AUTHOR:
Albert Weideman
https://orcid.org/0000-0002-9444-634X

AFFILIATION:
Professor of Applied Language Studies at the University of the Free State

CORRESPONDENCE TO:
Albert Weideman

EMAIL ADDRESS:
albert.weideman@ufs.ac.za

DATES:
Published: 29 Mar 2019

HOW TO CITE THIS ARTICLE:
Weideman, A., 2019. Degrees of adequacy: the disclosure of levels of validity in language assessment. KOERS — Bulletin for Christian Scholarship, 84(1). Available at: https://doi.org/10.19108/KOERS.84.1.2451

Degrees of adequacy: the disclosure of levels of validity in language assessment

Abstract

The conceptualization of validity remains contested in educational assessment in general, and in language assessment in particular. Validation and validity are the subjective and objective sides of the process of building a systematic argument for the adequacy of tests. Currently, validation is conceptualized as being dependent on the validity of the interpretation of the results of the instrument. Yet when a test yields a score, that is a first indication of its adequacy or validity. As the history of validity theory shows, adequacy is further disclosed with reference to the theoretical defensibility ("construct validity") of a language test. That analogical analytical disclosure of validity is taken further in the lingually analogical question of whether the test scores are interpretable, and meaningful. This paper will illustrate these various degrees of adequacy with reference mainly to empirical analyses of a number of tests of academic literacy, from pre-school level tests of emergent literacy, to measurements of postgraduate students' ability to cope with the language demands of their study. Further disclosures of language test design will be dealt with more comprehensively in a follow-up paper. Both papers present an analysis of how such disclosures relate to a theoretical framework for responsible test design.

Key words: *language testing; validity; test design; theory of applied linguistics*

Opsomming

Grade van toereikendheid: die ontsluiting van vlakke van geldigheid in taaltoetsing

Om geldigheid te konsepsualiseer bly 'n betwiste saak in opvoedkundige meting in die algemeen, en in taalassessering in die besonder. Geldigmaking en geldigheid kan respektiewelik opgevat word as die subjektiewe en objektiewe kante van die sistematiese argument wat gevoer kan word vir die toereikendheid van toetse. Tans word geldigmaking gekonseptualiseer as afhanklik van die interpretasie van die resultate van die instrument. Tog is dit so dat wanneer 'n toets 'n punt oplewer, dit 'n eerste aanduiding is van sy geldigheid. Soos die geskiedenis van geldigheidsteorie ook aantoon, word daardie toereikendheid verder ontsluit met verwysing na die teoretiese regverdiging (konstrukgeldigheid) van 'n taaltoets. Daardie logies-analitiese ontsluiting van geldigheid word verder geneem in die analogies linguale vraag: Is die toetsresultate interpreteerbaar en betekenisvol? Hierdie artikel illustreer hierdie verskillende grade van geldigheid met verwysing na empiriese analises van toetse van akademiese geletterdheid, vanaf voorskoolse toetse van ontluikende geletterdheid tot by metings van nagraadse studente se vermoë om die eise van akademiese diskoers te hanteer. Verdere ontsluitings van taaltoetsontwerp word vollediger hanteer in 'n opvolgartikel. Beide artikels bied 'n analise van hoe sulke ontsluitings verband hou met 'n teoretiese raamwerk vir verantwoordelike toetsontwerp.

Kernbegrippe: *taaltoetsing; geldigheid; toetsontwerp; teorie van toegepaste taalkunde*

1. Validity and validation: continuing contestation

In language testing, as in educational assessment in general, there is ongoing contestation about validity and validation. This paper and its follow-up (Weideman 2019b) take as their starting point the consideration that a lack of conceptual clarity may lie at the centre of this debate, a concern that has not yet been sufficiently investigated or attended to. What is more, what conceptualization is attempted often assumes the form of adding further concepts and ideas to the already overburdened notion of 'validity'. Such additions, like 'usefulness' (Bachman 2001: 110; Bachman & Palmer 1996: 17), 'fairness' (Kunnan 2000; 2004), or 'meaningfulness' – associating validity primarily with the interpretation of the scores of tests (Kane 2011: 3) – may be well intentioned, but they fail to achieve a systematic clarification of the concept. The extent of the ensuing conceptual confusion can be seen in further claims, such as the one by Fulcher and Davidson (2007: 15) that the "notion of test 'usefulness' provides an alternative way of looking at validity", or the proposal of Xi (2010: 167), that calls for the "integration of fairness into validity". In his contribution to a symposium on the principles of test validation at the annual Language Testing Research Colloquium (LTRC), Tannenbaum (2018), standing firmly in the 'interpretive' paradigm promoted by Kane (2011), even went so far as to claim: "I equate meaningfulness with validity."

These are all clearly contradictory notions: Bachman and Palmer's (1996) definition of 'usefulness', far from providing an alternative way of looking at validity, in fact includes construct validity as a component part. Conceptually, a component of something cannot also be the entity of which it is a part, in the same way that the sail, the keel or the rudder of a boat, though constituent parts of it, cannot be the boat, or, to take a more abstract example, that none of the three sides of a triangle can be the triangle. In language testing, the confusion starts when meaningfulness, or usefulness, or fairness is promoted to 'primacy' (Kunnan 2000: 1) or "most important consideration" (Bachman & Palmer 1996: 17; Bachman 2001: 110). In such a case there seem to be only two alternatives: either expand the concept of validity to encompass all of these, and perhaps more (ignoring the contradictions that thus arise), so that validity embraces everything, or subsume validity under, or equate it with the new concept that has been promoted to primary consideration in test design.

