Avasha Rambiritch

http://dx.doi.org/10.4314/jlt.v46i2.7

Challenging Messick: Proposing a theoretical framework for understanding fundamental concepts in language testing

A B S T R A C T As applied linguists, an important part of our work constitutes the design of language courses, language tests and sometimes even language policies. Clearly, these applied linguistic artefacts, especially language tests (which are the focus of this article) have far-reaching, sometimes negative, effects on our students. As applied linguists what is there in the literature on language testing to guide the work we do, to ensure that our designs have some positive effect? What have the experts in the field of language testing presented us with to ensure that important questions related to the social dimension in testing (issues related to transparency, integrity, accountability, fairness and ethics) are not ignored in the design and administration of language tests? What this article will attempt to do is to show that questions about the social dimension of language testing cannot be adequately answered by Messick (1980; 1989a; 1989b), a conventionally accepted expert in the field. Instead these questions can be answered in a "third idea, other than validity" (Weideman 2009: 239), as outlined by Weideman, an idea that does not foreground one concept but rather identifies a number of fundamental considerations for language testing.

Key words: applied linguistics, language testing, social dimension, social defensibility, validity, construct validity, theoretical analysis, framework, constitutive, regulative

1. Introduction

In defining the characteristics of all tests, Davies (1990: 17) states that a test "is intended above all to clarify the difference in the matter under test, in what is being tested (proficiency, aptitude, achievement) among the candidates". This need to "clarify the difference" means, among other things, that comparative figures and data need to be studied. In the field of language testing, this has led to a heavy reliance of applied linguistics on the field of psychometrics.

In fact, according to McNamara and Roever (2006: 1), "…psychometrics became the substrate discipline, prescribing the rules of measurement, and language was virtually poured into these pre-existing psychometric forms." Thus, psychometrics became the basis of language testing, the most important and for some the only way in which a test could be validated. However, language tests need more than psychometric data to be considered valid, not least because "language is rooted in social life and nowhere is this more apparent than in the ways in which knowledge of language is assessed" (McNamara & Roever, 2006: xiv). McNamara and Roever's observation that "a psychometrically good test is not necessarily a socially good test" (2006: 2) is an important one because a core concern here is the social responsibility that test developers have, not just to the test takers but to everyone affected by the test – supervisors, parents, test administrators and society at large.

The aim here is to show that questions related to the social dimension of testing cannot be adequately answered with reference to the work of Messick (1980; 1989a; 1989b), a conventionally acknowledged expert in the field (see Cook, Schmitt-Cascallar & Brown, 2005; Kane, 2006; McCallin, 2006 and Becker & Pomplun, 2006 for a presentation of further arguments about validity and validity evidence). Instead these questions can be answered in a "third idea, other than validity and usefulness" (Weideman, 2009: 239), as outlined by Weideman, an idea that does not foreground one concept but rather identifies a number of important considerations for language testing.

2. The need for a theoretical analysis or justification for applied linguistic designs

It is useful to begin by acknowledging that the field of language testing falls within the scope of applied linguistics. Weideman defines applied linguistics "as a discipline that devises solutions to language problems" (2006: 72). In this view applied linguistics presents the solution in the form of a design or plan, which is in turn informed by some kind of theoretical analysis or justification. The need for a theoretical analysis or justification is outlined by Weideman in a paper entitled 'A responsible agenda for applied linguistics: Confessions of a philosopher' (2007b). He observes here that while applied linguistic work can be so absorbing that very often one does not "take the time to stand back and take stock, or understand fully the disciplinary foundations we stand on" (2007b: 30), this ignorance quite easily sets one up for falling victim to theoretical fashions (2007b: 30). If we see applied linguistics as a discipline concerned with design, we must also "develop a theory of applied linguistics which shows what constitutive and regulative conditions exist for doing applied linguistics designs" (Weideman, 2007b: 29). Weideman explains that this theory or "agenda" (2007b: 29) needs a "certain yardstick" (2007b: 29). He has chosen the term 'responsible' as a measure of this. In addition to this, the discipline of applied linguistics needs a "broader theoretical framework" (2007b: 29), by which he means that one-sided emphases (e.g. on purely empirical determinations of validity) create undesirable design considerations. The theoretical framework he is referring to therefore allows the articulation of "a responsible agenda for applied linguistics" (2007b: 30).

