This is the post-print version of a manuscript accepted by Elsevier for publication in *Language & Communication*, 9p. Available online at http://dx.doi.org/10.1016/j.langcom.2016.12.004.

Does responsibility encompass ethicality and accountability in language assessment?

Albert Weideman[1]
University of the Free State

Abstract

A central conceptual issue in language assessment in general, and in the work of Alan Davies in particular, is never fully resolved. How is responsible language test design related to ethicality? This unfinished business goes back to the unresolved debate about validity and validation, that has resulted in a loss of conceptual clarity about sound language assessment. The contradictions inherent in declaring first validity to be the overarching concept, and then attempting to expand it by promoting other ideas to be the prime considerations, further confuse the issue. This contribution argues that such expansion is unhelpful. A way out is to look at language test design as being responsive to certain typical and general conditions. In that relationship, between the (subjective) making of the test as an artefact that presents us with a response to certain design principles, and the designed object, the actual test itself, lies a potential way out of the impasse.

KEYWORDS: language testing; accountability; ethicality; transparency; fairness

Diligence, quality and responsible test design

Though this contribution will deal with how some ideas and principles of language test design conventionally considered to be important are conceptualized, it is written primarily from the perspective of a language test developer. Language assessment design presents an onerous set of responsibilities for applied linguists who work in that subfield, and few would argue with the premise that the design of language tests has to be done with the greatest deliberation, diligence and care. Language tests have to be designed and developed in a way that is worthy of professional conduct (Davies, 1997). Responsible test designers thus regard their work as a profession; it is not surprising that Davies and Elder (2005: 800) observe: "What professions exist to do is to combine field expertise with a proper moral and social concern to act responsibly in normal settings." The combination they refer to

1. E-mail address: albert.weideman@ufs.ac.za; Office of the Dean, Humanities, University of the Free State, Nelson Mandela Avenue, Park West, 9301 Bloemfontein, South Africa.

here is the theme of this paper: the relation between the technical know-how to make a language test, and its sometimes unarticulated and at times even obscure connection with other dimensions, concepts, and ideas that inform test design and the subsequent use of tests.

The quality of the assessments designed by language testers becomes prominent where tests of language ability are used for medium to high stakes purposes, and acutely so when they are administered on scale. Many high stakes tests one finds at the hinge-points of education or access to employment opportunity, for example just before the end of pre-tertiary education, or before entry into higher education or the world of work (cf. too Joseph, 2016). Especially where secondary school exit examinations are administered as public, government-initiated tests, they are almost without fail high stakes assessments. Their results will inevitably be used as evidence of potential employability in the world of work, or for admission to tertiary education. What increases their impact even more is that they are often nationally organised - and in the case of some large scale, commercial language assessments, internationally. In all such cases, the quality of the assessment on a national or global scale further significantly increases their potentially beneficial or disadvantageous impact. It is no coincidence then that their worth and quality should be defensible both publicly and among experts.

There is a consistent thread in the work of Alan Davies and those who collaborated with him on all of these themes and issues. This contribution will analyse some of the most significant discussions in the work of Davies *c.s.* from the perspective of those applied linguists who are engaged in the design and development of language tests. It will ask what the enduring meaning of these discussions might be for the design of language assessments. It is noteworthy, in the first instance, that Davies not only wrote on language assessment as a subfield of applied linguistics (for this, cf. McNamara & Roever, 2006: 255; cf. too McNamara, 2006; Weideman, 2006), but in such discussion consistently brought accountability for such applied linguistic designs into play. Referring to the three prime applied linguistic artefacts (Weideman, 2014, 2016), namely language courses, language tests and language policies, Davies (2008: 298) states that "applied linguistics is prepared in its curricula and its assessments and in its planning… to be accountable", adding that it accomplishes that, amongst other things, "by theorising practice".

How does that accountability fit in with, or relate to responsible test design, however? How do all the various concepts and ideas, such as transparency, accessibility, fairness, ethical codes (Boyd & Davies, 2002), and standards (Davies, 1997, 2008) cohere, if at all? Does accountability depend purely and solely on predetermined or professional standards? Or does it perhaps depend even more on locally relevant and contextualized codes of practice? What are the necessary, and what the sufficient safeguards for fairness in language testing? This contribution will offer a possible alternative perspective to the ideas currently in the mainstream, one that is potentially more integrated and possibly clearer conceptually. It will do

so by proposing that the idea of responsible test design offers a coherent framework of principles that encompasses the notions of transparency, utility, accessibility and accountability – in fact the ethicality of assessment generally. It will proceed, like Davies, from the assumption that language test design falls within the domain of constructing applied linguistic artefacts. It may also provide further insight into the role of standards in achieving responsible design.