An example of the expansion of validity can be found in the observation by Davies (2008: 491), with reference to the Code of Ethics of the International Language Testing Association (ILTA), that "... what we are seeing in the professionalizing and ethicalizing of language testing is a wider and wider understanding of validity". In that case ethical considerations in language testing, such as treating those taking a test fairly, with care and compassion, and with due regard for the consequences of making the results of the test known, are mere extensions of the concept of validity. It is not as if language test professionals are unaware of the conceptual difficulties of such extension: Davies and Elder (2005: 799) acknowledge, for example, that, by introducing the notion of consequential validity or test 'impact', Messick has added "to the problems of validity by extending its scope into the social and the ethical". Thus, as we have already noted above, Xi (2010: 167) also calls for the "integration of fairness into validity". Yet this extension is contradicted by statements where the opposite is claimed, and where, instead, ethics includes validity: "While I can accept that ethics in language testing does include validity, whether it is wider in scope remains an unresolved question", is the observation made by Davies (1997: 335). A further example of the alternative, where validity is swallowed up by another concept or idea, is Tannenbaum's (2018) claim, already referred to above, that meaningfulness *is* validity. It is therefore not surprising that some wish to steer a middle course: Kane, who sits on the interpretative side of the paradigmatic fence in his defence of the current orthodoxy, claims that fairness and validity are different emphases, but 'intertwined' (2010: 177f.). Like the call for the integration of fairness into validity, however, that claim, when examined, does not lead to conceptual clarification, or to an understanding of the nature of the intertwinement.

At most, it shows that there may be further options than the two described here.

Instead of choosing either of these routes, this paper and its follow-up (Weideman 2019b) will explore another alternative: that the various additions to and further interpretations of validity may be conceptualized as disclosures of that concept. Such disclosure is both a historical unfolding of the concept over time, and a systematic, conceptual opening up of the notion. This paper will employ the terms 'disclosure', "opening up" and 'unfolding' as synonymous, and in the theoretical and methodological sense in which they were originally conceived (Dooyeweerd, 1957: 284f., *et passim*), and subsequently elaborated (Dooyeweerd, 2012a: 74; 2012b: 104, 158) in order to be applied to historical analyses in general. Such a characterization of validity, however, needs a fairly robust theoretical and conceptual framework, and one that may not yet have been sufficiently developed in applied linguistics. In order to make sense of the notion of 'disclosure', we in fact need a theory of applied linguistics (Spillner, 1977) that can do justice to both the historical and the systematic dimensions of the discipline.

This paper will not claim that achieving greater conceptual clarity will resolve all the contradictions and issues thrown up by the debate on validation and validity. Rather, it will primarily attempt to demonstrate that one may achieve a more meaningful and also theoretically defensible understanding of the evolution of the concept of validity. In addition, we should note at the outset that each reconceptualization of that concept that will be discussed here can be shown to emanate from different paradigmatic orientations. Though these orientations will be referred to where relevant, there is not enough room here for a full discussion of paradigmatic bias or ideology (Weideman, 2018) in each conception of validity to be discussed. The focus will therefore in the first instance not be on that, but rather on clarifying the underlying concepts and ideas that inform the gradual unfolding of our understanding of 'validity' over time, and the way that those insights, despite ideological bias, have deepened our understanding of it. In a word: by offering a systematic analysis of the historical development of the concept of validity over time, it will attempt to interpret the contradictions referred to here as potentially complementary, though in the end not sufficient, perspectives on that concept.

The first part of the argument below will show how, even in the now abandoned early conceptualization of validity, there was already a recognition of the relation between the subjective intentions of the test designer and the objective quality or adequacy ('validity') of the test. The subsequent acceptance that there is a subjective side to the process of validation, in which a persuasive argument for the validity of the measuring instrument (the designed object) is brought together, is therefore not as novel as it may be presented in the current orthodoxy. In the section following that, the methodology used is set out, before the several further disclosures of validity are analysed, taking a cue from the observation made many years ago by Thyne (1974: 5) that "... validity is not all-or-nothing, but a matter of degree..." To conclude, the issue of conceptual clarity is once again brought to the fore. In the follow-up paper (Weideman, 2019b) the question will be posed whether the process of 'validation' might perhaps not be more productively conceived as one of checking whether the test was responsibly designed. Throughout, statements made by commentators on language testing will be taken as serious claims, without resorting to speculation (e.g. "analytical sloppiness"), or coming up with excuses ("unthinking choice of terminology") for why their statements are sometimes contradictory. In the end, one cannot advance in conceptualising responsible language test design without striving for clarity about what that involves.

2. The validity of an instrument, and the process of its validation

If we look at the way that the current orthodoxy defines validity, we note that most of the discussion focuses on a claim made by Kane (1992:527): that validity "is associated with the interpretation assigned to test scores rather than with the scores or the test." Leaving aside for the moment the problematic differences in formulation, we may observe that, historically, the origins of the claim are often traced back to Messick's statement (1980:1023; cf. too 1981:18) that "Test validity is … an overall evaluative judgment of the adequacy and appropriateness of inferences drawn from test scores." Phrased differently, the further claim is made that the test itself cannot then be valid, but only the interpretations drawn from the scores it produces. In that view, as Davies and Elder (2005:809) remark, "validity is not tucked up in the test itself but rather resides in the meanings [interpretations] that are ascribed to test scores". Validity is not, and cannot in this view be a characteristic of the test, since it is dependent on the interpretation offered for test scores, and here the contestation begins.

