the theoretical or analytical mode, which allows the design of the measuring instrument or test to be theoretically defended on the basis of rational argument. In its turn, and characteristic of work done in this mode, such a rational justification is based on adequate evidence. Among the many modes of experience, therefore, two stand out as terminal or critically important functions:

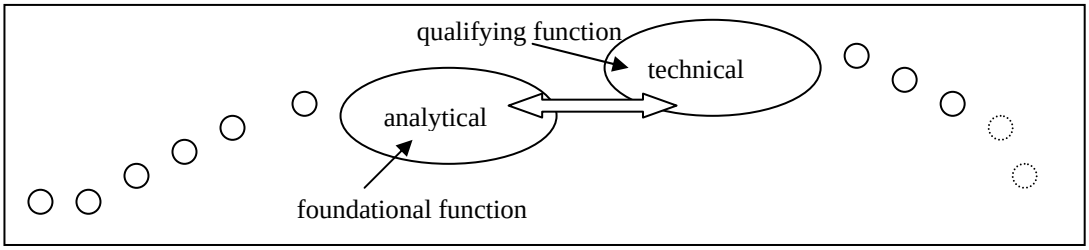

Figure 6: Terminal functions of an applied linguistic design

The relation between the leading, technical function of a test and its founding, analytical function is reciprocal. That is, in the design of an applied linguistic instrument, the technical imagination of the designer indeed leads the whole endeavour, but at some point in the design process the development of the artefact must open itself up to critical modification and even correction by analytical and theoretical considerations and rational argument. As Schuurman 1972: 46) phrases it:

> Het theoretische vlak is de *basis* van het technische ontwerpen. De technische verbeelding neemt daar haar uitgangspunt. Deze verbeelding of fantasie is het zwaartepunt van het ontwerpen; zij heeft een geobjectiveerd ontwerp tot richtpunt. Het ontwerpen voltrekt zich dientengevolge als een wisselspel van fantasie en theorie, en het mondt uit in de *intentionele* technische vorming van een ontwerp… (freely translated from the Dutch: The theoretical level is the *basis* of technical design. The technical imagination (of the designer) takes that as its starting point. This imagination or fantasy is (however) the gravitational centre of designing; it has as its focus an objectified design. As a result, we find in designing an interplay between fantasy and theory, which eventually achieves the *intentional* technical formation of a design…)

As in the foundational analysis given here of the technical dimension of experience, I follow the views put forward by Schuurman (1972; cf. too Van Riessen 1949) regarding the process of design, as in Figure 7 below. Schuurman (1972: 404) identifies three stages; for the sake of applying his observations to the field of applied linguistics, I have added two more.

1) In the first stage, there is a growing awareness of a language problem, leading to its more or less precise identification. At this stage, there is nothing "scientific" about what is being done. In the case of TALL and TAG, for example, the problem was identified by university administrators and authorities, and had to do with concerns for certain students being given access to university education, in a time of massive enrolment growth, who were presumed to have risk in terms of language. In many (though, as it turned out, certainly not in all) cases, these were students for whom the language of instruction was an additional, in other words not a first, language.

2) In the second stage, the designers bring together their technical imagination and the theoretical knowledge that potentially has a bearing on the problem identified. In the case of TALL and TAG, most of this initial design work involved bringing together expertise in language test development, but especially language course design, since the design of language intervention was thought to be critically important.
3) There is an initial (preparatory) formulation of an imaginative solution to the problem. At this stage there may be some experimentation with designed materials. During this phase, in the case of the University of Pretoria, the alignment, for example, of test construct with the outcomes of the designed curriculum and the enhancement of learning and language development had not been thought through properly, and anticipated the next phase of the design. In that phase, finally,
4) a theoretical justification is sought for the solution proposed and designed. The theoretical considerations underlying the development of TAG and TALL, and in particular the selection of an appropriate and defensible construct, have been set out in Van Dyk and Weideman (2004a). As a result of this analysis, the desirability of aligning this construct with the learning and the teaching outcomes of the designed intervention that followed (Weideman 2007a: esp. xi-xii) became increasingly more prominent and necessary.
5) Finally, if either the theoretical enquiry as to the adequacy and appropriateness of the solution, or the piloting of the preliminary product or products, shows up initially unanticipated weaknesses and flaws in the design, the designers, armed now with both more theoretical and practical information, re-apply their minds and imagination, and redesign the solution. In the case of TALL and TAG, the articulation of the construct in the form of a blueprint, and the matching of the components of the construct with the blueprint of the test (Van Dyk & Weideman 2004b), as well as the specifications of item types (potential subtests) also formed part of this final phase. In sum, during this phase of the design the technical solution may again be modified in order to align it more closely with an adequate and appropriate theoretical justification, and there may be further piloting and trial runs of both the test and the designed intervention.