Unresolved conceptual issues

In discussing how standards may or should affect language assessment, Davies (2008) deals specifically with standards as a set of criteria for assessing the appropriateness of the language tests that are designed to measure language performance. Standards as procedures and yardsticks for test design are thus his prime concern, as well as that of this discussion. Recognizing that reliability, validity, appropriateness, utility, accessibility, and theoretical defensibility are necessary prerequisites for tests, Davies (2008: 491f.) also emphasizes that these may not be sufficient: a public demonstration of their worth that clarifies their political, ethical and juridical dimensions is also required. The idea of consequential validity (Messick, 1988, 1989: 20, 88; Davies, 1997: 335; Davies, 2011; Davies & Elder, 2005: 798; McNamara & Roever, 2006; Weideman, 2012; Boyd & Davies, 2002: 304, 306), of gauging the social, political and economic impact of language tests, clearly is what Davies (2008) has in mind. Perhaps, if one considers his discussion and analysis as a whole, that discussion may even be shown to encompass the broader notion of the appropriateness of the public provision of language instruction and the assessment at school of such language instruction. Elsewhere, that is sometimes referred to as the contextual or ecological validity of a course or an assessment (Arzubiaga, Artiles, King & Harris-Murri, 2008).

At the same time, we should note that this discussion (Davies, 2008) refers to and builds upon earlier ones in Davies's work, especially his involvement in helping to draft, support, and promote the Code of Ethics of the International Language Testing Association (ILTA) (Davies, 1997: 336). In these analyses, the weakness of the language testing profession to enforce adherence to a set of standards is noted (Boyd & Davies, 2002: 307), but also that the cause is not lost: "For the emerging profession of language testing it is not too late to build in openness to its professional life" (Boyd & Davies, 2002: 312). As we can see, in Davies's perspective the transparency of professional practice remains a powerful tenet of his arguments about accountability. Thus openness, a readiness by the professional to provide "clearer public information about the professional expertise needed for language test construction" (Davies, 1997: 338), is introduced as a means of language test designers being publicly held to their own professional standards. Openness or transparency becomes the basis for accountability, the other closely related, significant theme in Davies's analyses: as can be expected, Boyd and Davies (2002) explicitly take accountability as their central point.

These discussions about ‘standards’ therefore immediately become related to a whole range of principles for evaluating the quality of language tests. That range might include all the familiar issues: transparency, accessibility, utility, accountability, fairness, care, integrity, trust, humility and the like. These are in fact the main threads, too, of the ILTA Code of Ethics that Boyd and Davies (2002) discuss as the proposed backbone of professional conduct for language testers. Yet the conceptual question that is neither asked nor answered when bringing them into such a code or its subsequent discussion is whether they may be treated as disparate issues (individually articulated ‘principles’, with annotations, as in the Code), or whether (more plausibly) they should also be treated as related, and if the latter, what that relationship is. The problematic nature of the relationship becomes evident when we come across statements such as: “While I can accept that ethics in language testing does include validity, whether it is wider in scope remains an unresolved question” (Davies, 1997: 335). Or consider this claim, again with reference to the ILTA Code of Ethics: “… what we are seeing in the professionalizing and ethicalizing of language testing is a wider and wider understanding of validity” (Davies, 2008: 491).

The conceptual problem in these statements is that, in line with the current (post-Messick) orthodoxy in language testing, validity may then be seen to encompass everything. But does it? The further trouble is that for those who are steeped in this orthodoxy, it may well be impossible even to ask that question or to see that point. So beguiling are our professional beliefs that even critical voices such as that of Davies may be tempted to leave them unexamined. While remaining aware of Messick “add[ing] to the problems of validity by extending its scope into the social and the ethical”, Davies and Elder (2005: 799), for example, remain somewhat ambivalent on this point. On the one hand, it is claimed – contrary to post-Messick orthodoxy - that “in some sense validity does reside in test instruments” (Davies & Elder, 2005: 797), and that it is therefore “not just a trick of semantics… to say that one test is more valid than the other for a particular purpose” (Davies & Elder, 2005: 798). This clearly conceptualises the validity of a test as a separate and unique ingredient and indication of its quality, in contrast to the inherently contradictory view taken by the current orthodoxy, to be returned to below, that validity is dependent on the interpretation of the results of a test. On the other hand, referring to validation studies, they conclude, though this time in line with the current orthodoxy, that the inconclusive nature of test validation studies demonstrates that “validity is not tucked up in the test itself but rather resides in the meanings [interpretations] that are ascribed to test scores” (Davies & Elder, 2005: 809). What is more, they remark that an “unintended consequence” of abusing the purpose of a diagnostic test by using its results for selection or access to (the scarce resource of) a course is “a threat to the test’s validity” (Davies & Elder, 2005: 808). The simple question to be asked here is: “Why would it not rather be a threat to the test’s ethicality?” The claim that the inappropriate and harmful use of test results undermines its validity only makes sense if validity is considered, first, to

encompass ethicality, and, second, to be an inherent quality of a test. If validity and ethicality were each principles of test design, the harmful employment of test results would clearly in the first instance have been just that: an abusive and therefore unethical use of the outcome of the measurement. However, the abusive use of a measuring instrument does not in the first instance say anything about the quality or technical force and capacity – what is usually referred to as the validity - of that instrument. If I abuse the purpose of a measuring tape by using the measurements it yields to harm or belittle others, that says nothing of the quality or accuracy of the measurement, of its validity.

The objection to this line of reasoning is usually to repeat (without further examination) the definition of validity as the adequacy and appropriateness of the interpretation of the results of a test. This definition sees validity as being dependent on a validation process that systematically brings together a multiplicity of sets of evidence to support such interpretations. The form of this systematic process is generally understood to be an argument (Kane, 1992, 2001, 2010, 2011; Van der Walt, 2012). As I hope will become clear below in the discussion of this currently orthodox view, repeating a definition that is widely and uncritically accepted does not quite rid it of its contradictions and the resulting lack of conceptual clarity.