The solutions to language problems that applied linguists design or devise are presented in the form of designs or plans – these could be the designing of language courses or the development of language tests (Weideman 2006: 72). According to Weideman's theoretical framework, the plan presented has two terminal functions: "a qualifying or leading function, and a foundational or basis function" (2006: 72).

Presented schematically:

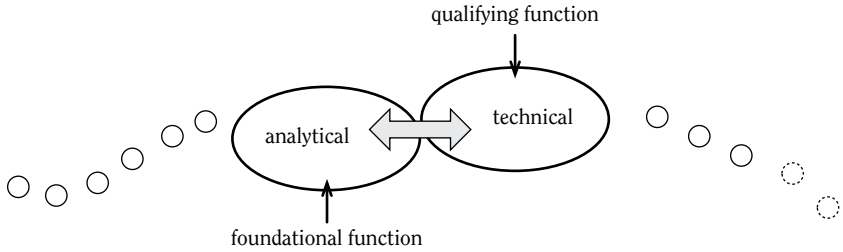

Figure 1: Leading and foundational functions of applied linguistic designs (Weideman, 2006: 72)

According to this schematic representation, the leading or qualifying function of a plan presented as an applied linguistic solution to a language problem is to be found in the technical aspect of design. The plan or design finds it foundational function in, or is based upon, the analytical or theoretical mode of experience. Explained simply, this theoretical framework suggests that:

a. The theory provides a rationale for the design but does not control it;
   and
b. The design therefore takes precedence, not the theory. While the theory is important it does not "prescribe" (Weideman, 2007b: 41) the design.

The context for this argument is provided by arguments against treating technically qualified objects, such as language tests, as mere applications of science. This does not mean that the technical design and development of an applied linguistic instrument, such as a language test, has nothing to do with scientific analysis, but that the role of scientific analysis is to provide subsequent theoretical justification for the design.

Weideman explains that "the context in which such a designed solution is implemented invariably has a social dimension, and that applied linguistic designs have ethical dimensions, since they affect the lives of a growing number of people" (2006: 72). The following observations are relevant:

a. Applied linguistic work should be backed by some foundational framework to ensure that the notions of responsibility and integrity can be articulated in a theoretically coherent and systematic way;
b. In the framework Weideman suggests that the plan or design has a leading or qualifying function and a foundational or analytical function;
c. The design cannot ignore the social and ethical dimensions present in the articulation of solutions to language problems.

3. Defining 'constitutive' and 'regulative'

The employment of the theoretical framework referred to above serves to articulate coherently and systematically issues of responsibility and integrity – as well as to make allowances for other dimensions, such as the social and the ethical. This theoretical foundation is derived from a specific way of looking at the world. Not only is applied linguistics conceived of as a discipline of design, but the outcomes of such designs are characterised as technical objects

by the philosophical framework being employed. Just as we do not exist in isolation, technical products such as tests do not exist on their own. They exist in the technical as well as in other modes of experience – they are not removed from but related to these other modes. Weideman states: "The conviction is a fairly simple one: nothing is absolute, and …, though one may distinguish between uniquely different modes of doing and being, all of these are connected to everything else" (2007a: 599). The technical dimension, in fact, coheres with a number of other dimensions of our existence. These relations yield two sets of concepts and ideas:

1. A set of **'constitutive'** concepts – defined here as "grounded upon". The technical design is grounded upon a set of concepts such as reliability/validity, and these are founding or constitutive concepts, as indicated in the figure below:

