Albert Weideman

http://dx.doi.org/10.4314/jlt.v45i2.6

Academic literacy tests: design, development, piloting and refinement

A B S T R A C T Though there are many conditions for drafting language tests responsibly, this contribution focuses first on how to operationalise a set of three critically important design principles for such tests. For the last thirty years or so, developers of language tests have agreed that the most important design principle emanates from our ability to give a theoretical justification for what it is that we are measuring. Without this, we eventually have very little ground for a responsible interpretation of test results, which is a second, though not secondary, principle for language test design. There is a third principle involved, which is that the measuring instrument must be consistent and stable. The paper investigates how a blueprint for an academic literacy test may be conceptualised, how that could be operationalised, and demonstrates how pilot tests are analysed with a view to refining them. Finally, that leads to a consideration of how to arrive at a final draft test, and how valid and appropriate interpretations of its results may be made. Since the three conditions for language tests focussed on here are not the only design principles for such applied linguistic instruments, the discussion is placed in a broader philosophical framework for designing language tests that also includes a consideration of some of the remaining design principles for language testing.

Introduction

The scholarly work of Johann van der Walt that we are honouring in this journal has made its most significant mark in that sub-field of applied linguistics that we know as language testing. Not only has he published widely on assessing language ability, but his influence also extends further, through the analyses done by his postgraduate students. In addition, he has been involved in what might well be described as the coalface of language assessment, since over many years he has made a valuable contribution to the school exit examination for English additional language learners. His work has most influenced my own in a substantial article (Van der Walt & Steyn 2007), especially in assisting to conceptualise more clearly the notion of the technical validity of a language test (Weideman 2009a). That important condition – to some the most

important and overriding design principle for language tests, of what Messick (1980: 1019) calls that "complex and holistic concept …[of] … test validity" – is, however, not the only design criterion for language tests. If we are to progress towards conceptual clarity in language testing, we may have to go beyond the alluring notion of a unifying single concept such as validity.

In the present discussion I intend to take the notion of what I have been calling a responsible framework for the design, development and refinement of applied linguistic instruments further, in discussing, first, three critically important design principles for language tests, and, next, articulating how these three and other conventionally important design principles can fit into a conceptual framework that does justice to all of their varying emphases. In calling for the articulation of a responsible framework of conditions for an applied linguistic instrument such as a test as regards **design** and as regards **development**, one is actually saying two things: first, that one needs a conceptual key to understand what it is that one is designing (and, by extension, measuring), and, second, to know which flexible process or perhaps less flexible, set procedures one is going to use in developing an instrument in terms of a coherent set of principles. These two, design and development, must therefore be simultaneously considered for a deliberate, thoughtful instrument to be constructed.

A first look at a framework of design principles for language tests

The first three design principles that will be discussed here concern (1) the theoretical defensibility of the definition of the ability being measured, or the construct on which the test is based, better known as the construct validity of a test (cf. Messick 1980, 1988; and the *American Standards for educational and psychological testing* [American Educational Research Association 1999: 9]); (2) the consistency and stability of the test as a measuring instrument; and (3) whether and how one might give appropriate and adequate interpretations of the test results.

The framework into which they fit is in the first instance a technical, design framework. Tests are technically qualified instruments; they are stamped or characterised by the technical ability of professional applied linguists to give shape to, to form, to devise, to plan and to develop an instrument that can be used to assess language ability. Viewed from the angle that they are characterised by their technical design, tests are similar to other types of applied linguistic artefacts, such as language curricula and courses, and language policies and language management plans. In these artefacts, we see an interplay in the design and development process between technical norms and technical facts. In that interplay, we encounter two levels of applied linguistic designs: a prior, conditioning (or norming) artefact, and a factual, or end-user format of the design, which is determined by, or at least should be brought into alignment with the principles set out in the former. The following table (Weideman 2011: 14) summarises these interactions between normative and factual designs in various applied linguistic subfields:

Table 1: Levels of applied linguistic artefacts

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

In the case of language tests, this distinction is useful because it articulates that tests, as end-user formats of the design, are subject to specifications which in their turn derive from a theoretically justifiable idea or construct of what is being tested. That condition of construct or theoretical defensibility, in turn, relates first to the technical design framework into which it can be cast, as will be illustrated below.

