Validation and the further disclosures of language test design

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: 30 Mar 2019

HOW TO CITE THIS ARTICLE:
Weideman, A., 2019. Validation and the further disclosures of language test design. KOERS — Bulletin for Christian Scholarship, 84(1). Available at: https://doi.org/10.19108/KOERS.84.1.2452

Abstract

The subjective validation of language tests and their objective validity remain contested. The debate is clouded by a lack of conceptual clarity: every measure of the adequacy of a test is either subsumed under validity, or other test features are promoted to prime consideration. An alternative is to recognize the technical dimension of experience as the leading function of applied linguistic artefacts such as language tests. In its coherence with other dimensions of reality, the technical aspect generates echoes of those others. These references to other facets are the basis of technically stamped, applied linguistic concepts. While a previous analysis referred to a number of foundational concepts in test design, this paper examines the traces of the lingual, social, economic and ethical aspects within the technical. These are all regulative technical ideas, acting as lodestars in the design of language tests. They are disclosures of the technical meaning of test design. Moreover, they allow applied linguistics to conceive of them as design principles, which must be given concrete shape and form in the actual development of language tests. Acknowledging this state of affairs enables the applied linguist to develop a robust and non-reductionist theory of applied linguistics.

Keywords: *validity; validation; theory of applied linguistics; design principles*

Opsomming

Daar is geen eenstemmigheid oor wat die subjektiewe validering van taaltoetse asook hul objektiewe geldigheid behels nie. Die debat word vertroebel deur 'n gebrek aan konseptuele duidelikheid: elke maatstaf vir die effektiwiteit van 'n toets word óf ingetrek onder geldigheid, óf verhef tot belangrikste oorweging. Die alternatief is om die tegniese dimensie van ons ervaring as die leidende funksie van toegepaste taalkundige artefakte te erken. In sy samehang met ander dimensies genereer die tegniese aspek analogieë van die ander. Daardie verwysings na ander aspekte vorm die basis van tegnies gekwalifiseerde toegepaste taalkundige begrippe. Waar 'n vorige analise na 'n aantal funderende begrippe verwys het, ondersoek hierdie bydrae die linguale, sosiale, ekonomiese en etiese spore in die tegniese aspek. Hierdie analogieë is almal regulatiewe tegniese idees, wat optree as rigtingwysers in die ontwerp van taaltoetse. Hulle is ook ontsluitings van die tegniese betekenis van toetsontwerp. Terselfdertyd maak hulle dit vir die toegepaste taalkunde moontlik om hulle as ontwerpbeginsels te beskou wat in die ontwikkeling van taaltoetse konkrete beslag en vorm te kry. Om hierdie stand van sake ernstig op te neem is vir die toepgepaste taalkundige die beginpunt van die ontwikkeling van 'n robuuste en nie-reduksionistiese teorie van die toegepaste taalkunde.

Kernbegrippe: *geldigheid; geldigmaking; teorie van die toegepaste taalkunde; ontwerpbeginsels*

1. Conceptualizing the elementary concepts of language testing

The validation of language tests and their validity remain contested, not the least because of a lack of conceptual clarity. This is the second of two analyses of the degrees of adequacy in language tests, in which the argument for the benefits of conceptual clarity is taken further. The first analysis (Weideman, 2019) has attempted to show that if we burden the concept of validity by associating it exclusively with the interpretations of test scores, as is done in the current orthodoxy, we run into a number of contradictions.

Instead, it was argued that one should see validity not only as the adequacy of the technical object, the language test that accomplishes the measurement, but as a technical concept that can be disclosed in various ways. That disclosure of several degrees of adequacy is related in the first instance to the technical subject-object relationship that is evident in the subjective process of validation and the determination of the validity of the technical object, the designed test of language ability that is being used as the measurement instrument. It is related, secondly, to the way that the leading technical function of the language test coheres with a number of other dimensions of our experience.

The technical aspect has as a nuclear moment the idea of design (Strauss, 2009:127; Schuurman, 2009:417), which itself is not further definable. That key idea allows us not only to distinguish the technical function from other aspects of experience, but also to recognize it as the qualifying or leading aspect of applied linguistic plans, those deliberate and intentional designs that characterize not only language tests, but also language courses and language policies. Moreover, for theorists the technical aspect of experience defines the scope applied linguistics: it is a discipline of design.