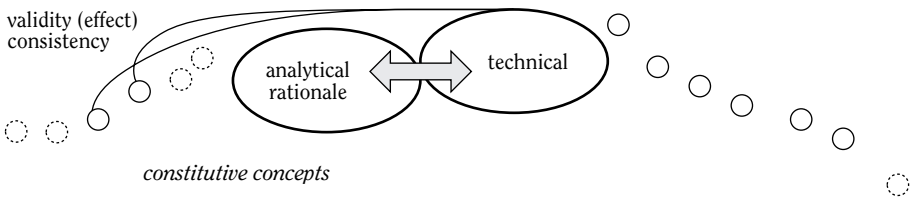

Figure 2: Constitutive concepts in applied linguistics (Weideman, 2007b: 42)

2. The technical mode that qualifies the applied linguistic instrument that a language test constitutes does not exist on its own, but is guided by a set of **regulative** ideas – what can be defined as leading or guiding conditions. The regulative conditions ensure that issues related, for example, to the social, are anticipated in the design of the test as indicated in the figure below:

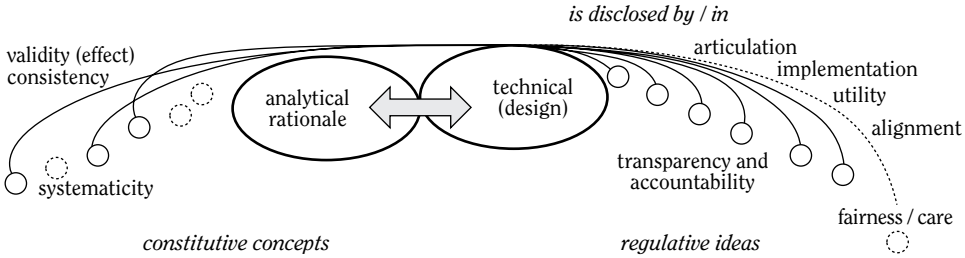

Figure 3: Constitutive concepts and regulative ideas in applied linguistic designs (Weideman, 2007b: 44)

The theory does not "dictate or prescribe" (Weideman, 2007a: 599) the design but is used to provide a rationale for it. Equally important is the concern here for the social and ethical dimensions related to the "designed solution" (2007a: 599) and that "the solution when implemented must also have ethical (and other) dimensions, i.e. must be transparent, accountable, theoretically and politically defensible and promote the interests of those affected by it" (2007a: 599). According to Weideman, validity and reliability are constitutive concepts for the characterisation of the applied linguistics artefact, e.g. a language test or a course design. A test must do what it is designed to do and it must also be consistent. While these are the "necessary conditions for its design" (2007a: 602) they do not function in isolation but in harmony or accordance with other factors, qualities or modes such as the lingual, the social,

economic, aesthetic, juridical and ethical dimensions of reality, and the way that these are reflected in concepts and ideas such as, respectively, the technical interpretability of the scores/outcomes of the test, the implementation of the test, its technical utility, alignment with needs of students and administrators, transparency, accountability and fairness.

4. Fundamental concepts in language testing

The framework referred to here is based on a "representation of the relationship among a select number of fundamental concepts in language testing" (Weideman, 2009: 241). The two main functions that have already been identified above are the technical mode and the analytical dimension. These do not function in isolation. The relation between the two is 'reciprocal' (Weideman, 2009: 244). The technical mode interacts not only with the analytical mode, however, but is also connected with all other modes, as can be seen in Table 1 below. Weideman points out that the technical unity of multiple sources of evidence, the reliability of a test, its validity and its rational justification are foundational or constitutive applied linguistic concepts (2009: 247). These may also be designated necessary requirements for tests (Weideman, 2009: 247). Important is the fact 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" (Weideman, 2009: 247). This, according to Weideman, is 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 (Weideman, 2009: 247) as opposed to focusing specifically only on one concept, such as validity.

This technical dimension of the applied linguistic design also links with the lingual, social, economic, aesthetic, juridical and ethical aspects. According to Weideman, the links between the technical, qualifying function of the test design and other aspects yield the ideas of technical articulation, test implementation or use, technical utility, the alignment the test has with learning and teaching language, its public defensibility or accountability, and its fairness or care for those taking the test (Weideman, 2009: 247).

This theoretical foundation or framework can be understood more easily if viewed in the form of a table (see page 113).