A diagrammatic representation of the process would therefore look something like this (Figure 7):

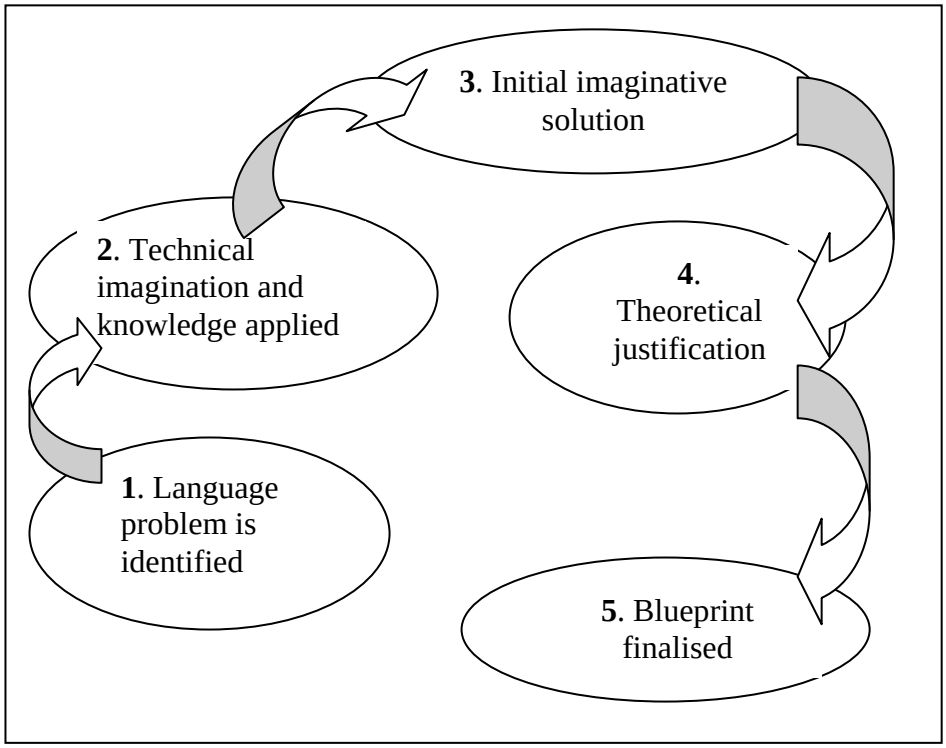

Figure 7: Five phases of applied linguistic designs

What is critical for the further development of the argument of this paper, and the description of the foundational framework that underlies it, however, is the observation that the technical qualifying function of an applied linguistic instrument, such as that of a designed test, interacts not only with the analytical dimension of experience, but connects and is connected with all other modes. For example, the analogical relation between the technical and the numerical becomes evident in notions of using a unity within a multiplicity of technical conditions and a diversity of factual (empirical) "sources" of technical evidence (a dense 'mosaic', in Messick's terms) to achieve the necessary result in the technically qualified process of validation. This analogical relation is probably the conceptual foundation for the proposal to have a "unitary" or unified approach to test validation. So, too, the analogical relation between the technical and the kinematic aspect of experience, which is originally characterised by consistent movement, yields the concept of technical consistency, or what in testing parlance is called reliability. Moreover, as is the case with some other technically achieved analyses that proceed from analogical concept formation, such a concept as the technical consistency of a test can be expressed in terms of numbers or dimensions, in which the links with both the numerical and spatial aspects are once more evident. Of a two-dimensional representation of the reliability of a test the scattergraph plotting its factor analysis (as in Figure 1, above) is a good example. But the measure(s) of the reliability of a test can of course also be expressed in terms of a purely numerical index, such as Cronbach's alpha, or Greatest Lower Bound. For TALL,

one of the two tests of academic literacy under discussion here, the average reliability index across five different institutional and randomly selected administrations looks as follows (Table 1):