This view was challenged by Popham (1997), and subsequently by Borsboom, Mellenbergh and Van Heerden (2004). Their objections to eliminating validity as a property of a test and by implication to making validity subject to interpretation appear to defend the claim of the initial, primitive and undisclosed concept of validity: that a test has validity if it measures what it set out to measure. Their detractors, like McNamara and Roever (2006:250f.), respond that clinging to that initial conceptualization of validity will "strip validity theory of its concern for values and consequences and … take the field back 80 years."

The implication of the initial view of validity as being a mark of the quality of a test if it measures what it was intended to measure is that, in producing a score, the test as the measuring instrument demonstrates that it has the instrumental or technical *force* to produce a result, which is an *effect* in line with the intention of its designer. That conceptual relation, between technical force and technical effect, has a two further implications. The first of these is that such cause-effect relationships originally belong to the physical dimension of reality, an observation that I shall return to in more detail below. The second is that if the outcome (effect) of the application of the measuring instrument (the language test) is in line with the designed intention of the test developer, that shows that there is a subject-object relationship: between the technically stamped, subjective design intention of the test maker, and the effect of the technical object (the test) that has been employed to make the measurement. Thus a technical subject-object relation already lies at the foundation also of that first, as yet undisclosed concept of validity: a language test is the outcome of subjective agency and design intention. Those who design the test cannot escape their full humanity and subjectivity in the planning and designing of an instrument, the technical object that they intend to use to measure language ability.

The effect of the measurement, the score, is indeed itself also a technical object, but it is not 'objective' in the popular sense of the word, viz. 'scientific' or unbiased. The technical shaping of any language test has all the marks of the subjective intentions of its human designer, of having been intentionally and purposely planned and prepared to measure the language ability it set out to measure. Though the design might at some stage – and not as early as might be thought (Weideman 2019a) – be informed by and even subsequently theoretically justified with reference to applied linguistic insight, that detour into science (Schuurman, 1972:361-362) taken by the language test designer does not make the test 'scientific': it remains a technically-characterized artefact, and its results, though expressed in numbers ('53%', 67%, 82%) or levels ('A', 'B', 'C', 'D'), are themselves technical objects. But such results are not 'objective' in the sense of being sacrosanct or virtually devoid of error, as they are purported to be in the modernist paradigm. The results mean nothing on their own – in that sense they are mere numbers or digits. They need interpretation by a competent technical agent, and that interpretation, since it is done (or initiated) by a human professional, is a subjective technical procedure. That is the legitimate response of postmodernist and related anti-modernist paradigms to the pretence of objectivity in modernism. I believe it is also the explanation for most of the contestation that surrounds the interpretive view of validity proposed by Kane and others, following Messick, when it is pitted against the initial, undisclosed perspective.

While it is a wholly legitimate consideration that the results of a language test need interpretation, the interpretive view is nonetheless wrong in implying that the primitive, initial view of validity does not recognize the human subject, the designer of the technical instrument, the language test. A language test is not intentionally designed other than subjectively, though it does produce a result, which may be conceptualized as a technical object.

Assigning 'validity' to a test, as one of its essential characteristics, has been widely discussed and contested. Fulcher and Davidson (2007:279), for example, encourage language test designers to consider whether the current orthodoxy, in denying that validity is a feature of a test, might not be mistaken:

> If a test is typically used for the same inferential decisions, over and over again, and if there is no evidence that it is being used for the wrong decisions, could we not simply speak of the validity of that particular test – as a characteristic of it?

Making a similar point about tests building a reputation over time, Davies and Elder (2005: 797) observe that "in some sense validity does reside in test instruments". It is, in their words, 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). These observations become even more relevant when we observe how those who see validity only in the interpretation of results use alternative terms such as 'adequacy' (Messick, 1981: 10; 1980: 1023) or 'effectiveness' to describe the quality of a test, or employ circumlocutions, like a "test … accomplishing its intended purpose" (Messick, 1980: 1025), or speak of tests that are "purported to tap aspects" of a trait (Messick, 1989: 48; 50, 51, 73). Sometimes even those who support the current orthodoxy would speak, without any reference at all to 'interpretation', of a test being "a valid measure of the construct" (McNamara & Roever, 2006: 109; see also 17), or of some items or even subtests being "construct irrelevant", and therefore invalid. No amount of interpretation can make clearly irrelevant items or subtests valid.

Can this dilemma be resolved? The current orthodoxy sees validation as a process of systematic argumentation, often built on hypotheses (Davies & Elder, 2005: 802ff.; Van der Walt, 2012). The argument brings together several data sets that support the decisions made, through plausible inferences, with reference to language test results (Kane, 2011:13; see also Van Dyk, 2010). Is the result of this process not the outcome implied by Fulcher and Davidson (2007) and Davies and Elder (2005:796), namely that the test is then shown, by a thorough process of validation, to be valid? This view is formulated by Davies (2011: 38) thus: "[W]e validate a test and then argue that it is valid."

Perhaps the resolution lies in viewing the subjective process of validation and the objective validity of the language test as two sides of the same coin: the technically qualified artefact, the designed language test, is a technical object that has to be shown, through a subjective process of technically stamped argumentation called 'validation', to be able to work. Objective instrumental validity is therefore the product side of the subjective process of language test design, of its subsequent production of results, and of its eventual validation. There is a technical subject-object relation at work here, and that needs to be conceptually fleshed out. After setting out the methodological framework for the rest of the analyses next, I shall then return to how this view of validation and validity can be further conceptualized in considering validity and its subsequent disclosures in language testing.