How, in theorizing the field, are fundamental applied linguistic concepts and ideas conceptualized? The analyses presented here proceed from the starting point that applied linguistics is a discipline of design (Weideman, 2017), recognizing too that concept formation in that field will contribute to a theory applied linguistics. The methodology employed in both the former (Weideman, 2019) and in this analysis provides an explanation of how we can conceptualize applied linguistic 'primitives', those fundamental, elementary concepts and ideas on which the discipline is based. Those elementary technical concepts and ideas are also fundamental to language assessment, which is an important subfield of applied linguistics.

Though unique, the technical aspect of our experience is not its only, or even its most important dimension: it is one among many, therefore related to all the others, and by that token not absolute. In its coherence with other dimensions of reality, the technical aspect of our experience allows applied linguistic theorizing to identify the various analogies to other dimensions that are reflected, as referred or echoed meanings, within the technical. For example, in the case of the concept of the technical validity of a language test, that analogical concept is a reflection, within the technical, of the physical aspect. In the latter, the notion of cause and effect, of the application and working of a force to produce an outcome, is an original physical one. In the working and operation of a technically qualified (intentionally designed) language test, however, the notion of validity becomes an analogical one: that of technical validity, a concept expressing the working or force of an instrument intentionally designed to measure language ability. The technical or instrumental force is designed to produce an effect. That effect, the result of the measurement, is a test score. In the same way, the kinematic dimension of reality is reflected within the technical as technical consistency or reliability, since in the originally kinematic sense we encounter the idea of regular movement. Hence the elementary applied linguistic concept of the technical reliability of a language test. Similarly, in reflecting the numerical aspect, we encounter in our designs the fundamental applied linguistic concept of a technical unity within a multiplicity, a concept that will again be referred to below. Figure 1 shows the various dimensions of experience, from the numerical to the certitudinal, and a provisional formulation of their reflected, analogical meanings within the technical (in *italics*). Those analogies that emanate from dimensions that precede the technical, in the order they are presented in Figure 1, are called retrocipations (founding, constitutive reflections), and those reflections of dimensions following the technical are termed anticipations (forward echoes). The retrocipations are echoes in a constitutive, founding direction, and the anticipations are forward-looking analogies in the regulative direction. Retrocipations are thus founding ('necessary') elementary technical concepts, while anticipations are regulative ideas: lodestars that lead and disclose the design, so that, with reference to the lingual, social, economic, ethical and confessional dimensions, the instrument becomes technically meaningful, appropriate, useful, fair and reputable.

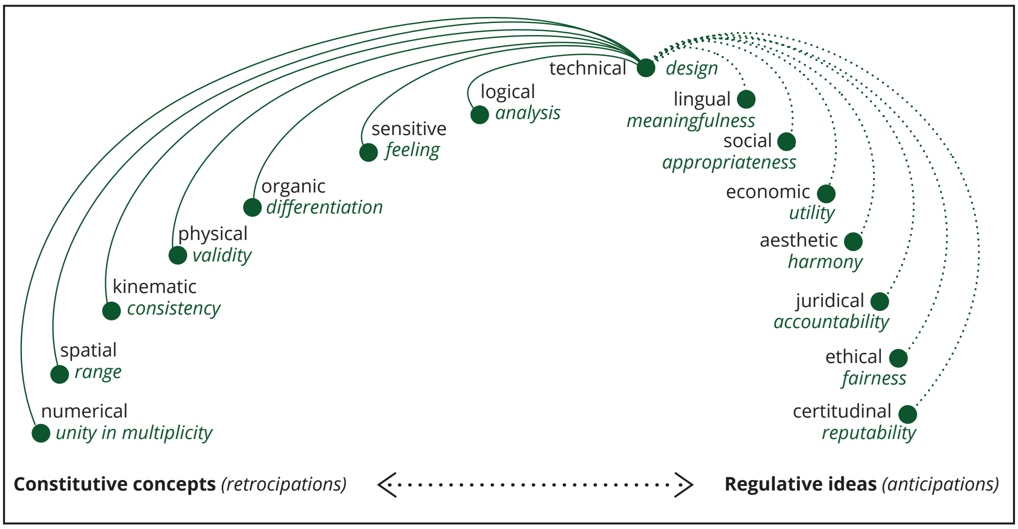

Figure 1: Coherence of the technical dimension with others (and their *traces*)