Validity and validation revisited

Despite the possible conceptual problem identified in the previous section, Davies and Elder (2005) also provide a possible way out, by articulating an alternative perspective on validity and validation. They do so in terms of an analogy: just as doing justice makes full use of the law, so validation is identified as the process through which validity is achieved (2005: 796). I interpret this, also with reference to some components of the surrounding argument which have already been referred to above, to mean that the subjective process of validation may yield insight into the objective quality of a test that is called its validity. Subjective validation, undertaken either in the form of an argument with hypotheses (Davies & Elder, 2005: 802ff.; Van Dyk, 2010; Van der Walt, 2012) or by examining the plausibility of inferences drawn from the test's results (Kane, 2011: 13), on the one hand, and, on the other, objective validity, are then two sides of the same coin. As Davies (2011: 38) phrases it: "[W]e validate a test and then argue that it is valid." Together with the subjective validation process goes the warranted, adequate and appropriate force of the measurement that is expressed in the results (objects) that such measurement yields.

What is even more significant, and has nowhere been further discussed yet, as far as I know, is that Davies and Elder (2005: 796) at the outset offer a remarkable insight that has the potential of putting us on a conceptually much more productive path: "Validity is self-contained: Its definition is reflexive ('validity is validity') just like those other great abstractions: beauty, truth, justice." This means both that

validity is in a significant sense indefinable, and also that it is a distinguishable and therefore separate criterion for language tests. The first implication is that, should we widen validity to include other, equally distinguishable (but potentially also indefinable) criteria for language test design, by bringing all manner of disparate concepts in under its umbrella, confusion would result. The second, less palatable implication for those who believe that viewing validity as a quality of a test is no longer acceptable, is that such an indefinable characteristic, if not acknowledged as such, is likely to obscure our conceptualisation. It is likely to return not only to haunt those in pursuit of conceptual clarity, but also to confuse those who work to design and produce actual language tests.

So this is the difficulty that the current orthodoxy never quite confronts. Having identified validity with the interpretation of the results of a test, it has in fact closed off any potential conceptual alternative. Consider for a moment, however, that proponents of the current orthodoxy, starting off with its originator, Messick, then have to employ all manner of conceptual circumlocutions to replace validity. If validity is no longer a quality of a test, no longer "tucked up in the test itself", then its elimination is revenged by its return in the guise of a synonymous concept: thus Lee (2005: 2) will refer not to validity as such, but to the 'effectiveness' of the use to which a test can be put, or of a test being "valid in a specific setting" (Lee, 2005: 3), remarking, furthermore, that through verbal protocols we may investigate the consequences of a test, because these "should be considered valid and useful data in their own right". Data may therefore be valid (without being subjected to interpretation), but not the measurement itself. In the current orthodoxy, care is taken to avoid referring to validity as a quality of a test. Thus a test is not valid, but "measures the construct": "… if we ensure that a given test measures the construct … we say that the resulting scores provide an empirically informed basis for decision-making" (Lee, 2005: 4). Lee is not alone: Kane (2011: 5) continues to speak of the "validity of the criterion", and of "sources of invalidity" without considering that if a criterion can be valid, and the use of a test invalid, these are conceptually nothing less than characteristics of the measure or of the event. While abandoning validity as a quality of a test, McNamara and Roever (2006: 17) continue to assume that a "test is … a valid measure of the construct" (McNamara & Roever, 2006: 109), and to speak about "items measuring only the skill or the ability under investigation" (McNamara & Roever, 2006: 81). In these pronouncements there is little mention of the interpretation of the scores derived from these items, which in the current orthodoxy is where validation is located. In fact, Messick himself often uses circumlocutions such as a "test … accomplishing its intended purpose" (Messick, 1980: 1025), or speaks of tests "purported to tap aspects" of a trait (Messick, 1989: 48; 50, 51, 73).

In these synonymous concepts and circumlocutions, current analysts are thus merely following in the footsteps of the scholar who provided the conceptual opportunity to eliminate validity in following claim (Messick, 1980: 1023; cf. too 1981: 18): "Test validity is … an overall evaluative judgment of the adequacy and

appropriateness of inferences drawn from test scores." From validity being equated with an "evaluative judgment" to being identified with 'interpretation' is but one small, apparently unproblematic step: "Validity is associated with the interpretation assigned to test scores rather than with the scores or the test" (Kane, 1992: 527). 'Judgement' is replaced by and transformed into 'interpretation', and the scores can no longer be valid, though their interpretation can. Objections to the elimination of the concept of the objective technical force that a test result yields, such as the earlier ones of Popham (1997) or the later ones voiced by Borsboom, Mellenbergh and Van Heerden (2004), are brushed aside with remarks about how this would "strip validity theory of its concern for values and consequences and … take the field back 80 years" (McNamara & Roever, 2006: 250f.). Remarkably – and fortunately, if one is aiming for conceptual clarity - McNamara and Roever at the same time claim that that very same "validity theory has remained an inadequate conceptual source for understanding the social function of tests" (2006: 249)! In fact, they encourage "language testing research… to move beyond the limits of validity theory" (2006: 40) if it wishes to engage adequately with the consequences or impact of language tests (as Messick had envisaged), as well as with the social, political and ethical implications of such assessment. Below, I shall examine several such attempts to "move beyond" validity theory, and demonstrate that they too may lead us up a garden path.