3. Methodology

The methodology to be used here to unravel the various conceptualizations of validity is an emergent theory of applied linguistics, that views language testing as an important subfield of that discipline, and applied linguistics as a whole as a discipline of design (Weideman, 2017). Since earlier descriptions already articulate the core of such a theoretical framework, the possible objections to it, and the advantages of employing it (Weideman, 2009), and since it has already been applied in a number of other studies (Van Dyk, 2010; Rambiritch, 2012; Pretorius, 2015; Keyser, 2017; Steyn, 2018; and to some extent also by Pretorius, 2018), I shall here confine myself to a number of essential distinctions that it allows the applied linguist to make.

The analytical starting point of this framework of applied linguistics is the abstraction, from a recognized applied linguistic artefact such as a language test, a language course or a language plan, of its leading or characterizing function, and to recognize that function as the technical dimension of experience (Strauss, 2009:127). The methodology is non-reductionist in its analytical intentions, since no dimension of experience is considered to be absolute, able to hold sway over other dimensions, or subsume them without antinomy and contradiction. Each dimension identified is characterized by a nuclear meaning, an idea that sets it apart from others. A language test is a technically stamped instrument because it has intentionally been designed to measure language ability. Schuurman (1972: 384) calls design the "centre of gravity" of the technical modality; an alternative formulation would be to conceive of 'design' as being the nuclear, or defining moment of the technical modality. However, that moment of design, the characterizing idea of the technical aspect that we abstract from an applied linguistic artefact such as a language test, is not a concrete object, like the test, but an aspect of it. That abstracted technical moment is a dimension that can theoretically be lifted out (abstracted) and thus subjected more closely to analytical scrutiny.

In the process of theoretical abstraction, however, the integrity of our experience re-asserts itself: the technical is by no means the only dimension of our world, and it is inextricably connected with all other dimensions of reality, that resist such a pulling away. That re-assertion is a reminder of the wholeness of our experience, which also has other dimensions, such as the aspects of number, of space and of movement, as well as the physical dimension, the organic, the sensitive, the analytical, the lingual, the economic, the aesthetic, the juridical, the ethical and the certitudinal modalities. In attempting to abstract the technical, we discover, upon analysis, that the other dimensions of experience leave traces, or analogies, within the modality that has been isolated for scrutiny. These echoes of other dimensions within the technical sphere, methodologically termed analogies, can be either retrocipations or anticipations, depending on whether the analogical moment refers to an aspect preceding the technical (for retrocipatory echoes), or is evidence of a trace of an aspect following it (for anticipatory reflections).

What is more, the conceptualization of such analogical reflections of the other dimensions within the technical yield the basis for concept-formation within the discipline of applied linguistics. Each analogical moment enables us to discern an elementary, fundamental or primitive concept or idea. The retrocipatory analogies within the technical that refer to earlier dimensions (in this case the numerical, spatial, kinematic, physical, organic, sensitive and logical) are methodologically termed technical concepts. These are constitutive, founding ('necessary') concepts, and include conceptual primitives relating, for example, to the reliability, validity, differentiation, intuitive appeal ("face validity") and theoretical defensibility or construct validity of the applied linguistic design. The anticipatory reflections, in turn, are technical ideas. They are, furthermore, termed regulative technical ideas, since, as will be shown in the analysis following, they become the lodestars that disclose the meaning of design. Examples would be the technical significance, efficiency or utility of an applied linguistic design, or its fairness and reputability. The constitutive concepts and regulative ideas remain characterized, however, by the aspect that is being theoretically abstracted, and is being subjected to scrutiny and analysis. The 'reflected' meanings of the other modalities therefore, in the case of the technical, are echoes within the technical sphere of those other dimensions of our experience.

Together, the sets of elementary technical concepts and ideas that will be discussed subsequently are illustrations of how the theoretical framework cursorily outlined in the previous paragraphs works; of the systematic insight that they bring to the essential concepts of validity and the subjective process of validation; and of the way that these conceptual primitives are further disclosed and opened up.

4. The undisclosed concept of validity

The undisclosed concept of validity pre-dates the views of Messick on validity (1980; 1981, 1988; 1989), views that are today considered by some to be nothing less than its culmination (Xi, 2008:179). This elementary concept that preceded it is nonetheless worth examining on its own, in its undisclosed conceptualization: that a language test is valid if it measures what it set out to measure.

This formulation clearly ties the technically qualified design of a language test to the analogical reflection of the physical dimension of reality, where one first encounters the notion of energy-effect. In its original meaning within the physical sphere, we observe the effect of a force that is applied, that operates or works; in the current case, however, one is dealing with an analogical, reflected concept. The analogical technical concepts in this case refer to the effect of the application of the measuring instrument: a result or score that is technically accomplished (in the sense of intentionally, by design). The working or operation of the language test as technically qualified measure is a cause of the designed effect (the score obtained). All of these descriptions rely on analogical technical concepts: when a language test therefore operates, or works, and yields a result, it does what its designed intention was, and is, in that as yet undisclosed sense, valid or, in Messick's terminology (since he wishes to avoid the term 'validity') adequate. The technical adequacy of a language test relates to the effectiveness of the test (it produces a result or score). Phrased differently: by measuring the language ability it was intended to do, it operates with the technical force required, and is then technically adequate, at least in an initial sense. These are clearly all retrocipatory analogical physical moments within the technical. They do not avoid or eliminate validity, but clarify it further, conceiving of validity or adequacy not only as a recognizable feature of the technical object, but also as a primitive, in the sense of a founding, fundamental technical concept that is itself not further definable. As Davies and Elder (2005:796) observe: "Validity is self-contained: Its definition is reflexive ('validity is validity') just like those other great abstractions: beauty, truth, justice."