The conceptualization of such analogical reflections of the other dimensions of experience within the technical yields the basis for concept-formation within the discipline of applied linguistics. This paper will examine further disclosures of language test design in light of several regulative technical ideas, in particular those of the technical significance (an analogical lingual reference), the technical appropriateness or social and interactional fit of the language test with the institutional environment in which it is being employed (the social analogy), the technical utility (an economic anticipation) and the fairness of a test (an ethical disclosure of the technical). Specifically, it will examine whether such disclosures of language test design can meaningfully still be conceived of as measures of validity.

2. Validity: an overburdened concept

The overburdening of the concept of technical validity within language test design is evident when we consider that it has a multiplicity of meanings in the literature on this: that tests should measure consistently (a kinematic analogy within the technical sphere of design); that they should work in order to yield a result (a physical analogy); that their construct should be theoretically defensible (the logical link); and that their results, once they have become available, are also subject to interpretation (their leading aspect of design referring here to lingual expression and signification). That, in the preceding analysis (Weideman, 2019) has been interpreted as referring to the several degrees of the adequacy of language tests. It has also been noted how the interpretation of the results crucially rests upon the theoretical rationale for the language test. Without an articulated notion of what the language ability is that one is testing, it becomes impossible to give a meaningful interpretation of results. In the current consensus, such meaningful interpretation lends validity to the inferences that can justifiably be made and argued for from the test results.

The previous analysis has illustrated how the concept of validity, both in its primitive or undisclosed sense as the technical effectiveness of the test, and in its conceptualization as construct validity (the theoretical ground for the ability being tested), was disclosed when it was conceptualized as a quality of a language test that is opened up by the technical idea of meaningfulness. The latter disclosure crucially depends on the regulative idea of the technical interpretability of the results, which is a lingual analogy within the technical, denoting how inferences from test scores (the objective effects of the measurement) can be expressed and made significant and meaningful. Table 1 shows how scores on a highly reliable set of undergraduate tests, the Test of Academic Literacy Levels (TALL) and the Toets vir Akademiese Geletterdheid (TAG), may be, and has been interpreted by those affected by the test (students, lecturers, administrators):

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 1: TALL/TAG score interpretations: levels and associated risk

The lingual analogy is evident too in the observation that interpretations such as these are made on the basis of a detailed record of previous test results, and their use. That record connects a number of measures: of average performance over the years; of the reliability levels of tests and their associated standard deviations; and of the scores (expressed numerically) that are associated with each level in Table 1. So in order to interpret scores, one needs to have sight not only of the construct that is measured, but also of how meaningfully and usefully such scores can be and have been interpreted. That kind of analogical lingual employment of the results of the language test is enriched by comparative historical data, usually expressed in numbers. If one has a score, its technical significance can be expressed with reference to a wide range of data. Part of the contestation surrounding the current theory of validity is that it makes validity wholly dependent on interpretation. It appears to be blind to the observation that there is a technical subject-object relation at play here: in order to give a subjective technical interpretation to a score, the competent technical agent doing the interpretation still needs the score, a technical object. What is more, there are questions of adequacy to be considered on both the subjective side ("How competent is the interpreter, the person making the inference?"), and on the objective side ("How effective is the score as a measurement of ability?").

In the following section, I return briefly to these lingual analogies and the technical ideas they give rise to. In this further analysis, this paper will attempt to examine several further disclosures of design, with reference to economic and ethical echoes in the design of language tests. Though these disclosures are often gathered conceptually under the umbrella of 'validation', the critical question will be asked whether they are not perhaps more meaningfully conceptualized as disclosures of responsible language test design, and whether that might not perhaps be a conceptually clearer term to use for demonstrating, through argument, the adequacy of a language test. Finally, the analysis attempts to demonstrate the robustness of the theoretical framework adopted here, and its associated methodology. It also aims to illustrate the benefit of employing a non-reductionist approach.