As I have put it in a previous analysis, avoiding the use of "validity as descriptive of a test therefore merely returns in another guise, that of adequacy, or in similarly synonymous terms for the concept of effectiveness" (Weideman, 2012: 5). This is in fact Messick's solution. If one reads closely the texts of the Messick (1981: 10, 1980: 1023) formulations, it soon becomes evident that the concept of 'adequacy' has replaced 'validity', and that the other key idea is 'appropriateness'. But for Messick they are two distinct and distinguishable concepts. In my analysis, they refer, respectively, to two different analogical technical concepts. Adequacy, as an alternative for validity,

> is a concept that is linkable directly to the effects (or effectiveness) of applying a technical instrument such as a language test. The technical adequacy of a test refers to its force to measure what it claims to be measuring, its effectiveness, which is the classic definition of validity. A measuring instrument is adequate if its result (definable as the effect of the measurement which is caused by the application of the instrument) has the desired force. In its original physical sense, the concept of force is expressed in terms of cause and effect. In the analogical technical sense, it is used when we are dealing with technically qualified instruments such as tests; the measurement, when applied, acts as technical cause to achieve a certain technical effect, that is, to obtain a result. (Weideman, 2012: 5)

Its companion term in Messick's work, appropriateness, is not strictly speaking a concept, but in philosophical terms rather a concept-transcending idea (Strauss, 2009: 195). Appropriateness has to do with the social fit of a test, and it is an idea

that refers in this instance analogically to one of the social dimensions of the technically stamped, leading function of a test. For a further articulation and explanation of these terms, I refer to a number of earlier analyses (Weideman, 2009, 2012 and 2014). Appropriateness, the analogical social idea relating to test design and use, shows that Messick had indeed introduced an idea that allowed us to take account of the social dimensions of language test design. However, the concept of the technical adequacy of the instrument that he additionally promotes becomes merely a substitute concept for validity.

Why has validity theory been found wanting?

We have noted above the conclusion by McNamara and Roever (2006: 40) that language testing should "move beyond the limits of validation theory." What would such "moving beyond" entail?

I shall refer here to just three examples that illustrate the conceptual dilemmas that may accompany such an undertaking. The first of these has already been noted: Kane's redefinition of validity (1992: 527) that excludes the notions of adequacy and appropriateness in Messick's original definition, and promotes the interpretation of the results to prime position. This "interpretative argument" (cf. too McNamara & Roever, 2006: 24f.) takes us beyond validation by defining the technical object that should be validated not as the test, but as the "interpretation or use" (Kane, 2011: 3) of scores (the effects of the measurement). But surely validating an interpretation of a score cannot be same as validating either an instrument or the score that it yields as an effect of its measurement. That this cannot be the same is evident as soon as we consider that no amount of interpretation (valid or not) can make sense of a measurement produced by an ineffective, inadequate instrument. Of course Kane is correct, in the 'interpretative' turn that his proposal promotes, that a score, though an objective effect of the measurement, can never have any meaning on its own. There is no doubt that it needs interpretation. But the interpretation of a technical effect (which is the result of the measurement, usually expressed quantitatively either in a number or in a hierarchy of levels) relies on that objective effect being available, and not only valid, but also limited (within a certain range), consistent, and so forth. The shift from validating not the result of the measurement, but the interpretation of the result, is just that: a shift. However, a score needs to be a result of adequate quality to be technically interpretable. It may be, as one commentator on an earlier draft of this paper pointed out, that Kane and others who have alerted us to the importance of the interpretability of results – a point that I shall be returning to below – would fully agree with this last statement. Yet in a close re-reading of Kane's later (2011) explanations, it still appears to me as if the shift from a valid measurement to a validation of the interpretation of the score yielded by the measurement is uppermost, or at least primary.

A second and well-known attempt at moving beyond validity is that of Bachman and Palmer (1996; Bachman, 2001). Somewhat contradictorily, their first move is to shift the emphasis from test validity to test usefulness, by declaring usefulness, and not validity, as the "most important consideration in designing and developing a language test" (1996: 17; Bachman, 2001: 110), before then happily returning to an acceptance of Messick's primary consideration, namely construct validity: "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). What is the prime consideration to be, then? Test utility or test construct validity? What we should note, additionally, is that, as in the case of the first example given above, Messick's (1980) original formulation is subtly amended. Bachman and Palmer (1996) speak not of the judgement of the adequacy of the inferences drawn from test scores, as Messick does, but of the meaningfulness of the interpretations we can make about scores. It is interesting that some commentators do not see the modification at all, thinking that Bachman and Palmer are here simply aligning themselves with Messick: Xi (2008: 179) even declares that the idea of test usefulness makes "Messick's work more accessible". Other commentators, however, do appear to notice the conceptual minefield that this presents. Fulcher and Davidson (2007: 15), for example, have observed that Bachman and Palmer's "notion of test 'usefulness' provides an alternative way of looking at validity, but has not been extensively used in the language testing literature". Conceptually usefulness can, however, never be the same as validity (or validation). In fact, when Bachman and Palmer (1996: 18) describe their model of test usefulness, usefulness *encompasses* validity:

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

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

One may speculate about why validity, together with a number of other qualities of test design, is incorporated under usefulness, though it does serve to mask the dissimilarity of their views with those of Messick, provided one does not read too closely. Yet, as I have remarked previously, "surely no one would disagree that part (construct validity) of a whole (usefulness) that is made up of several other parts cannot conceptually be the same as that whole. The norms of logic do not allow that…" (Weideman, 2012: 3). A driving shaft can be part of a vehicle, and so can its engine, or its body, or its wheels, or any other part of the whole vehicle. But the wheel, or any other part, cannot be the vehicle. In fact, this "moving beyond" validation theory brings us no closer to a clearer conceptualization. Moreover, it does not provide us conceptual access to the technical conditions that language tests must conform to, those ideas and concepts that guide our design of these instruments.

A third example of where the conceptualization of language test design has moved beyond validity theory may be found in Kunnan's assertion of the "primacy of fairness" (2000: 1; cf. too Kunnan, 2004). But does that assertion indeed take us beyond validity theory? So strong is the influence of validity theory that Xi (2010: 167), for example, continues to frame fairness in terms of validity, calling for the "integration of fairness into validity". In a slightly modified manner Kane (2010: 177f.) views fairness and validity as different, though related ('intertwined') emphases in evaluating test quality. Again the original concept has been modified, with a resultant loss rather than a gain in conceptual clarity.

One may therefore ask: is Messick's contribution to validity theory indeed its culmination, as claimed by Xi (2008: 179)? If it were, why would subsequent conceptualizations then seek to modify it, asserting that interpretability, usefulness or fairness must be promoted to the prime concern? Those modifications may serve to explain why analysts have considered the operationalisation of validity as formulated by Messick (1980, 1981, 1988, 1989) as being neither feasible (Davies & Elder, 2005: 789; Xi, 2008: 179) nor adequate to address social concerns in language test design and administration (McNamara & Roever, 2006: 249; Rambiritch, 2012).

There is no doubt that fairness, ethicality, accessibility, interpretability, usefulness, accountability, validity and other components of sound language test design are important. The trouble lies in their sometimes confusing and contradictory conceptualization. In the next sections, I suggest a possible alternative.

Typical principles of language assessments

If Davies is correct in viewing language assessment design, along with language course design and language planning, as being firmly part of applied linguistics, the next step is to acknowledge that these three kinds of design, that usually result in concrete or factual language tests, language development courses and articulated language policies, are in one sense similar, and in another all typically different. They are similar, we know, in that each such design is stamped by the technical dimension of experience: that leading function of design that is characterized by the indefinable nuclear moment of shaping, forming, planning, facilitating or arranging (Weideman, 2014, 2016: chapter 11). At the same time, that leading technical function of the design is, in applied linguistic conceptualization, connected to a number of others. Most prominently, there is an analogical connection of the technical with the logical or analytical dimension of the design. Applied linguistic designs need to be theoretically justified. In the theoretical defensibility of a design we find another terminally important function of the planned arrangement, intervention or measurement, what I have called its founding function (Weideman, 2009, 2014). I shall return to this point below.

What is noteworthy, furthermore, is that in these three major applied linguistic artefacts – tests, courses and policies - there is a relationship in each entitary category between a condition-setting design, and the factual outcome of those designs, as in the following table (Weideman, 2011b):

Prior, conditioning artefact	**End-user format of design**
language curriculum	language course
construct and test specifications	language test
language policy	language management plan

Table 1: Levels of applied linguistic artefacts

In each of these categories, the one level of artefact acts as a normative condition for its end-user format (cf. too Rajagopalan, 2016). So the test construct and the specifications that flow from it set out conditions for the factual test. The test has to conform to these requirements, and so also with the other artefacts.

The first noteworthy point about this characterization of applied linguistic designs is that their normative and factual shapes are indissolubly linked. Norm and fact in these designs are two necessary levels; one cannot imagine the one without the other. When, in designing a language test, we are not open and clear about the language ability it measures, its construct, it has been constructed without the theoretical deliberation that it needs in applied linguistic terms. It may still be a language test, and it may still have been produced with much technical (design) imagination and creativity, but it does not qualify as an applied linguistic artefact, since it has not taken the detour into defending the theoretical base on which it has been built. If it is not an applied linguistic design, it may even lack any such base. It may therefore also be deficient in another than analogical analytical sense, since it may lack the technical transparency, the openness about the rational basis for the design that, as Davies has exhorted test designers, will allow us to become accountable also to its users or to the public at large.

There is a related, second noteworthy issue to the above. If language interventions have both a normative (conditioning) and a factual (end-user format) side to their designs, we may extrapolate from that observation that the actual making of tests, as is the case with other designs in the field (e.g. courses and plans), are generally responses to criteria, principles, conditions or requirements. We may also infer that the relationship between fact and norm is not only pervasive, but unbreakable. Hence my preference to refer to responsible design as the factual production of, in this case, a technically qualified instrument to measure language ability, in response to norms or conditions. In this conceptualization, the idea of "responsible design" therefore goes back to the original meaning of the term. If I had employed more conventional terminology, I would probably have used the adjective 'responsive', since that is the meaning intended: factual designs respond to

normative conditions. Yet the idea of ‘responsible’ to me is stronger, conveying also the onus that rests on the designer, and the range of choices that designers have to make, not the least of which is to be publicly accountable. If they make design choices that are illogical, invalid, not transparent, uninterpretable, not useful, unaccounted for, uncaring and uncompassionate, they are not merely unresponsive, in other words, but in light of where language test design has progressed, irresponsible.