When this is acknowledged, the initial, as yet undisclosed meaning of validity in language testing therefore enables the language tester to make a claim that the technical force of a test has been adequate to yield an effect. That cause-effect relationship is an analogical technical one. It derives from the original physical domain, but is a conceptual echo of that domain within the technical dimension of design, and thus within applied linguistic concept formation, since it is a discipline of design. The schematic representation (Figure 1) sets out all of the modalities that cohere with the technical sphere (the numerical, spatial, kinematic, right through to the certitudinal), and their derived, analogical meaning within the technical, as potentially reflected (*in italics*). In the case of the reflection of the physical sphere within the technical, we can conceptualize the analogical meaning in the form of a constitutive technical concept, as one relating to technical effectiveness, validity or adequacy. That link, the coherence of the technical and physical modalities, enables us to conceptualize technical validity in general, and then apply it to the typical case of a technically stamped, designed instrument to measure language ability, a language test.

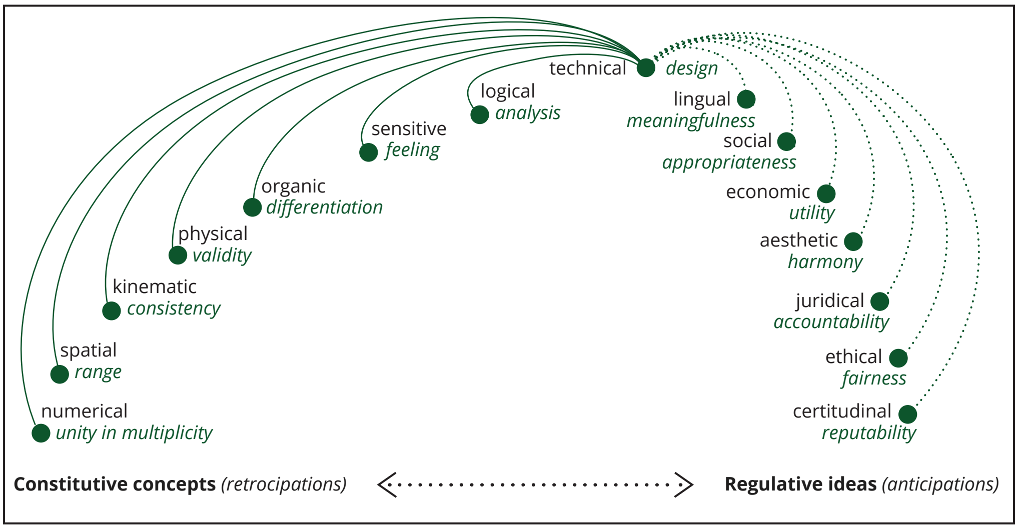


Figure 1: Coherence of the technical dimension with others (and their *traces*)

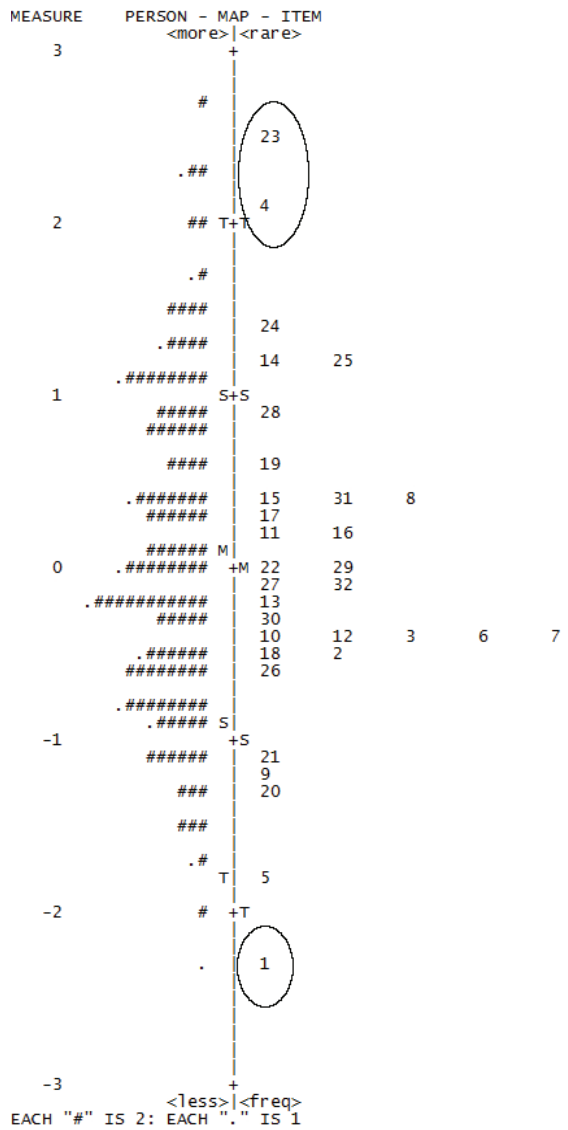
How do language testers know that their designed instrument works? An early indication of validity is found in analyses that attempt to determine whether the test as a whole works, and whether each component part (subtest) operates and performs adequately in relation to the others. In a validation argument for a language test, hypotheses may on this basis be formulated about the parameters for test and subtest performance. Let us take as an example the test-subtest correlation, as well as the subtest intercorrelations, in the refined pilot version of the Toets van Akademiese Geletterdheid vir Nagraadse Studente (TAGNaS) (Keyser 2017), set out in Table 1.