Table 1: Reliability indices for TALL (2004-2008)

Version of test (administration)	Reliability (Cronbach's alpha)
2004 (Pretoria)	0.95
2005 (Northwest)	0.94
2006 (Pretoria)	0.94
2006 (Stellenbosch)	0.91
2008 (Pretoria)	0.94
Average	**0.94**

The fact that all of such technical measurements can be expressed in numerical terms is a clear indication that they have their basis in some of the "natural" dimensions of experience, such as the numerical, the spatial and the kinematic.

Another constitutive concept for test design is made possible, as I have noted above, by the connection between the leading technical function of the design and the physical sphere, which is characterised by the operation of energy-effect. This allows us to conceptualise the technical power or force of the designed instrument, or that which in testing terminology has been articulated in the notion of validity. Messick's notion of "adequacy" is, I would argue, closely related to that original analogical understanding of the technical validity of the test, of the test having the technical force to "measure what it is supposed to measure." Clearly, since this is an analogy that lies in the foundational direction (see Figure 8, below), it is also in that original understanding a restrictive notion, and one that anticipates a broadening of that understanding – in the case of language testing a broadening that relates to the regulative conditions for technical instruments, as will be demonstrated below.

The connections between the leading technical aspect of the test design and those modes preceding it, such as in Figure 8 below, are analogical relations that are made in a foundational or constitutive direction. The analogical relations between the technical design and those dimensions preceding the technical are mediated, therefore, through the founding, analytical function of the design. This is why, in the theoretical justification that is sought for the design of a test, the original understanding of test validity – a reference to the connection between the physical aspect and the technical – comes to be re-interpreted as the validity also of the hypothetical ability that is being measured by means of the designed instrument, the language test. This conceptual reinterpretation underlies the identification by Messick and those following him of the "construct validity" of a test as such a critically important, indeed necessary or constitutive condition for test

design. In language testing, the theoretical justification of the technical design is done on the basis of a hypothetical or theoretically informed and articulated construct of what is being tested. The degree of theoretical defensibility achieved has come to be called – perhaps somewhat confusingly – the "construct validity" of the technical instrument. Figure 8 below is a graphic representation of these relationships, which form the conceptual prompts for foundational concepts relating to applied linguistic designs.