The table and the theoretical framework it articulates seem to suggest that if conditions such as consistency, validity, theoretical and social defensibility, transparency, accountability and fairness, are anticipated in the design of a test, then that test will fulfil the requirements of being a (psychometrically and socially) good test. It also suggests a move away from earlier beliefs about language testing – one view being purely asocial and psychometric, and, in the other, a move away from using assessment instruments to answer questions about the social aspects related to testing.

What this framework does is highlight a number of important concepts in language testing. This view allows for a more open and flexible way of designing and using tests rather than the restriction of an "overarching or unified" (Weideman, 2009: 239) concept. It is important at this stage to consider closely the work of Messick with a view to determining whether the framework he proposes adequately answers questions related to the design and administration of tests that are valid and reliable as well as socially responsible and transparent.

Table 1: Constitutive and regulative moments in applied linguistic designs

Applied linguistic design	Aspect / function / dimension / mode of experience	Kind of function	Retrocipatory / anticipatory moment
is founded upon	numerical	constitutive	unity within a multiplicity of sets of evidence and conditions for (test) design
	kinematic		internal consistency (technical reliability)
	physical		internal effect / power (validity)
	organic		technical differentiation
	feeling		technical perception and intention
	analytical	foundational	design rationale (construct validity or theoretical defensibility)
is qualified by	technical	qualifying / leading function (of the design)	
is disclosed by	lingual	regulative	articulation of design in a blueprint / plan
	social		implementation / administration
	economic		technical utility, frugality
	aesthetic		harmonisation of conflicts, resolving misalignment
	juridical		transparency, public defensibility, fairness, legitimacy
	ethical		accountability, care, service

(Weideman, 2007a: 602)

5. Messick on validity

Messick's article on validity, published in *Educational Measurement* (1989b) is still considered one of the most significant writings in the field. It is a work that has been and still is widely quoted by experts in the field of language testing. The article, simply entitled 'Validity', constitutes a major change in the way validity, and more specifically, construct validity began to be understood. In the opening lines to this article he states that

> validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness of inferences* and actions based on test scores or other modes of assessment (Messick, 1989b: 13).

He explains that validity is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use, stating that "what is to be validated is not the test or observation device as such, but the inferences derived from test scores or other indicators" (Messick, 1989b: 13). Messick stresses that validity is a matter of degree, not an all

or nothing measure or property (1989b: 13). Equally important is his contention that validity is an ongoing process, an "evolving property and a continuing process" (Messick, 1989b: 13).

A key point that Messick raises in this article is the view of validity as a "unitary concept" (Messick, 1989b: 13). Since as early as the 1950's, test validity has conventionally been broken up into three types: content validity, predictive and concurrent criterion-related validity, and construct validity. For Messick, however, content validity and predictive/concurrent criterion-related validity do not qualify to "bear the name 'validity' and to wear the mantle of all that name implies" (Messick, 1980: 1015). He justifies this by stating that because content validity provides "judgemental evidence in support of the domain relevance and representativeness of the content of the test instrument, rather than evidence in support of inferences to be made from test score" (Messick, 1989b: 17), it is not validity at all. He dismisses criterion-related validity as a type of validity because it

> relies on selected parts of the test's external structure. The interest is not in the pattern of relationships of the test scores with other measures generally, but rather is more narrowly pointed towards selected relationships with measures that are criterial for a particular applied purpose in a specific applied setting (Messick, 1989b: 17).

But while he dismisses content validity and criterion-related validity as types of validity, he does not dismiss the value of the evidence they provide. Instead he suggests "the use of labels more descriptive of the character and intent of each aspect" (Messick, 1980: 1014).

For Messick the compartmentalising of validity into different types "leads to confusion and, in the face of confusion, oversimplification" (Messick, 1980: 1014). A consequence of this, according to Messick, is the assumption on the part of test users that any one type of validity would do, so that once evidence of one type of validity is forthcoming, one is relieved of responsibility for further inquiry (1980: 1014). The point Messick makes is that there are not different types of validity, but different kinds of evidence, and that the points associated with each of these terms are important ones, but that their distinctiveness is blurred by calling them all 'validity'. A worse consequence, as indicated earlier, would be the belief on the part of test users that any one type of validity would be sufficient to validate a test. As stated earlier, Messick does not dismiss the value of the evidence that content-validity and criterion-related validity provide, stating that this evidence does contribute to score meaning.