Subtest		Total test	Subtest 1	Subtest 2	Subtest 3	Subtest 4
Scrambled text 1	1	0.58				
Scrambled text 2	2	0.58	0.41			
Vocabulary (single)	3	0.69	0.31	0.35		
Vocabulary (double)	4	0.46	0.22	0.17	0.28	
Grammar & text relations	5	0.87	0.29	0.28	0.41	0.34

Table 1: Test-subtest and subtest intercorrelations in refined pilot version of TAGNaS Test 1

Van der Walt and Steyn (2007) advise that the parameters for subtest intercorrelation coefficients should be set fairly low, between 0.3 and 0.5, since each subtest is supposed to be testing a slightly different dimension or component of language ability, in this case of academic literacy. The test-subtest correlations, on the other hand, one would wish to be higher, desirably above 0.7, since each subtest has to make as substantial a contribution to the test as possible, i.e., have the maximum technical force or effect when applied. These are the assumed desirable parameters for the *technical adequacy* of the workings of the components of the test. In this pilot of TAGNaS, we notice that only two subtests, the Vocabulary (single) and the Grammar & text relations components, correlate that well with the overall test (shaded). However, in all but two cases (also shaded) the subtest intercorrelations are within the desired parameters. In their validation argument, the language test designers must in this case either attempt to recast the problematic items, or to repilot others, or perhaps choose to argue for a slight relaxation of the parameters. In the latter case, they may offer arguments relating to the few items making up the non-performing subtest (Vocabulary [double], that has only four items), the shortness of the test overall (in line with its purpose as a first level, screening test it is only 38 items in length), or even their experience with similar test designs. In each case, these arguments will, in the terminology of validation studies, be backed by warrants: sets of evidence that are systematically presented and evaluated.

The adequacy or validity of the instrument – its technical force – can be argued for and illustrated in numerous other ways. With larger data sets, for example, further evidence to identify test components or items that may corrupt the measurement (and therefore make it technically less adequate) may be brought into play. A Rasch analysis (Linacre, 2018) can be employed, as is suggested by Van der Walt and Steyn (2007), to determine the degree of fit between the general ability of individual candidates and the general level of difficulty of test items. In the case of the second tier of the TAGNaS pilot, it can be demonstrated that there is an adequate matching between person ability and item functioning. Keyser (2017) shows just how high the degree of fit there is in this language test: the item values stretch between -2.3 logits (with Item 1 as the lowest) to 2.5 logits (Item 23) in the variable map (Figure 2). That is well within the parameters of between -3 and 3 logits suggested by Van der Walt and Steyn (2007). What is more, the evidence of the misfitting items (1, 4 and 23) produced by this Rasch analysis is further confirmed by earlier analyses employing Classical Test Theory. There are thus several warrants for the technical adequacy of this test.



```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    3                +
                     |
                     |
                  #  |
                     |
                     |  23
                     |
                .##  |
                     |
                     |  4
    2            ## T+T
                     |
                     |
                 .#  |
                     |
               ####  |
                     |  24
              .####  |
                     |  14     25
          .########  |
    1               S+S
              #####  |  28
             ######  |
                     |
               ####  |  19
                     |
           .#######  |  15     31     8
             ######  |  17
                     |  11     16
             ###### M|
    0     .########  +M 22     29
                     |  27     32
       .###########  |  13
              #####  |  30
                     |  10     12     3      6      7
            .######  |  18     2
           ########  |  26
                     |
          .########  |
             .##### S|
   -1                +S
             ######  |  21
                     |  9
                ###  |  20
                     |
                ###  |
                     |
                 .#  |
                    T|  5
                     |
   -2             #  +T
                     |
                     |
                  .  |  1
                     |
                     |
                     |
                     |
                     |
                     |
   -3                +
               <less>|<freq>
EACH "#" IS 2: EACH "." IS 1
```

Figure 2: Person-item fit: Interpretasie van visuele inligting (Interpretation of visual information), TAGNaS first pilot

The arguments for the validation of a test are sets of evidence brought together to demonstrate the validity or adequacy of the language test as measuring instrument. Attempts to present them as not providing that, since tests do not possess validity, is a conceptual obfuscation.

5. Evidence for the reliability or technical consistency of a language test

To many who develop arguments in order to validate the design of a language test, the reliability with which such a test measures is simply part of its validity. However, as the expositions above, including those in Figure 1, show, technical consistency relates not to the analogical physical concept of technical adequacy, but precedes and thus supports it. The notion of consistency arises originally in the kinematic dimension of experience, the nuclear moment of which is regular movement. Analogically, within the technical sphere that retrocipation yields the concept of regularity and consistency, but now in a designed, instrumental, and technical sense. Since the technical consistency or reliability of a language test conceptually precedes its validity, it is supportive of its adequacy, rather than identical to it. Despite test reliability currently being treated as part of validity, those two are nonetheless still distinguished. Therefore language test designers will seek to find a test that measures language ability as consistently as possible, and may even use that as one of the claims or hypotheses about the test when it is being validated.

Test reliability is measurable, and the conservative index usually employed is Cronbach alpha, though there are also other, less strictly conceived indices, such as Greatest Lower Bound (CITO 2005: 17, 30, 37), as in the following illustration, taken from a TiaPlus (CITO 2005) analysis of a pilot test of the Assessment of Language for Economics and Finance (ALEF), a test designed to assess the language ability levels of prospective professionals-in-training in those domains. The respective subtests are Vocabulary in context (S1); Text comprehension (S2); Interpreting graphic & visual information (S3); Register & text type (S4); Grammar & text relations (S5); and Scrambled text (S6). The values of the two reliability indices employed are in bold.