3. Technical meaningfulness and significance

In Schuurman's (2009:417-418) view, the technical sphere is first disclosed by the articulation of the design in blueprints and preparatory prescriptions for its development. That kind of technical expression is another lingual analogy, and is evidenced in language assessment design also as the further articulation of the construct through developing a set of specifications for the test that does justice to the construct by operationalizing it. An example of such design specifications is to be found in Table 2.

There is also another set of lingual analogies within the technical that deal with the technical significance of the test results. We can ask: Are the results meaningful? Table 1 is a first indication that they may be so. But one may, for example, wish the test results not only to be a measure of a general or specific ("academic literacy") language ability and the risks associated with low levels of ability, but also to give an indication of how even more specific weak or underdeveloped facets of that ability can be identified. Thus we find ourselves in the realm of gleaning diagnostic information from language test results. The results denote what still needs attention to achieve the desired level of academic literacy. For example, Pot's (2013) analysis of the results of the Test of Academic Literacy for Postgraduate Students (TALPS) indicated that students about to engage in postgraduate work struggle primarily with the structuring of an argument. Building on that finding, Drennan (2019) has included in her design of an Assessment of Preparedness to Present Multimodal Information (APPMI), a writing readiness test, several subtests that measure the ability to build arguments. The test measures that either directly (by means of the subtests "Organisation of text", "Organising and categorising information visually" and "Making academic arguments/building argument") or obliquely ("Interpreting graphic and visual information", "Understanding text type and communicative function", "Text comprehension", "Grammar and text relations"). As the design specifications for APPMI in Table 2 show, the test is aimed at gauging exactly the kinds and ranges of sub-abilities that may show the level of readiness to engage in academic writing of senior undergraduate students in the social sciences who intend to proceed to postgraduate study. The test is designed to inform the subsequent teaching of academic writing, and the technical design intention is expressed clearly in the specifications.

Subtest	Number of items	Weighting / mark
Organisation of text	5	5
Understanding academic vocabulary [two word format]	6	12
Interpreting graphic and visual information	8	8
Organising and categorising information visually	8	8
Understanding text type and communicative function	8	8
Text comprehension	20	20
Making academic arguments / Building arguments	8	16
Grammar and text relations	16	16
Text editing	7	7
Totals	**86**	**100**

Table 2: Specifications for an Assessment of Preparedness to Present Multimodal Information (APPMI)

A further analogical lingual trace within the technical may be found in the quest to determine whether test results have anything to say about future performance. Sebolai (2018) has investigated this predictive validity thoroughly for the undergraduate academic literacy tests being referred to in this paper, finding that TALL was a better indicator of overall academic performance especially for those first year students of a university of technology who had taken English as an additional language at school. That technically relevant information goes a long way to placing students on the appropriate academic literacy development course.

4. Technical appropriateness and fit

While the previous analysis (Weideman, 2019) has given due attention to the way that Messick's work highlights the technical adequacy of the measurement, there is another significant concept in his definition of validity as “an overall evaluative judgment of the adequacy and appropriateness of inferences drawn from test scores” (1980:1023). That concept is appropriateness. One suspects, given the contexts and the period in which these statements were being made, that ‘appropriateness’ for many would refer to the technical fit of the working of the measuring instrument, already referred to above. The concept of technical appropriateness is echoed by Bachman and Palmer (1996:21) in their remark: “Construct validity pertains to the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores.” Yet Messick is also credited, as we have noted, of drawing attention to the social and ethical dimensions of assessment, and these references take us beyond the technical adequacy or ‘fit’ of a test in the analogical physical sense.