A third implication of this characterization relates to an issue left unarticulated by the simplification that such a table as the one above inevitably carries with it. This is that there are further prior conditions than merely the test construct and specifications. These conditions refer to the codes of ethics, the standards, or the codes of practices that additionally shape and guide the design process. That is an important point, since it brings to the fore the fact that in making a test, we assume, first, that it is different from the other designed entities under discussion. A test is different from a language development intervention, the aim of which is language instruction that makes possible language learning. It is different too from a language policy, the latter being an official arrangement facilitating language use within an institution. Each is different from all of the others. The intentions for making the design, and the purposes to which each has to be put, make for typically different sets of norms and conditions to be applied in their respective designs.

To these typically different sets of conditions belong, in the case of the design of language assessments, the codes of ethics, standards, codes of practice and other norms that may be agreed upon professionally from time to time. It is an indication of how seriously the designers of language assessments take their responsibility that such codes and standards have been articulated, and may even, in the context of the three different designed entities being discussed here as applied linguistic artefacts, be somewhat unique.

In addition to the typical principles that apply to test design, and to which professional test designers may be held accountable, there is a further set of norms for the design of applied linguistic artefacts. This category of norms or design principles relates not to the concrete, factually different artefacts, each with their own unique identity, but to a set of general design conditions for technically qualified artefacts. I turn finally to a discussion of these.

General principles of test design

The principles for the design of the technically qualified measurement instrument that we call a language test derive from the relations or connections that its leading technical function has with all other modes of reality (Weideman, 2009). These analogical connections are important, because they provide the systematic basis from which applied linguistic conceptualization proceeds. Several of these analogical connections have already cropped up in the discussion thus far.

The connection of the leading technical function of test design with that of the sphere of energy-effect, the physical modality of our experience, has been evident in the discussion about the technical (in the sense of designed, or planned) *validity* of such a test. When we conceptually explore it, that analogical concept informs us that the effect of employing a test causes a score to be generated. The administration of the instrument has a certain technical force or effect, resulting in a score. That technically stamped force or adequacy of the measurement, also referred to as its effectiveness or validity, is therefore of great importance to language testers, and (even when they wish to avoid it, as we have noted) they formulate on the basis of this concept the principle of test validity. If a test is not a valid measure, that is one indication of it not being a responsibly designed instrument. Of course, a test must also yield a consistent measurement. That requirement or design principle derives from the link between the leading technical dimension of a test and the kinematic mode of experience, a mode in which we originally encounter the ideas of regularity and consistency. Thus test *reliability* becomes another, and indeed related principle of test design. As has been suggested by a reviewer, these two principles may indeed be related respectively to the "evaluation inference" in Kane's proposed argument chain (Kane 2011: 4, 8) as well as – perhaps more obliquely - to what is called the "generalization inference" (Kane 2011: 8, 10). The latter is definitely also related to the spatial substratum of the technical measurement, that is, to the principle that requires that the *range* of the generalization must be specified for a score interpretation to hold (Kane 2011: 10).

We encounter a third, and once again related analogical technical concept when we begin to ask about how the score that the test has generated – the effect caused by the measurement - is transformed into a result that is interpretable. Surely a number, or a level (category A, B, C…) on its own is meaningless? To arrive at a sensible technical interpretation of the score, we need to conceptualize the analogical relation between the technical mode and the lingual modality, in which meaning, expression and signifying are primary. Perhaps, as the same reviewer has commented, that may be close to what, in an argument-based approach, might be termed an explanation inference. For a sensible, meaningful interpretation of a score that yields a useful result, we also need to bring another analogy into play, that has already been referred to above: the technical rationale for the test, its construct, that we find in the interplay between the analytical and the technical, and in which the instrument indeed finds its (theoretical) grounding or basis function. Without a theoretically defensible definition of what it is that we are measuring, the test results would be uninterpretable. Hence the emphasis in the orthodox view of test validation on construct validity, as the technically justifiable basis for the interpretation of the results.

It should by now also be clear that the principles identified thus far, relating to test reliability, validity, theoretical defensibility and technical meaningfulness, are

distinguishable and distinct design conditions, but are actually also tightly related. That too, in my view, is what the current orthodoxy on test validation has tried to capture conceptually, in the requirement of a unitary concept of validity that requires a coherent argument to be systematically presented for demonstrating the quality of a test. It does so by bringing into the argument a multiplicity of data sets and analyses (in contrast to the pre-orthodox view of breaking validity into various components: construct, content, concurrent, predictive; as if any or either of them on their own would be sufficient: cf. Davies & Elder, 2005: 798). Yet we should note that the concept of a (technical) unity within a potential multiplicity of conditions and data is unthinkable without teasing out its conceptual basis. That conceptual basis is to be found in the analogical relation between the technical and the numerical modes. It is in the numerical aspect that we first encounter the notions of the one and the many. Once again, a normative requirement derives from the connection that we make between the technical mode and another, in this instance: the numerical.