For Messick, for example, our test design work is dominated by the notion that language testing needs an overall, 'unitary' framework that it finds in the idea of validity (Messick 1988: 35, 40f.; 1981: 9; 1989: 19). That conception has for over 20 years been the dominant, orthodox one (see also Van Dyk 2010: 178ff.), yet it is today certainly not the only one that holds sway. Bachman and Palmer's (1996: 18) idea of promoting the criterion of technical utility, summarised below in Figure 1, presented it with an early challenge (and see also Borsboom, Mellenbergh & Van Heerden 2004):

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

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

The question of whether one should at all strive for such a unitary framework is a difficult one to answer. Elsewhere (Weideman 2009a) I have argued, once again with an appeal to the nature of applied linguistics as a discipline of design, that, if one has to have a unitary conceptualisation, it would be stamped by the guiding technical design function of applied linguistic artefacts, as articulated in Figure 2 below (Weideman 2009a: 244). In this representation, the technical mode of experience leads and guides the analytical analogy within its sphere of influence:

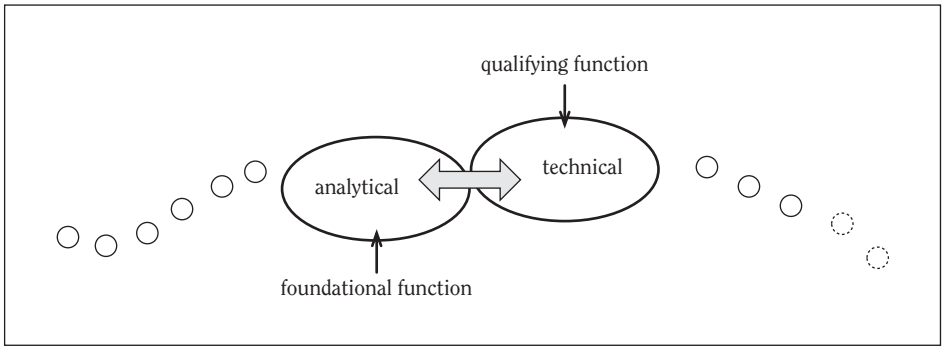

Figure 2: Terminal functions of an applied linguistic design

Phrased differently, in the case of applied linguistic designs, the analytical dimension of experience, in its interplay with the leading technical design function, makes it possible for us to provide a theoretical foundation or justification for the instrument we are designing. This insight is critical, amongst other things, for developing a notion of validity that takes us beyond the conceptual problems that the current orthodoxy has brought us up against (Weideman 2009a). In the following as well as in the final sections, we return to the framework introduced here.

A theoretically defensible construct for measuring academic literacy

As an illustration I shall use here material from that kind of language test design about which most has been written in South Africa: the design and development of the academic literacy tests known as the *Test of Academic Literacy Levels (TALL)* or, in their Afrikaans versions, as the *Toets van Akademiese Geletterdheidsvlakke (TAG)* (cf. for example Butler 2009; Geldenhuys 2007; Le 2011; Marais & Van Dyk 2010; Van der Slik 2008; Van der Slik & Weideman 2005, 2007, 2008, 2009, 2010; Van der Walt & Steyn 2007; Van Dyk, 2010; Van Dyk & Weideman 2004a, 2004b; Weideman 2003, 2006a, 2006b, 2009a; Weideman & Van der Slik 2008; Weideman & Van Rensburg 2002). Johann van der Walt not only enthusiastically participated in the design of these tests that are owned by the Inter-Institutional Centre for Language Development and Assessment (ICELDA), of which he is also a board member, but has done some analyses of test use as well.

For TALL and TAG, the desirable and theoretically defensible construct had to be one that conceived of academic literacy as the ability to handle academic discourse at university level. Simultaneously, academic discourse was conceived of as a distinctly and materially different type of discourse (Weideman 2009b), that is embedded in human interaction within an institutional context. Rather than viewing language ability as limited to mastery of sound, form and meaning, Bachman and Palmer (1996: 61ff.) have pointed out that tests should be based on an interactional perspective of language ability, which emphasises the ability to negotiate meaning within specific contexts. Similarly, Blanton (1994: 221), even while she emphasises that the nature of academic discourse is one in which written language is valued as the major currency, notes that academically literate behaviour involves interacting with texts:

> Whatever else we do with L2 students to prepare them for the academic mainstream, we must foster the behaviors of 'talking' to texts, talking and writing about them, linking them to other texts, connecting them to their readers' own lives and experience, and then using their experience to illuminate the text and the text to illuminate their experience (1994: 228).