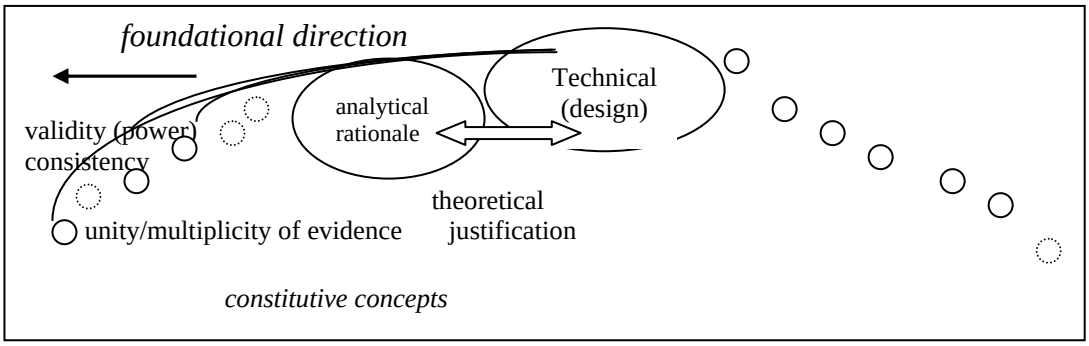

Figure 8: Foundational concepts of applied linguistic designs

All four of the examples given so far – the technical unity of multiple sources of evidence (or technical systematicity of a test), its technical reliability, validity, and rational justification, are, in this model, foundational or constitutive applied linguistic concepts. In that sense, they are also necessary requirements for tests, an observation which is borne out not only by the history of language testing and the way that these concepts have been treated in the mainstream literature, but also by the enduring interest in them despite their conflation, in recent discussions, that consider all to be aspects of "validity". It is important to note that each of these "necessary" or foundational concepts yields a (technically stamped) criterion or condition for the responsible use or implementation of the technical instrument (cf. Schuurman 1977, 2005; Davies & Elder 2005: 810f.). This is the reason why we say that tests should be reliable, valid, and built on a theoretical base that is defensible in terms of a unity within a multiplicity of sources of evidence.

Technical subject-object relations revisited

What is also important to note is that this kind of formulation of the necessary conditions for responsible language test design is often done from the perspective of the instrument being designed, the technical artefact that is the product, or technical object, of the design. What one should never forget, of course, is that like all other objects, language tests function in a technical subject-object relation. This means that, though the instrument we use will yield certain results, in intentionally producing certain objective effects (the test scores), that technical power to produce these results does not mean that such measurement outcomes can exist autonomously, and have meaning on their own. Indeed, as recent discussions, specifically of definitions of "validity", have shown, the objective effects that tests

intentionally, i.e. by design, produce, can only have meaning when they are interpreted by technical subjects – chief among them probably being the users of test scores, the administrators, who need to base decisions on such technical effects.

This places a tremendous responsibility on test designers, who have to take steps to ensure that when their test is used, its results are interpreted in a technically adequate and appropriate way. This, in my opinion, is the enduring meaning of Messick's work (1980, 1981, 1989) that has been referred to above. However, my introduction here of the distinction between technical object (the adequate or valid measurement or test score) and technical subject (who, by conducting a process of validation, should make an appropriate interpretation of this) should go some way towards eliminating the confusion between the force of the objective measurement and its subjective technical meaning.

Regulative conditions for language testing

In addition to having foundational concepts that relate the technical dimension to preceding aspects of experience, this guiding aspect of applied linguistic designs also anticipates, in the other direction, its linkages with the lingual, social, economic, aesthetic, juridical and ethical aspects that follow it. These forward-looking, anticipatory links between the technical, qualifying function of the test design and these other aspects yield the ideas of technical articulation, test implementation or use, its utility, its alignment with, for example, learning and teaching language, its public defensibility or accountability, and its fairness or care for those taking tests. The following figure (Figure 9) sets out these regulative ideas within the model I have been referring to:

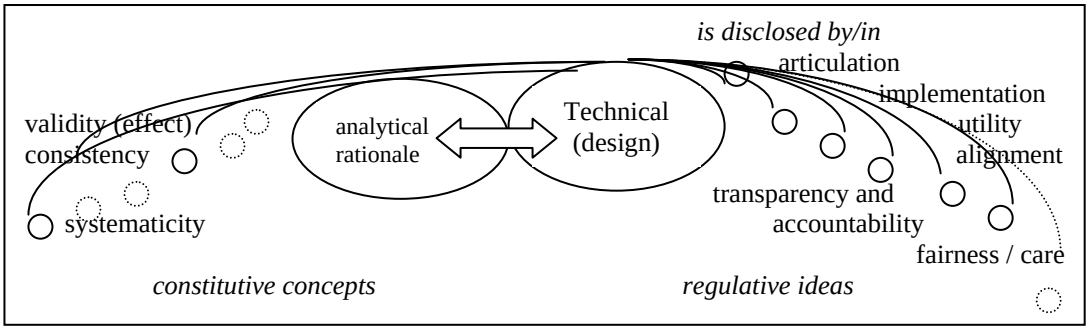

Figure 9: Constitutive concepts and regulative ideas in applied linguistic designs

In the case of the design of a language test, the particular technically qualified applied linguistic object that we are using as example, the relationship between the leading technical function of its design and the lingual dimension becomes evident when we encounter the articulation or *expression* of that design in the blueprint of a test. The further articulation of such a blueprint in the form of a set of specifications for the test in general, as well as for its subsections, item types and items in particular (Davidson & Lynch 2002, Van Dyk & Weideman 2004b) similarly depends on the relation between the technical and lingual aspects of

experience, as does the appropriate technical *interpretation* of the test scores (somewhat confusingly also defined as "validity" in the current orthodoxy).