However, he sees construct validity as based "on an integration of any evidence that bears on the interpretation or meaning of the test scores" (Messick 1989b: 17). Messick states that

> construct validity is indeed the unifying concept of validity that integrates criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships. The bridge or unifying theme that permits this integration is the meaningfulness or interpretability of the test scores, which is the goal of the construct validation process (Messick, 1980: 1015).

Messick's view on validity is, according to Van der Walt and Steyn (2007), "a more naturalistic, interpretative one" (2007: 139).

> Messick saw assessment as a process of reasoning and evidence gathering carried out in order for inferences to be made about individuals and saw the task of establishing the meaningfulness and defensibility of those inferences as being the primary task of assessment development and research (McNamara & Roever, 2006: 12).

Messick introduces the social dimension into this picture by arguing two issues:

> That our conceptions of what it is that we are measuring and the things we prioritise in measurement, will reflect values, which we can assume will be social and cultural in origin, and that tests have real effects in the educational and social contexts in which they are used and that these need to be matters of concern for those responsible for the test (McNamara & Roever, 2006: 13).

Messick used a, by now well-known, matrix to summarise his theory on validity:

Table 2: Facets of validity as a progressive matrix

	TEST INTERPRETATION	**TEST USE**
EVIDENTIAL BIAS	Construct validity (CV)	CV + Relevance/Utility (R/U)
CONSEQUENTIAL BIAS	CV + Value Implications (VI)	CV + R/U + VI + Social Consequences

(Messick, 1989a: 10)

Messick explains this matrix by stating that the

> validity of test interpretation and test use, as well as the evidence and consequences bearing theron, are treated here in the unified faceted framework…, because of a conviction that the commonality afforded by construct validity will in the long run prove more powerful and integrative than any operative distinctions among the facets (Messick, 1989b: 20).

Construct validity is to Messick the "integrating force that unifies validity issues into a unitary concept" (1989a: 10), it "binds the validity of test use to the validity of test interpretation" (1989a: 10) and "binds social consequences of testing to the evidential basis of test interpretation and use" (1989a: 10).

Messick has long been accepted as the expert on construct validity. His (1989b) article has been called the "most cited authorative reference" (Shepard, 1993: 423) on the topic of validity. Yet a close reading of the literature on validity, construct validity and language testing reveals that Messick's views have not been unquestioningly accepted by all experts in the field. Concerns and questions are addressed carefully, probably for fear of upsetting the applecart on which has carefully been placed the concept of construct validity. The reasoning behind this could be, according to Weideman, "the massive influence of the views of Messick and the institutional base that he represented" (Weideman, 2009: 239).

Whatever the reason, Messick's views do raise a number of concerns. A first concern lies in the challenging and complex nature of Messick's work. Today the work of developing tests no longer lies in the hands of psychometrists and measurement specialists alone. Taylor (2009: 22) points out that there are growing numbers of people involved in selecting or developing tests and they often find themselves doing this without much background or training in assessment. The present day emphasis on the importance of testing means that many professionals in a variety of fields have to play the role of test developer – such as the language teacher who wants to

design a test to test the writing levels of her class, but has no formal training in designing tests. Her first step then would be to consult the literature available on the designing of language tests – leaving her with the daunting task of unravelling Messick's concept of validity. Despite the availability of all of Messick's writing, McNamara and Roever still write about the need to make language test development and validation work "more manageable" (2006: 33) while Shepard refers to the "complexity of Messick's analysis" (1993: 427). While Messick has made an influential contribution to the field of testing, his work is not easily accessible to the lay person who needs to understand the field of testing, nor does he present us with a framework or guidelines to assist in the designing of tests that are accessible and transparent. Why then the huge influence of Messick in the field of testing? The main reason probably is that language testing was for a long time focused on psychometrics, and that Messick's predecessors worked firmly within that tradition. Could it be that Messick's consideration of the social consequences in testing came at exactly the time when the field needed such a change? Is it possible that Messick's theory was so readily accepted and became so influential because it seemed as if he was offering a new way of looking at and evaluating/assessing tests, one that included a consideration of the consequences of the uses of test scores, one that "extends the boundaries of validity beyond test score meaning to include relevance and utility, value implications, and social consequences" (Shepard, 1993: 424)? Herein lies a second concern. Despite Messick's incorporation of the social dimension in his theory of construct validity, there is still a heavy reliance on the collection of empirical data and statistical measures. Messick stresses this in his work. He states that