It is not difficult to see where the view of language articulated here by Blanton is different from that of 50 years ago. It emphasises that language is interaction rather than structure; it is an open, rather than a closed view of language. Blanton's useful definitions of academic discourse, however, are not easy to operationalise. Our experience in searching for a construct that could be converted into a set of technical specifications for what needs to be tested, led the designers of TALL and TAG eventually to a definition of academic language ability which states that students who are academically literate must be able to:

- understand a range of academic vocabulary in context;
- interpret and use metaphor and idiom, and perceive connotation, word play and ambiguity;
- understand relations between different parts of a text, be aware of the logical development of (an academic) text, via introductions to conclusions, and know how to use language that serves to make the different parts of a text hang together;
- interpret different kinds of text type (genre), and show sensitivity for the meaning that they convey, and the audience that they are aimed at;
- interpret, use and produce information presented in graphic or visual format;

- make distinctions between essential and non-essential information, fact and opinion, propositions and arguments; distinguish between cause and effect, classify, categorise and handle data that make comparisons;
- see sequence and order, do simple numerical estimations and computations that are relevant to academic information, that allow comparisons to be made, and can be applied for the purposes of an argument;
- know what counts as evidence for an argument, extrapolate from information by making inferences, and apply the information or its implications to other cases than the one at hand;
- understand the communicative function of various ways of expression in academic language (such as defining, providing examples, arguing); and
- make meaning (e.g. of an academic text) beyond the level of the sentence.

(Weideman 2007: xi-xii)

This construct of academic literacy can be operationalised in the following set of specifications (adapted from Van Dyk & Weideman 2004b: 18-19):

Table 2: Test specifications and task types

Specification	Task type(s)
Vocabulary comprehension	Vocabulary knowledge tests Cloze procedure
Understanding metaphor & idiom	Text comprehension passages
Textuality (cohesion and grammar)	Scrambled text Cloze procedure Text comprehension passages
Understanding text type (genre)	Register and text type tasks Interpreting and understanding visual & graphic information Scrambled text Cloze procedure Text comprehension passages
Understanding visual & graphic information	Interpreting and understanding visual & graphic information (potentially:) Text comprehension passages
Distinguishing essential/ non-essential	Text comprehension passages Interpreting and understanding visual & graphic information
Numerical computation	Interpreting and understanding visual & graphic information Text comprehension passages
Extrapolation and application	Text comprehension passages (Interpreting and understanding visual & graphic information)
Communicative function	Text comprehension passages (possibly also:) Cloze procedure, Scrambled text
Making meaning beyond the sentence	Text comprehension passages Register and text type tasks Scrambled text Interpreting and understanding visual & graphic information

The alignment of the construct with the task types eventually selected, after piloting, for TALL and TAG, is of course essential, as is the continued critical scrutiny and examination of the construct itself. The theoretical defensibility of the definition of academic literacy operationalised first in specifications for task and item types, and next in the actual test itself, depends on its being in accord with recent and relevant insight into the kind of language being assessed. In sum: the construct of academic literacy used in these tests is theoretically defensible because it gives the test designer and test users a current, rational explanation of what it is that gets measured.

A second principle: consistency and stability

The technical consistency or reliability of the tests of academic literacy that are used as examples here is conventionally calculated by utilising a reliability index such as Cronbach's alpha or, for more heterogeneous constructs, the lesser known index known as Greatest Lower Bound (cf. Van der Slik & Weideman 2005). There are 10 different versions of TALL and eight versions of TAG; both tests, in every single one of their various administrations over the past seven years, far exceed the benchmark of a Cronbach's alpha of 0.7. TALL, in fact, has maintained an average of well above 0.9 over any five administrations in any one or more of the five institutional contexts where it has been administered in the past seven years.