What the conception of technical fit, in its initial, closed understanding does not yet take into account is the technical appropriateness of administering a language assessment in specific institutional contexts. The opened up, differentiated notion of technical appropriateness that one then encounters is an analogical social one: How well does this or that language test fit the institution, its needs, and the kinds of social interactions that constitute it? A test of language ability that is designed to measure the communicative competence of an aspirant salesperson will not be an appropriate measure of the ability of a test taker to handle the demands of academic discourse, for example. In testing academic literacy, there is also much debate currently about discipline specific testing. How soon should university students be sitting for a test of academic literacy that is generic, and when would the administration of a discipline or field specific test be more appropriate? While TALL and TAG are undergraduate tests, and examples of assessments of a general initial competence in academic discourse, APPMI is intended to measure the field-specific academic literacy of senior social sciences students. In the case of TALL, a generic approach may be warranted, amongst other things, by the lack of specific field differentiation among first year students, as well as by logistical considerations. APPMI, on the other hand, derives from a specific sub-institutional need, where social sciences departments wish to assess language ability that is directly relevant to the language needs of those students who intend to pursue postgraduate studies. In each case, the technical appropriateness of the test matters. Of course, once adopted, the equalization of such differential tests of academic literacy becomes another challenge to the test designer, an early indication that trade-offs may have to be made in the design and development of such tests of language ability.

On the other hand, a test like the Test of Emergent Literacy (Gruhn & Weideman, 2017), designed for measuring emergent language ability among 4-5 year olds, and the work by Steyn (2014) on a Test of Early Literacy (TEL) for 8-9 year olds, have been designed to be technically appropriate for administration to those groups, with specifications that differ markedly in level from those for measuring the ability to handle discourse at tertiary level.

The implementability of a language test, the facility of its interaction with the human test takers who are required to sit for it, is yet another social dimension of a language test. That relates closely to the economic analogies within the technical sphere, to which we turn in the next section.

5. Further unfolding: but are they disclosures of validity?

Declaring usefulness, not validity, to be the “most important consideration in designing and developing a language test” (Bachman & Palmer, 1996: 17; Bachman, 2001: 110), has, as we have noted, not deterred other commentators from interpreting this as another understanding of validity. From the methodological viewpoint taken in this paper, however, the technically stamped usefulness or utility of a language test relates in the first place to a set of economic anticipations in the technical dimension. Those analogies make it possible to conceptualize the ideas of technical usefulness, efficiency and utility. From the prior discussion it is also clear that for a language test to be useful, it would first have to be appropriate and implementable (its social side), to generate meaningful, interpretable results (the lingual analogies), and to be based on a sound, theoretically defensible construct (a reference to the analytical). The idea of the technical usefulness of a language test relates, in turn, specifically to analogical links with the economic sphere, allowing those who design and use language tests to raise questions about the careful and frugal use of resources for assessment, about the technical means to achieve designed ends, about logistic facility in administering a test, and about efficiency in obtaining its results. Whether these are still disclosures of validity, however, now begins to look doubtful.

Similarly, when Kunnan (2000:10) declares fairness to be “a critical central component” of language testing, in fact one that has ‘primacy’, we encounter the disclosure of language test design by anticipatory ethical issues. Once more, Messick’s idea of “consequential validity” or test impact pioneered these concerns, but at the same time still links the ethical closely to the concept of validity. Conceptually speaking, however, these analogical ethical ideas are not validity issues. Rather, they allow us first to conceptualize technical fairness in general, and then to ask specifically whether the test treats those whose language ability is being measured fairly, without bias, with due care and compassion, and with love and respect.

Testing whether the items that comprise the language test show Differential Item Functioning (DIF) is one quantifiable measure of fairness. DIF analyses help the test designer to identify those items that discriminate against candidates simply because they belong to a certain group, gender or class. DIF statistics are measures that identify what may be technically irrelevant factors that influence the measurement of language ability. The tests of academic literacy levels being used as illustration here have been found to be largely free from such discriminatory items (Van der Slik & Weideman, 2010).

Another measure of fairness can be found in the determination of how many candidates sitting for a test may have been misclassified as a result of test inconsistency. Taking the administration of TALL 2011 as an example (Table 3), we can see that, in four scenarios (Alpha or Greatest Lower Bound based; same test or parallel test – CITO 2005:17-18), of the more than 5000 test takers, a maximum of 204 candidates might in the worst outcome have been misclassified. That calculation enables the test designers and administrators to offer a second chance test to at least 102 candidates below the cut-off score, on the basis of there being an even chance of being misclassified above or below that decision point.