Test designs lead to the technical production of the test, that includes test development, experimentation (cf. Schuurman 1972: 46), sometimes in the form of trial runs or test piloting, redevelopment and any number of successive stages (cf. Figure 7, above). Eventually, all of this activity leads to the implementation of a test when it is administered to a population of prospective testees, the persons whose language ability is being measured. This administration ties the technical instrument to its social context and use, the importance of which is evident in ongoing discussions in language testing circles on the social dimensions of test use, or the consideration of test impact, as it is called. It goes without saying that a consideration of how tests impact, through their use, on larger social and education systems (Bachman & Palmer 1996: 29f., 34; Shohamy 2001a) is critically important. In the same way that, in the framework that we are using here, the technical validity or force of a test crucially depends on its reliability, the technical use of tests in the wider social world depends on an appropriate interpretation of test scores, which is an idea that conceptually connects, as we have noted, the technical and lingual aspects. The employment of the term "appropriate" in relation to "interpretation" and "use" in the work of Messick and others anticipates the meaning that test effects (results) will have in a specific, defined social context – hence the call for the validation process of the designed instrument always to be contextually appropriate.

It also follows from this analysis that there rests a potentially burdensome responsibility on test designers to anticipate as best they can the possible negative and positive consequences of tests, and the way that scores are indeed either inappropriately or appropriately interpreted. These are regulative conditions for test design.

The connection between the technical function of a test and the economic dimension becomes evident when we consider the idea of utility. Since the design and production of any technical object implies the availability of technical means to achieve technical ends or purposes, it follows that producing and using language tests should consider, for example, the resource implications of using one design of a test instead of another. So, for example, though this is not the only reason for doing so, most tests of academic literacy try to ensure "face validity" – the apparent relevance of a test for laypersons – by including substantial sections on academic writing, which have to be handmarked. The *Test of Academic Literacy for Postgraduate Students* (TALPS) developed at the University of Pretoria is one good example. It is often the case that such subsections consume inordinate amounts of expert time (cf. Davies & Elder 2005: 806f.), a very scarce technical resource, and may undermine the technical consistency of a test. This is one of the reasons why, in the case of TALL and TAG, the two undergraduate tests, the test designers have come up with arguments for justifying the use of multiple choice questions only, and ensuring that there is an adequate rationale for questions in such a format also testing (productive) writing. Bachman and Palmer (1996: 35f.)

categorise such considerations under the rubric of “practicality”, one of the components of their overall model (Figure 2, above) of test usefulness.

The technical utility of a test is the first regulative idea relating to language test design that confronts us with the notion of having to weigh up various factors in deciding on and using tests. Such weighing up generally brings to the fore one or more design decisions, as we have seen above (Figure 1), in the choice that the designers of TALL and TAG face between the appropriateness and relevance of a rich and varied definition of academic literacy as construct, and the technical homogeneity of the test. Similarly, there are trade-offs possible between reliability and utility (a longer test is potentially more reliable, but may consume too many scarce resources; cf. too Davies & Elder 2005: 806), or between efficiency and effectiveness, or even between conflicting political interests. Weideman (2006: 84) notes that the weighing up that precedes decisions on trade-offs is related first to the analogical relation between the technical, qualifying aspect of the design and the economic, but, second, also to the aesthetic, juridical and ethical dimensions of the test design. In all of these considerations, one encounters a number of fundamental concepts that allow us to conceptualise the foundations not only of the subfield of language testing, but of the whole of applied linguistics.