> the essence of this unified view is that the appropriateness, meaningfulness, and usefulness of score-based inferences are inseparable and that the integrating power derives from empirically grounded construct interpretation (Messick, 1989a: 8),

and that the "watchword for educational and psychological measurement is to maximise the empirically grounded interpretability of the scores and minimise construct irrelevancy" (Messick, 1989b: 89). A key question here is whether these are sufficient to answer all questions about the process of measuring language ability. The intention here is not to ignore or discredit the contribution that empirical considerations or constitutive elements have made to the field of language testing, but instead to question whether these are sufficient to answer all questions about the process of measuring language ability.

Despite the influence of Messick's work, others in the field have raised questions and concerns. McNamara and Roever explain Messick's validity matrix as follows in table 3.

McNamara and Roever translate the matrix to make clear "the way in which Messick's theory takes theoretical account of the aspects of the social dimension of assessment" (2006: 13). More importantly, they point out a glaring flaw in the matrix. In explaining Messick's validity matrix, McNamara and Roever state that "aspects of the social context of testing are more overtly present in the model, in the bottom two cells of the matrix" (2006: 13). This for them then raises the question of the

> relationship of the fairness oriented dimensions of the top line of the matrix to the more overtly social dimensions of the bottom line, a question it could be argued that Messick never resolved and remains a fundamental issue facing our field (McNamara & Roever, 2006: 13).

Table 3: Understanding 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 hence the sense we make of scores?	What happens in our education systems and the larger social context when we use tests?

(McNamara & Roever, 2006: 14)

McNamara and Roever here raise an important point. Messick has stressed the unifying and integrating nature of his view of construct validity. On closer inspection of his matrix, one is forced to admit that there is no close integration or unifying of the different concepts. While Messick's matrix asks us to consider questions about the social dimension of language testing, these questions have been relegated to the bottom row of the matrix. The empirical and social still exist, but may in such a view continue to operate as separate entities in the field of testing.

Shepard (1993) also raises a number of concerns with Messick's validity matrix. Unsurprisingly, a first concern has to do with the faceted nature of his matrix. Shepard states that the faceted presentation "allows the impression that values are distinct from a scientific evaluation of test score meaning" (1993: 427), an argument similar to the one raised by McNamara and Roever above. She states that the separate rows in Messick's table make it appear that one would first resolve "scientific questions of test score meaning and then proceed to consider value issues" (1993: 427). She also observes that the sequential segmentation of validity gives researchers tacit permission to leave out the very issues that Messick has highlighted, because the "categories of use and consequence appear to be tacked on to 'scientific validity', which remains sequestered in the first cell" (1993: 427). Shepard's final point on this is related to what she refers to as the "complexity of Messick's model" (Shepard 1993: 429), stating that both the model and the chapter on construct validity stress that construct validity is a "never-ending process" (429) and that while this may be true, the "sense that the task is insurmountable allows practitioners to think that a little bit of evidence of whatever type will suffice" (429).