	Test	S1	S2	S3	S4	S5	S6
Number of testees	127	127	127	127	127	127	127
Number of items	80	18	20	125	5	5	20
Average test score	38.02	11.30	11.23	5.63	2.33	1.57	5.97
Average P-value	47.53	62.77	56.14	46.92	46.61	31.34	29.84
Co-efficient alpha	**0.93**	**0.82**	**0.77**	**0.68**	**0.79**	**0.68**	**0.93**
Greatest Lower Bound	_	**0.92**	**0.89**	**0.80**	**0.88**	**0.79**	**0.98**

Table 2: Extract from summary statistics: ALEF pilot (2019)

As is evident, the test overall is highly reliable: any measure above 0.9 in the case of co-efficient alpha for the test as a whole is considered to be indicative of a very high degree of consistency. What is more, the subtests that contribute more substantially to the overall reliability, namely Vocabulary in context (S1) and Scrambled text (S6), can also be identified, and perhaps inform subsequent revisions to the design of such a language test. In sum, the analogical kinematic link within the technical allows us to conceptualize and understand the various meanings of the notion of the technical consistency of a language test.

6. Construct validity and score interpretation as disclosures of the degree of adequacy of a test

The credit of one particular disclosure of the concept of test validity must go to Messick's promotion of validity as a unitary concept (1980, 1981, 1989), that encompasses and surpasses in importance all previously identified 'components' of validity (e.g. content, criterion and other kinds). This unifying concept is to be found in construct validity, definable as the theoretical idea underlying the language ability that is being tested. In the examples being referred to here that idea or construct is academic literacy, defined as the mastery of academic discourse in which distinction-making and analytically qualified argument building are central (for further definitions and discussion, see Read, 2015; Pot & Weideman 2015; Patterson & Weideman 2013a, 2013b). Only if one knows, and can articulate clearly what the ability is that is being measured, and can preferably do so with reference to a theoretical idea of that language ability, can the 'validity' of the measurement become theoretically defensible. Construct validity is analogically linked, therefore, to the logical dimension of experience, as being the theoretical rationale that supports the technical (designed) measurement.

The construct of a language test, apart from the requirement of having to be theoretically defensible, also has to be operationalized: the idea of language ability that it proposes must be further articulated and specified, so that a blueprint for the test emerges (Weideman, 2019a). Schuurman (1972:385) calls the articulation of such a blueprint for the design an important initial disclosure of the design: the opening up of the technical function of design to anticipate its expression in the form of a complete plan is a reflection, within the technical, of the lingual mode of experience. The logical retrocipation within the leading technical function of the test and the lingual anticipation therefore work together. In signifying the design intentions in detail in the test blueprint, the test designer ensures that the construct itself is enhanced and disclosed.

The theoretical justification of the ability being measured ("academic literacy") remains a foundational (constitutive) condition, however, since it offers the basis from which another lingual anticipatory moment is accomplished: the interpretation of the test scores. Without a theoretically defensible and intelligible construct, there are no grounds for interpreting the scores of a language test. Without interpretation with reference to the language ability being measured, language test scores are meaningless instead of meaningful. For example, in the case of the academic literacy tests that are used as illustrations here, the interpretation of their scores would be meaningless without reference to the construct of academic literacy, defined as the ability to handle the demands of language for learning, or, in the formulation already given above, the mastery of academic discourse in which distinction-making and analytically qualified argument-building are central. Should the tests not measure that ability, they would clearly not have a defensible basis. When they do, they can, for example, meaningfully gauge the level of mastery, and allow scores to be interpreted with reference to the risk for students attached to each level. Table 3 gives an example of how scores on a highly reliable set of undergraduate tests, the Test of Academic Literacy Levels (TALL) and the Toets vir Akademiese Geletterdheid (TAG), have been interpreted:

Level / Code	Interpretation
1	Little to no risk of level of academic literacy interfering with academic performance
2	Less risk of level of academic literacy interfering with academic performance
3	Borderline: please consider taking a second-chance test, or self-classify as at risk
4	Clear risk of level of academic literacy interfering with academic performance
5	Very high risk of level of academic literacy interfering with academic performance

Table 3: TALL/TAG score interpretations: levels and associated risk

These interpretations are associated with ranges of test results, and derive from detailed recorded previous experiences, as well as initial referencing against the results of similar tests. Each level, in this record, is historically connected with average performance over the years, with different standard deviations recorded in various administrations also factored in, and the scores (numerical values) associated with each level. In the recording of these interpretations for such tests over time, we encounter yet another anticipation within the technical of the lingual aspect of language testing.

The methodological point is this: While no score has meaning on its own, since all scores need interpretation, the view of validity as residing only in the interpretation of test results may obscure the fact that we still need a score, an objective result, before any subjective interpretation can be done. To imply that the quality of the score does not matter, merely its interpretation, is to devalue the adequacy of the instrument. Without an adequate or valid instrument, no valid technical interpretation is possible. There is a technical subject-object relationship between the subjective process of score interpretation and a score that has been obtained by employing an adequate instrument. Once we have the score, its technical significance can be expressed.

7. Preliminary conclusion

It is probable, as has been noted in passing above, that the contestation about what validity is in the initial, undisclosed view, and what has been called here the interpretive stance, that the “linguistic turn” that influenced so many other disciplines also made itself felt in the views of Messick and his followers. An analysis of the echoes of the lingual aspect of experience within the technical sphere makes the discussion of these divergent views conceptually more accessible.