Especially in the regulative ideas of transparency and accountability, that relate the technical, qualifying dimension of the designed test to its juridical aspects, and the ethical idea of a test being so designed that it has as a consequence the care for and benefits of others, we have prime examples of regulative or sufficient conditions for language test design. A test has to be accessible and the basis for its design transparent politically, i.e. in terms of its public import and power. Its juridical dimension is also evident, of course, in its public defensibility. I fail to see, however, that where it unintentionally disadvantages others, that constitutes a “threat to the test’s validity” (Davies & Elder 2005: 808), instead of a breach of its ethical concern with the rights and interests of others, unless one uses one analogical technical concept (validity) as the privileged vantage point from which to survey all fundamental considerations in language testing. Of course, as the analysis above has shown, the constitutive concept of technical validity can be enriched by articulation of the theoretical idea or rationale for a test, as well as by the subjective technical interpretation of the results of a test. It can also be further enriched and opened up when one considers the notion of its social results or impact. But such enrichment does not constitute either a basis for privileging the concept, or for subsuming everything under it. It merely points to the unfolding of or opening up of the design to the regulative conditions for language testing.

I believe the value of the current framework lies in its separating out what is conceptually distinct, and by doing so enriching our theoretical understanding of the constitutive and regulative, necessary and sufficient, conditions for language testing. A further assessment of its benefits is made in the final section.

The function of a foundational framework for language testing

McNamara & Roever (2006: 253) are essentially correct in observing that, in order to see the current concerns of language testing more clearly, we "need an adequate social theory to frame the issues"; what is even more relevant to me in terms of the discussion above, is their observation that we also need "to break down the walls between language testing researchers and those working within other areas of applied linguistics, social science, and the humanities generally" (2006: 254). In order to do this, approaches to language testing in particular, and to applied linguistics more generally, need to tune into foundational considerations, the basic concepts and notions of their field. This is a philosophical undertaking, constituting an examination of the foundations of the field.

It is not a coincidence, I believe, that Messick himself (1989: 30f.) turned to the "philosophical foundations of validity and validation", referring to the perspectives of Leibniz, Locke, Kant, Hegel and Singer in order to gain clarity in this regard. It is a pity, nonetheless, that we see too little of such foundational discussion within applied linguistics and language test design. One less beneficial effect of such neglect is that much of our recent discussion within the field of language testing, with the possible exception, perhaps, of parts of Davies and Elder (2005), has remained firmly attuned to the current orthodoxy, leaving that essentially unchallenged.

This paper has attempted to make just such a contribution. Contrary to McNamara and Roever's (2006: 254) idea, quoted above, I would, however, not limit the task of "breaking down the walls" to those separating the human sciences only. As is clear from the analysis given above, in making applied linguistic and language test designs, we need to refer not only to the human or cultural dimensions of our experience – the way that our designs relate to ethical, legal, aesthetic, economic, social and lingual concerns, that yield regulative conditions for those designs – but also to the natural dimensions of experience, its numerical, spatial, kinematic and physical aspects, which give us the foundational or necessary requirements for tests to be adequate.

The function of the framework outlined above, then, is to contribute to the breaking down of disciplinary barriers across the human and natural sciences, along with bringing more sharply into focus certain specific concepts that are too easily conflated by the reigning orthodoxy in language testing.

Acknowledgements

I would like to record my heartfelt gratitude to my colleague, Jurie Geldenhuys, who sacrificed large chunks of his research leave to be of assistance to me in coming to an understanding of the concepts and discussions referred to here. I wish to thank, too, Avasha Rambiritch, colleague and doctoral student, who kept a steady stream of noteworthy literature directed our way. I would also like to thank Alan Davies in advance for managing the introduction of the colloquium where this paper will be presented. By continuing to ask me probing questions during a fruitful visit some years ago, he initially stimulated my thinking and insisted on answers to some very difficult questions, many of which unfortunately remain unanswered. Nonetheless, for that and for the views expressed here, I remain solely responsible.

References

Alderson, J.C. & Banerjee, J. 2001. Impact and washback research in language testing. In Elder, C., Brown, A., Grove, E., Hill, K. Iwashita, N., Lumley, T., McNamara, T. & O'Loughlin, K. (eds.). 2001: 150-161.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education. 1999. *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

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

Bachman, L.F. 2001. Designing and developing useful language tests. In Elder, C., Brown, A., Grove, E., Hill, K. Iwashita, N., Lumley, T., McNamara, T. & O'Loughlin, K. (eds.). 2001: 109-116.