Misclassifications			
Alpha based		**GLB based**	
- Rxx’ case: Percentage	3.8%	Percentage	3.3%
Number	204	Number	178
- Rxt case: Percentage	2.7%	Percentage	2.4%
Number	148	Number	128

Table 3: Potential misclassifications in administration of TALL (2011)

When we consider the “consequential validity”, the impact of a test, we ask: How fairly does the test measure? How compassionately does it treat those taking it? Was it designed with care and concern for them? In all of these, the ethical anticipations of language test design and use, as fair, caring and compassionate, are made possible conceptually. For the purpose of this discussion, the question is, however, whether it is not overstretching the notions of validation and validity to subsume all analogically lingual, social, economic and ethical ideas in language test design under the single conceptual umbrella of ‘validity’.

6. Pivotal technical concepts and ideas generate design principles

In the preceding, a number of pivotal technically stamped concepts and ideas relating to language test design have been discussed. The discussion, however, is by no means exhaustive: not every single technical primitive has been identified and articulated. The discussion has illustrated how such primitives are conceptualized: as analogical moments in the technical dimension of design, both in retrocipatory direction, as constitutive technical concepts, and in anticipatory direction, as regulative technical ideas. The analogies arise from the echoes of the other dimensions of reality within the technical sphere that has, as the leading function of language tests, theoretically been isolated for scrutiny. Besides the constitutive concepts of technical reliability, validity, and theoretical defensibility ("construct validity"), that relate respectively to the echoes of the kinematic, physical and logical aspects of experience, the analysis has also considered the technical interpretability (a lingual analogy), implementability (a trace of the social), usefulness (an analogical economic link) and fairness (the ethical anticipation). But there are indeed many more, which for limitations of space cannot be dealt with here, but will briefly be mentioned below, as they have been in Figure 1.

The further implication of such conceptualization, which also has to be treated here only cursorily, is that from each pivotal technical concept or idea applied linguists may derive a principle for their designs of language interventions. Thus we can say that our tests must be reliable, effective, theoretically defensible, meaningful, appropriate, efficient, and fair. But we might add to those principles the condition for language test design to be a unity within multiplicity of components, with reference to the numerical analogy in the technical. Taking the factor analysis of TALL 2006 as an example (Figure 2), we note that, while the test is arguably a unity (all the items test the first factor), there are subgroupings: items 1-5, all relating to the first subtest, lie slightly out, as well as the items in the last subtest, while the other subtests cluster more tightly around the zero line. It is for the test designer to argue that these groupings (1-5; 6-47; 48-60) are a technical unity within a multiplicity of components.