This is undoubtedly ironic, as this was exactly the reason why Messick saw a need for a unified view of validity – so that test developers will not simply use one type of validity to validate a test. Messick argues against the compartmentalising of validity into different types, claiming that it "leads to confusion and, in the face of confusion, oversimplification" (Messick, 1980: 1014). He also goes on to claim that the distinctiveness between them is blurred by calling them all 'validity'. These are, then, his motivating reasons for unifying the concept of validity. This raises concerns. Messick's main aim was unifying the concept of validity. When we use the words 'unify' or 'unifying' we refer to things that are the same as or uniform or not different or not varying. Is it possible that Messick's use of this term causes the very confusion he refers to above? If

the concept of validity is a unified one, then it would make sense to see everything under that concept as being or meaning the same. If content, criterion and face validity are unified or the same as or not varying it would potentially make sense to use any one type to validate the test – they are, after all, uniform or in Messick's words 'unified'. He states that "to speak of validity as a unified concept is not to imply that validity cannot be differentiated into facets" and that "the distinctions introduced may seem fuzzy because the facets of validity are not only intertwined but overlapping" (1989a: 9). Validity then is not just unified but both unified and faceted. In his attempts to unify the concept of validity, Messick has created a most complex network of arguments. It is hard to avoid the conclusion that Messick's attempts to view validity as a unified concept has to some extent created the very problems he wanted to avoid. This is compounded by the complexity of his writing which makes it inaccessible to many readers. There remains also a suggestion of too much of a reliance on the importance of empirical data. What the field requires is shared emphasis on empirical data as well as social effects.

Messick's claim that "validation is a continuing process" (1989b: 13) suggests that it is never-ending. Bachman and Palmer reiterate this view when they state that construct validation is an "on-going process" and that "we should not give the impression that a given interpretation is 'valid' or 'has been validated'" (1996: 23). Is it not possible that there could be an end to the process of validation? Should there not be a valid test at the end of the process? As Weideman asks, "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?" (Weideman, 2009: 242). Is it not acceptable to ask whether an instrument that has undergone a process of validation may be shown to be a valid test? Does the validation not demonstrate that it does what it was designed to do? The answer here should be an affirmative or a negative, however qualified the 'yes' or 'no' might be.

It is only when one unquestioningly accepts Messick's definition of validity and construct validity that one is not allowed to ask such questions. Fortunately, experts in the field are asking such questions. McNamara and Roever make two thought provoking statements: that although "validity theory investigated the technical qualities of tests in the interests of fairness, it did not address the wider social function of tests" (2006: 248), and that "despite Messick's efforts to build a unitary approach to validity that acknowledged the social meaning of tests, validity theory has remained an inadequate conceptual source for understanding the social function of tests" (2006: 249). Davies and Elder claim that "Messick's conceptual clarity can be analysed less charitably" (2005: 799), stating that because validity can only be achieved through validation, what Messick does is offload all the problems of validity onto validation, "leaving validity as an abstract and essentially empty concept" (2005: 799). Shepard's concluding concern with Messick's matrix deals with the question of subsuming everything under the umbrella of construct validity. She states that "most theorists agree that validation includes the whole of Messick's framework, not only the first box. But can all of the implied questions be subsumed under construct validation without degrading its scientific meaning?" (1993: 428).

It would be useful at this point to look at one further reinterpretation of Messick's matrix. Weideman, by "turning some of the terms around so that we make a small adjustment to the matrix" (2009: 239) suggests the following reading of Messick's distinctions:

Table 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

(Weideman, 2009: 239)

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

(1) The technical adequacy of inferences made from test scores depends on multiple sources of empirical evidence.
(2) 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.
(3) 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.
(4) 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 (Weideman, 2009: 240).

The matrix above is thought provoking for a number of reasons. The most obvious of these is that the matrix is not a validity matrix, and construct validity does not appear in every cell. Instead the matrix is concerned with the "relationship of a selection of fundamental considerations in language testing" (Weideman, 2009: 240). According to Weideman, the matrix above is clearly not "solely about validity" (2009: 240). He states that concepts in the matrix, "while obliquely related to the technical power of a test … rather articulates the coherence or systematic fit of a number of concepts related to language testing" (2009: 240). These concepts, as indicated in the matrix, would be the empirical evidence such as for reliability and validity, as well as ideas on the impact, usefulness/utility as well as the social justification and political defensibility of the test. The emphasis here is on a "select number of fundamental concepts in language testing" (2009: 240) rather than subsuming all concepts under validity. The unitary concept of validity does "blur the distinctions" (to use Messick's words), but for reasons different to Messick's. Subsuming everything under construct validity undermines the importance of these other concepts, whereas seeing them as a number of fundamental concepts in language testing, we acknowledge their contribution to the responsible design of language tests. In arguing against the need for a unitary concept of validity Weideman states that concepts like "technical appropriateness, technical meaningfulness (interpretation) of measurements (test scores), utility, relevance, public defensibility and the like must be conceptually distinguishable to make sense" (2009: 241). He observes that if one does not distinguish what is "conceptually distinct, the distinction so avoided subsequently obtrudes itself upon the conceptual analysis" (2009: 241).