Where do the other criteria referred to in the discussion above fit in? We have already noted that the technical appropriateness of the measurement that is so prominent in Messick's argumentation derives from the link between the technical and the social. A test is appropriate if upon implementation it has the intended and desired fit with the social milieu in which it has been employed (see Elder, McNamara, Kim, Pill & Sato, 2016). Usefulness, the idea prominent in Bachman and Palmer's (1996) concepts, of course relates to the connection between the technical and the economic mode, yielding the requirement of technical utility. It is interesting, again, that these and other requirements for test design hang together: a test has to yield a valid score, and its measurement has to be based on a theoretical rationale, before that score can be interpreted as a meaningful result of the measurement. Moreover, the interpretation has to be appropriate for the social or institutional context in which the test has been administered, and that is actually a precondition for its utility.

In the same way, the transparency or openness of the technical design to scrutiny not only by peers, but also by the public is a prerequisite for its accountability. Without another lingual analogy in the technical sphere, the provision of sufficient and significant information about the making, working and uses of the technical instrument, its juridical and political functions cannot be disclosed. The technical instrument has to be shown to do full justice to the ability that it is measuring. So indeed, as Kane (2010: 177f.) has observed, the various requirements, such as fairness and validity, are interrelated. They apply concurrently and sequentially: technical fairness, the compassionate, respectful, non-discriminatory, and caring use of language tests (or 'beneficence' as it is called in the preamble to the ILTA Code of Ethics; cf. too Green, 2014: 58) echoes the ethical connections with the technical, but it cannot properly be achieved, as the same code explains, before we have accessibility to technical data and information. At the same time, we should acknowledge that fairness so defined is not separate from, or a mere afterthought,

to the other requirements. The general principles of test design depend on one another, but, once disclosed, are simultaneously applicable.

There are further examples of general principles of test design, that have been more fully articulated and explained as constitutive and regulative conditions for test design (Weideman, 2009, 2014), or what may in shorthand form be described as necessary and sufficient conditions for assessment design. For a more complete explanation, I therefore refer to these. The principles identified here have been useful to me and fellow language test designers, because we could use them as starting points for measuring the worth of the instruments that we design. We could ask and check, for example, whether a test is properly differentiated (a principle that derives from the organic analogy within the technical), or whether it is intuitively appealing (its sensitive analogy, usually referred to as the face validity of the test), efficient (another economic analogy), imaginative and creative, and aligned with language development needs (possibly aesthetic analogies), along with the questions of adequacy/validity, theoretical defensibility, accessibility, accountability, fairness and others that are conventionally considered.

One of the principles mentioned only in passing here deserves further attention. That is that the technical range of a language test is limited. This is an analogically conceptualized condition for test design that clearly stems from the relation between the technical modality and the spatial. It is this condition, of the technical environment or space that defines the context of the test design and its eventual administration (see Elder et al., 2016), that has provided most of the backup for the argument against the modernist concept that the score of a test is valid merely because it derives from a supposedly scientific measurement. It is this modernist belief that is contested in arguments for validation, and against the view that validity is an inherent quality of a test. While both have some credibility, it is the principle of acknowledging the specific technical context of the administration of the test that shows a way out of the dilemma: if each test is administered, its results interpreted, and efficiently used in a specific context, there cannot be a test that is valid for all time and for all environments, the modernist pretence. It is the application of the principle that the measurement of a test is also contextually limited, that still allows us, on the other hand, to seek the validation of that measurement instrument, and to acknowledge, further, that over time and over several administrations, speaking of its validity in that limited sphere is “not just a trick of semantics” (Davies & Elder, 2005: 798).

The reference to the contextual specificity of a language test is relevant in another sense to the typical and general principles of test design outlined in this contribution. These principles are useful to test designers and developers, but they are just that: principles, that need further interpretation and specification in the process of test design. For example, the principle of adequate differentiation may lead test designers to consider more and other subtests to test a certain language ability, that earlier designs may have failed to bring in. But they may also conclude

that more rather than less internal differentiation in terms of the number of subtests employed is not necessary. The application in context of each principle needs to be responsibly considered. That responsible consideration is facilitated by the fact that they fit into a framework of applied linguistic design principles (Weideman 2016: chapter 11) that shows them as distinct but interdependent conditions for test design, production and use. In a word, these test design principles then fit into a theory of applied linguistics that provides the desired coherence to their application, and serves to overcome conceptual confusion. One comment received on an earlier draft of this is that these technically qualified concepts and ideas may well have been identified by other scholars, a point with which I not only fully agree, but which I think is also amply illustrated in the examples of such conceptualizations discussed above. This should not be surprising, since all of us work with the same states of affairs: we conceive, design, draft, administer and use language assessments. Yet to conceptualise these shared professional realities clearly and without conceptual confusion would greatly facilitate the design and implementation processes of the instruments we make. The argument of this contribution is that currently we have greater obfuscation than clarity, and that this holds back the responsible design of language tests. It proposes a way out of that conceptual quagmire.

Is there an encompassing, overarching consideration in language testing?