Having employed a particular conceptual methodology, the preliminary conclusion of the analysis is that the technically stamped reliability and validity of a test are two distinguishable facets of test quality that are often lumped together. Both are disclosed when we bring in the requirement that the construct of a test must be theoretically defensible. Finally, all three of these constitutive requirements for test design are further opened up when the technical meaningfulness of the results of a test is brought into play. While all work together, they are conceptually distinguishable, and cannot conceptually be brought together under the umbrella of validity without contradiction. When the further disclosures of requirements for language test design are examined, as I propose to do subsequently (Weideman 2019b), that conclusion will become even more significant.

References

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, p. 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. https://doi.org/10.1037/0033-295x.111.4.1061.

CITO. 2005. *TiaPlus user's manual*. Arnhem: M & R Department.

Davies, A. 1997. Demands of being professional in language testing. *Language Testing,* 14(3):328-339. https://doi.org/10.1177/026553229701400309.

Davies, A. 2008. Accountability and standards. (*In* Spolsky, B. & Hult, F.M. *eds.* The handbook of educational linguistics. Oxford: Blackwell, p. 483-494.) https://doi.org/10.1002/9780470694138.ch34.

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, p. 795-813.)

Dooyeweerd, H. 1957. A new critique of theoretical thought. Volume II. Amsterdam: H.J. Paris.

Dooyeweerd, H. 2012a. Roots of Western culture: pagan, secular and Christian options. Grand Rapids, MI: Paideia Press. [Series: Strauss, D.F.M. General *ed.* 2012. The collected works of Herman Dooyeweerd; Series B, Volume 15.]

Dooyeweerd, H. 2012b. Encyclopedia of the science of law: introduction. Grand Rapids, MI: Paideia Press. [Series: Strauss, D.F.M. General *ed.* 2012. *The collected works of Herman Dooyeweer*d; Series A, Volume 8/1.]

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

Kane, M.T. 1992. An argument-based approach to validity. *Psychological Bulletin,* 112(3):527-535. https://doi.org/10.1037//0033-2909.112.3.527.

Kane, M.T. 2010. Validity and fairness. *Language Testing,* 27(2):177-182. https://doi.org/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. https://doi.org/10.1177/0265532211417210.

Keyser, G. 2017. Die teoretiese begronding vir die ontwerp van 'n nagraadse toets van akademiese geletterdheid in Afrikaans. Bloemfontein: University of the Free State (M.A. dissertation.) URL: http://hdl.handle.net/11660/7704.

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, p. 1-14.)

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

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

Linacre, J.M. 2018. A user's guide to WINSTEPS Ministep: Rasch-model computer programs. Chicago: Winsteps.

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. https://doi.org/10.1177/0265532211417210.

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, p. 33-45.)

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

Patterson, R. & Weideman, A. 2013a. The typicality of academic discourse and its relevance for constructs of academic literacy. *Journal for Language Teaching* 47(1):107-123. https://doi.org/10.4314/jlt.v47i1.5.

Patterson, R. & Weideman, A. 2013b. The refinement of a construct for tests of academic literacy. *Journal for Language Teaching,* 47(1):125-151. https://doi.org/10.4314/jlt.v47i1.6.

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

Pot, A. & Weideman, A. 2015. Diagnosing academic language ability: insights from an analysis of a postgraduate test of academic literacy. *Language Matters* 46(1):22-43. https://doi.org/10.1080/10228195.2014.986665.

Pretorius, M. 2015. The theoretical justification of a communicative course for nurses: Nurses on the move. Bloemfontein: University of the Free State. (M.A. dissertation.)

Pretorius, M. 2018. Accommodative competence: a Communication Accommodation Theory approach to language and communication training in nursing. Antwerp: University of Antwerp. (Ph.D. thesis.)

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

Read, J. 2015. Assessing English proficiency for university study. Basingstoke: Palgrave Macmillan.

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

Spillner, B. 1977. On the theoretical foundations of applied linguistics. *International Review of Applied Linguistics,* 15(2):154-157.

Steyn, S. 2018. A theoretical justification for the design and refinement of a Test of Advanced Language Ability (TALA). Bloemfontein: University of the Free State. (M.A. dissertation.)

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

Tannenbaum, R.J. 2018. Validity aspects of score reporting. Contribution to Symposium 4, Language Testing Research Colloquium 2018 (Auckland): Re-conceptualizing, challenging, and expanding principles of test validation.

Thyne, J.M. 1974. Principles of examining. London: University of London Press.

Van Der Walt, J.L. 2012. The meaning and uses of test scores: An argument-based approach to validation. *Journal for Language Teaching,* 46(2):141-155. https://doi.org/10.4314/jlt.v46i2.9.

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

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

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. https://doi.org/10.2989/salals.2009.27.3.3.937.

Weideman, A. 2017. Responsible design in applied linguistics: theory and practice. Cham: Springer International Publishing. [Online]. DOI 10.1007/978-3-319-41731-8.

Weideman, A. 2018. Positivism and postpositivism. In C.A. Chapelle (*ed.*). *The encyclopedia of applied linguistics*. Wiley & Sons. [Online]. https://doi.org/10.1002/9781405198431.wbeal0920.pub2.

Weideman, A. 2019a. Definition and design: aligning language interventions in education. Soon to be published to *SPIL Plus*.

Weideman, A. 2019b. Validation and the further disclosures of language test design. Submitted to *Koers*.

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

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