6. Conclusion

The move by Messick (1989a; 1989b) to include a concern for the social consequences in the field of language testing is indeed a positive move. The framework proposed by Messick was one of the first to consider questions related to test use, social consequences and test fairness, while not disregarding the importance of empirical or statistical evidence. It is, however, his attempt to unify the field in terms of a single concept (validity) that raises an important question: Does this 'unification' adequately address concerns about the incorporation of the social dimension in language testing in the design and administration of language tests? It is the view here that it does not. Rather than this overarching concept of validity, what the field of language testing requires is that test developers see the relationship between fundamental concepts in language testing. In employing a framework that incorporates a concern for the empirical analyses of a test, as well as a concern for the social dimensions of language testing, one is forced to consider important questions related to every aspect of the test: the validity and reliability of the test, the reason for giving the test, the effect of the test on the test-taker, concerns about the design of fair tests, the rights and responsibilities of the test designer and the rights and responsibilities of the test-taker. The value of this framework, according to Weideman, "lies in its separating out what is conceptually distinct, and, by so doing, enriching our theoretical understanding of the constitutive and regulative, necessary and sufficient conditions for language testing" (2009: 249).

REFERENCES

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

Becker, D.F. & Pomplun, M.R. 2006. Technical reporting and documentation. In Downing, S.M. & Haladyna, T.M. (Eds) *Handbook of Test Development*. New York: Routledge. 711-723.

Cook, L.I., Schmitt-Cascallar, A.P. & Browm, C. 2005. Adapting achievement and aptitude tests: a review of methodological issues. In Hambleton, R.K., Merenda, P.F. & Spielberger, C.D. (Eds) *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*. New Jersey: Lawrence Erlbaum Associates. 171-192.

Davies, A. 1990. *Principles of language testing*. Cambridge: Basil Blackwell.

Davies, A & Elder, C. 2005. Validity and validation in language testing. In Hinkel, E. (Ed.) *Handbook of research in second language teaching and learning*. New Jersey: Lawrence Erlbaum Associates. 795-813.

Kane, M. 2006. Content-related validity evidence in test development. In Downing, S.M. & Haladyna, T.M. (Eds) *Handbook of Test Development*. New York: Routledge. 131-153.

McCallin, R.C. 2006. Test administration. In Downing, S.M. & Haladyna, T.M. (Eds) *Handbook of Test Development*. New York: Routledge. 625-652.

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

Messick, S. 1980. Test validity and the ethics of assessment. *American pathologist*, 35: 1012-1027.

Messick, S. 1989a. Meaning and values in test validation: The science and ethics of assessment. *Educational researcher*, 18(2): 5-11.

Messick, S. 1989b. Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed.). New York: American Council on Education & Macmillan. 13-103.

Shepard, L.A. 1993. Evaluating test validity. *Review of research in education*, 19: 405-450.

Taylor, L. 2009. Developing assessment literacies. *Annual review of applied linguistics*, 29: 21-36.

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.

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

Weideman, A. 2007a. The redefinition of applied linguistics: modernist and postmodernist views. *South African linguistics and applied language studies*, 24(1): 589-605.

Weideman, A. 2007b. 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. *South African linguistics and applied language studies*, 27(3): 235-251.

ABOUT THE AUTHOR

Avasha Rambiritch
Unit for Academic Literacy
University of Pretoria
Email: Avasha.rambiritch@up.ac.za