Bachman, L.F. & Palmer, A.S. 1996. *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

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

Davidson, F. & Lynch, B.K. 2002. *Testcraft*. New Haven: Yale University Press.

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 Associates: 795-813.

Dooyeweerd, H. 1953. *A new critique of theoretical thought*. 4 volumes. Amsterdam: H.J. Paris.

Elder, C., Brown, A., Grove, E., Hill, K. Iwashita, N., Lumley, T., McNamara, T. & O'Loughlin, K. (eds.). 2001. *Experimenting with uncertainty: Essays in honour of Alan Davies*. Cambridge: Cambridge University Press.

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

Hamp-Lyons, L. 2001. Ethics, fairness(es) and developments in language testing. In Elder, C., Brown, A., Grove, E., Hill, K. Iwashita, N., Lumley, T., McNamara, T. & O'Loughlin, K. (eds.). 2001: 222-227.

Hommes, H.J. van Eikema. 1972. *De elementaire grondbegrippen der rechtsweteschap: Een juridische methodologie*. Deventer: Kluwer.

Kane, M.T. 1992. An argument-based approach to validity. *Psychological bulletin* 112 (3): 527-535.

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. & Roever, C. 2006. *Language testing: The social dimension*. Oxford: 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. & Braun, I.H. (eds.). 1988. *Test validity*. Hillsdale, New Jersey: Lawrence Erlbaum Associates: 33-45.

Messick, S. 1989. Validity. In Linn, R.L. (ed.). 1989. *Educational measurement*. Third edition. New York: American Council on Education/Collier Macmillan: 13-103.

Pitoniak, M.J. 2008. Issues in high-stakes, large-scale assessment. Presentation to a colloquium hosted by Higher Education South Africa (HESA) and the Centre for Higher Education Development (CHED) of the University of Cape Town, 18 April.

Schuurman, E. 1972. *Techniek en toekomst: Confrontatie met wijsgerige beschouwingen*. Assen: Van Gorcum.

Schuurman, E. 1977. *Reflections on the technological society*. Jordan Station, Ontario: Wedge Publishing Foundation.

Schuurman, E. 2005. *The technological world picture and an ethics of responsibility: Struggles in the ethics of technology*. Sioux Center, Iowa: Dordt College Press.

Shohamy, E. 2001a. Fairness in language testing. In Elder, C., Brown, A., Grove, E., Hill, K. Iwashita, N., Lumley, T., McNamara, T. & O'Loughlin, K. (eds.). 2001: 15-19.

Shohamy, E. 2001b. *The power of tests: a critical perspective on the uses of language tests*. Harlow: Pearson Education.

Shohamy, E. 2004. Assessment in multicultural societies: Applying democratic principles and practices to language testing. In Norton, B. & Toohey, K. (eds.), *Critical pedagogies and language learning*. Cambridge: Cambridge University Press:72-92.

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.

Van der Slik, F. & A. Weideman. 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? Forthcoming in special issue of *Ensovoort*.

Van der Slik, F. & Weideman, A. 2008. Measures of improvement in academic literacy. Submitted to *Southern African linguistics and applied language studies*.

Van der Walt, J.L. & Steyn, H.S. jnr. 2007. Pragmatic validation of a test of academic literacy at tertiary level. Forthcoming in special edition of *Ensovoort*.

Van Riessen, H. 1949. *Filosofie en techniek*. Kampen: Kok.

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

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

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

Weideman, A. 2007b. The redefinition of applied linguistics: modernist and postmodernist views. *Southern African linguistics and applied language studies* 25(4): 589-605.

Weideman, A. 2008. Towards a responsible agenda for applied linguistics: Confessions of a philosopher. *Per linguam* 23(2): 29-53.

Weideman, A. & Van der Slik, F. 2007. The stability of test design: measuring difference in performance across several administrations of a test of academic literacy Forthcoming in *Acta academica* 40(1).