A reliability index gives one a measure of how consistently a test measures. The overall test reliability of course derives from the way that each of the subtests or task types, and the individual items that make up these subtests, perform. For example, for TALL and TAG, the conditions for a productive item include, first, its alignment with the construct, which is examined not only during design, but scrutinised again after being piloted. Second, these conditions include parameters for item performance both in terms of facility (the percentage correct answers elicited when being tested out) and discrimination. The latter are calculated by using a programme such as Iteman. Though one is aiming for 50% ease, an acceptable facility value for an item in TALL and TAG would lie anywhere between 0.2 and 0.8, meaning that between 20% and 80% of test-takers usually get the right answer. In effect, the designers of TALL and TAG would accept facility values from 0.15 up to 0.84. As to discrimination value, the ability of the specific item to discriminate between the top 25% and the bottom 25% scorers on the total test score, the test designers are looking for items that normally would yield a value of 0.3 (or at least 0.25) and more. The exceptions in this case are for high facility values where an item is an introductory one. The rationale is that an easy item to introduce a challenging subtest may give some positive motivation.

A third measure of consistency relates to the homogeneity of the test as indicated in a factor analysis. The more heterogeneous a test is, the more outlying items there are in this analysis. The following example (Figure 3), a graphic representation of TALL 2008 as administered at the University of Pretoria, shows that in the case of two subtests (items 1-5; 51-67), the items lie a little further away from the zero line associated with the measurement of a single factor, or homogeneous construct.

If one looks at the graph as a whole, however, the items still display a reasonable measure of association with one another. As Van der Slik and Weideman (2005) point out, as rich a construct

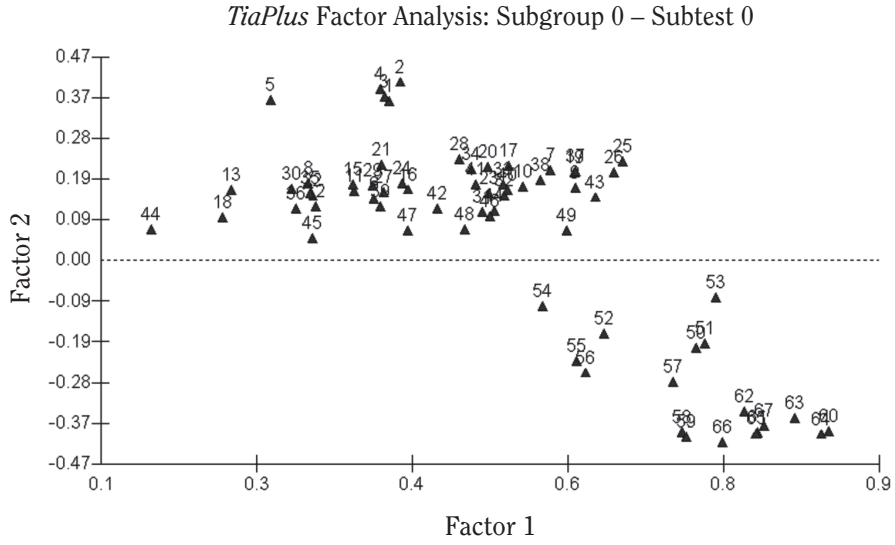

Figure 3: Factor analysis of TALL 2008 (UP)

of academic literacy as the one employed in these tests is likely to yield a more heterogeneous picture than was originally expected from analyses of the psychometric qualities of earlier tests, those of 50 years ago. Certainly, the two outlying sections can be claimed to be strongly associated with notions of academic literacy that are current today. This kind of analysis can be done with free software: Figure 3 above was done by means of TiaPlus (CITO, 2005).

Once these various analyses of technical stability have been carried out, test designers have an empirical yardstick, or in fact several, to evaluate which items (and sometimes subtests) should be eliminated, replaced, or modified. Depending on whether sufficient items remain after such sifting, and, most important, their representativity of the construct, a responsible decision can be taken on whether to proceed with further piloting and refinement, or whether the test is ready for use. Should the items that remain after culling not be representative of the construct, they either have to be augmented, or the weightings of the remaining ones adjusted, in order to cover the construct and specifications adequately. This refining process is never easy, quick or straightforward. Responsible test design takes time and patient analysis.