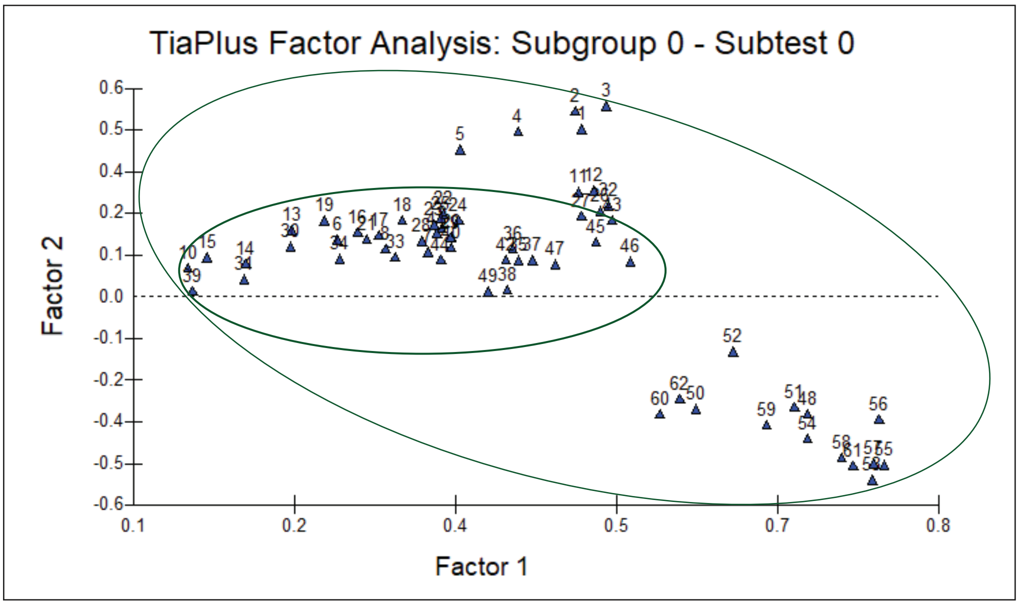

Figure 2: Factor analysis of TALL 2006

Similarly, we may argue that a language test has to have a specified range (a spatial echo), must be internally differentiated (a reference to the organic foundation of the technical design) in respect of its subtests (see Table 2), must have intuitive appeal (the sensitive

retrocipation, usually referred to as the "face validity" of the test); or that the designers must be accountable for their technical plans (the juridical), strive to build a reputable assessment (the conceptual trace of the certitudinal), and so on. Each analogical moment yields a design principle that must be taken into consideration, interpreted and given flesh in the design of language test. However, they remain principles that must be responded to in ways that in practice may differ contextually, yet must be accounted for. To call that 'validation' is to narrow the conception, and to confuse rather than clarify.
So the final question here is: If we can discover principles for designing language tests responsibly, should we not rather consider replacing the term 'validation' with "responsible test design"? We stretch the concept of validity almost to breaking point, and make the further concepts that are introduced meaningless if we either subsume validity under them, or confusingly equate them with validity.

This paper has attempted to demonstratehow the employment of a particular methodology may help us towards conceptual clarification that we badly need in language testing. The methodology itself is robust and attractive, in that it strives to be non-reductionist: no dimension of experience is absolute, or can be promoted to one that overbears all others, since all aspects, though uniquely discernible as different modalities, are interdependent, echoing in their interdependence the others. Such a methodology should be at the basis of a serious attempt to conceptualize and fashion a theory of applied linguistics.

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, 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.

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

Drennan, L. 2019. Assessing readiness to write: the design of an Assessment of Preparedness to Present Multimodal Information (APPMI). (Submitted as a chapter in Du Plessis, T., Read, J. & Weideman A. *eds*. Transition and transformation: Assessing academic literacy in a multilingual society. Forthcoming from *Multilingual Matters*.)

Gruhn, C.M.S & Weideman, A. 2017 The initial validation of a Test of Emergent Literacy (TEL). *Per Linguam,* 33(1):25-53. https://doi.org/10.5785/33-1-698.

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.) https://doi.org/10.1002/9781118411360.wbcla144.

Kunnan, A.J. *ed*. 2000. Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida (Studies in Language Testing; 9.) Cambridge: Cambridge University Press.

Messick S. 1980. Test validity and the ethics of assessment. *American Psychologist*, 35(11):1012-1027. https://doi.org/10.1037//0003-066x.35.11.1012.

Pot, A. 2013. Diagnosing academic language ability: an analysis of TALPS. Groningen: Rijksuniversiteit Groningen (M.A. dissertation.)

Schuurman, E. 2009. Technology and the future: a philosophica1 challenge. Translated by H.D. Morton. Grand Rapids: Paideia Press. [Originally published in 1972 as: Techniek en toekomst: confrontatie met wijsgerige beschouwingen. Assen: Van Gorcum.]

Sebolai, K. 2018. The differential predictive validity of a test of academic literacy for students from different English language school backgrounds. *Southern African Linguistics and Applied Language Studies,* 1-12. https://doi.org/10.2989/16073614.2018.1480899.

Steyn, S. 2014. The design and refinement of a test of early academic literacy. Groningen: Rijksuniversiteit Groningen. (M.A. dissertation.)

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

Van der Slik, F. & Weideman, A. 2010. Examining bias in a test of academic literacy: Does the *Test of Academic Literacy Levels* (*TALL*) treat students from English and African language backgrounds differently? *Journal for Language Teaching,* 44(2):106-118. https://doi.org/10.4314/jlt.v44i2.71793.

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. 2019. Degrees of adequacy: the disclosure of levels of validity in language assessment. Submitted to *Koers.* https://doi.org/10.19108/KOERS.84.1.2451.