Language testing research has long given prime importance to the validity and validation of language tests. It is unlikely that this ‘overarching’, unitary conceptualization of the quality of tests will yield any ground. As the discussion above has shown, all of the modifications have either come to nothing or have not yet gained enough traction to push validity out as encompassing everything: promoting usefulness, interpretability or fairness to prime position has in some respects made us gain ground, but validation and validity, the subjective and objective sides of the technical force of a test, still reign supreme. Moreover, they do so while they are clearly only one set of criteria among many for ensuring the quality of a responsibly designed measurement.

However settled that belief in validation and validity may be, the conceptual contradictions that that position generates have persuaded me that we nonetheless must make the attempt. In gaining conceptual clarity, there are concomitant gains for test design. If there is anything further to be learned here, it is that language test design may be viewed as a response to certain typical and general conditions. In the relationship between the (subjective) making of the test as an artefact that presents us with a response to certain design principles, and the designed object, the actual test itself, lies a potential way out of the impasse. Forced to choose, therefore, I would venture to propose that responsibility in design encompasses not only test reliability and validity, interpretability, usefulness and efficiency, but also ethicality, accountability, care and integrity, and indeed the reputability and

trustworthiness that a test of good quality may gain over time. In the idea of responsible design all of these are distinct though related principles for designing language tests.

References

Arzubiaga, A.E, Artiles, A.J., King, K.A. & Harris-Murri, N. 2008. Beyond research on cultural minorities: Challenges and implications of research as situated cultural practice. *Exceptional Children* 74(3): 309-332.

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.). *Experimenting with uncertainty: Essays in honour of Alan Davies*. Cambridge, UK: Cambridge University Press, pp.109-116.

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

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

Boyd, K. & Davies, A. 2002. Doctors' orders for language testers: The origin and purpose of ethical codes. *Language Testing* 19(3): 296-322.

Davies, A. 1997. Demands of being professional in language testing. *Language Testing* 14(3): 328-339.

Davies, A. 2008. Accountability and standards. In Spolsky, B. & Hult, F.M. (Eds.). *The handbook of educational linguistics*. Oxford: Blackwell, pp. 483-494.

Davies, A. 2011. Kane, validity and soundness. *Language Testing* 29(1): 37-42.

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

Elder, C., McNamara, T., Kim, H., Pill, J. & Sato, T. 2016. Interrogating the construct of communicative competence in language assessment contexts: What the non-language specialists can tell us. [In this volume].

Joseph, J. 2016. Extended/distributed cognition and the native speaker. [In this volume].

Green, A. 2014. *Exploring language assessment and testing*. New York: Routledge.

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

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

Kane, M.T. 2010. Validity and fairness. *Language Testing* 27(2): 177-182. DOI: 10.1177/0265532209349467.

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

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

Kunnan, A.J. 2004. Test fairness. In Milanovic, M. & Weir, C. (Eds.). *Studies in language testing,* 18. Cambridge: Cambridge University Press, pp. 27 – 45.

Kunnan, A.J. (Ed.). 2000. *Studies in language testing 9: Fairness and validation in language assessment.* Cambridge: Cambridge University Press.

Lee, Y-J. 2005. Demystifying validity issues in language assessment. *Applied Linguistics Association of Korea Newsletter*. October. Available http://www.alak.or.kr/2_public/2005-oct/article3.asp. Accessed 25 April 2008.

McNamara, T. 2006. 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.). *Test validity*. Hillsdale, New Jersey: Lawrence Erlbaum Associates, pp. 33-45.

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

Popham, W.J. 1997. Consequential validity: Right concern – wrong concept. *Educational Measurement: Issues and practice*. Summer 1997: 9-13.

Rajagopalan, K. 2016. Prescription, language politics and the field of applied linguistics: A tribute to Prof. Alan Davies. [In this volume].

Rambiritch, A. 2012. *Accessibility, transparency and accountability as regulative conditions for a post-graduate test of academic literacy*. PhD thesis, Bloemfontein: University of the Free State. Available: http://hdl.handle.net/11660/1571.

Strauss, D.F.M. 2009. *Philosophy: Discipline of the disciplines*. Grand Rapids, MI: Paideia Press.

Van der Walt, J. 2012. The meaning and uses of test scores: An argument-based approach to validation. *Journal for language teaching* 46(2): 141-155. DOI: http://dx.doi.org/10.4314/jlt.v46I2.9.

Van Dyk, T. 2010. *Konstitutiewe voorwaardes vir die ontwerp en ontwikkeling van 'n toets vir akademiese geletterdheid*. PhD thesis. Bloemfontein: University of the Free State. Available: http://hdl.handle.net/11660/1918.

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

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

Weideman, A. 2012. Validation and validity beyond Messick. *Per Linguam* 28(2): 1-14.

Weideman, A. 2014.Innovation and reciprocity in applied linguistics. *Literator* 35(1), Art. #1074, 10 pages. DOI: http://dx.doi.org/10.4102/lit.v35i1.1074.

Weideman, A. 2016. *Responsible design in applied linguistics: Theory and practice*. Forthcoming from Springer.

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

Xi, X. 2010. How do we go about investigating test fairness? *Language Testing* 27(2): 147-170. DOI: 10.1177/0265532209349465.