A third principle: appropriate interpretation of test results

The empirical results obtained with the kinds of analyses referred to in the previous section are supplemented by other analytical and statistical information, and point to how such quantitative information can be employed to arrive at responsible interpretations of test results. For example, the developers of TALL and TAG initially used historical data, derived from the employment of another, earlier test of academic literacy (cf. Weideman & Van Dyk 2004a), to set the cut-off point to determine who qualifies for the compulsory academic literacy intervention in which students should be placed who were identified by the test as being at risk as regards their level of academic literacy. The data they used were based on the fact that the earlier test was what is known as a norm referenced test, which is a kind of test in which results are calibrated against another, external benchmark, in this case, Grade 10 learners. The use of

this test had indicated that over a number of years those measuring at a level of language ability associated with that of Grade 10 learners or lower grades had stayed consistent at between 27% and 33%. The following table, from Van Dyk and Weideman (2004a: 4), captures this for the administration of the previous test at one tertiary institution:

Table 3: Summary of test results since 2000

2000		2001		2002		2003	
N = 4661		N = 5215		N = 5788		N = 6472	
≥Gr.11	≤Gr.10	≥Gr.11	≤Gr.10	≥Gr.11	≤Gr.10	≥Gr.11	≤Gr.10
N = 3356 (72%)	N = 1305 (28%)	N = 3495 (67%)	N = 1720 (33%)	N = 4212 (73%)	N = 1576 (27%)	N = 4615 (71%)	N = 1857 (29%)

Setting the cut-off point when using the scores of the new tests, TALL and TAG, was therefore based, among other things, on experience already gained, developed further, and meticulously recorded in subsequent years. The development of ever more sophisticated, additional analytical rationales for the responsible determination of cut-off points are described in detail in Van Dyk (2010: 166f., 268 f.), but again it should be noted that it took some time, and looking at the scores from many different possible points of view, before these decisions gained in stability and credibility.

What is most important to note, however, is that the interpretation of the test scores had to take into account that (a) a stigma could in the case of some students attach to the results of the tests, since, after all, they indicate potentially low levels of academic literacy in some; and (b) the test results, though generated in a highly reliable fashion, as was discussed above, were themselves not perfect. No test is 100% reliable. First, the test developers attempted to solve this problem by making the results available not only with reference to a cut-off point, which would have resulted in a simple pass/fail, but rather in terms of several risk bands, as in Table 4 below:

Table 4: Levels of risk associated with scores on TALL and TAG

Risk level	Interpretation
1	Very high risk
2	High risk/clear risk
3	Borderline (moderate risk)
4	Less risk
5	Little to no risk

The scores associated with the different bands were still derived from the cut-off point (that determined the highest score for level 2, and the lowest for level 4), but at least the publication of the results mitigated the impact somewhat, especially at level 2, or, if at level 1, indicated the seriousness of the risk. Second, the fact that the test was not entirely reliable was compensated for in the interpretation of test results by giving a second chance test to those who were deemed to be borderline cases (level 3).

The empirical basis for determining who was a borderline case and therefore eligible for a re-test was found in a calculation, again done through TiaPlus, which gives one an option of considering four different scenarios of potentially misclassified test-takers. The four scenarios derive from two measures of reliability, Cronbach's alpha and Greatest Lower Bound or GLB (cf. Jackson & Agunwamba 1977), as well as from a consideration of comparing scores on the test with a hypothetical parallel test, or from a comparison of observed scores with 'true' scores (for a discussion of these, cf. Van Dyk 2010: chapter 5). In this manner, by calculating the extent to which the reliability of the test undermines the possibility of obtaining an accurate score from this measuring instrument, the number of potential misclassifications can be calculated, as in Table 5 below (adapted from Weideman & Van der Slik 2008), for students from each of the two participating higher education institutions:

Table 5: Potential misclassifications on the English version of the academic literacy test (Percentage of this test population). In italics the corresponding intervals (in terms of standard deviations) around the cut-off points

	Institution 1	**Institution 2**
Alpha based:		
Correlation between test and hypothetical parallel test	432 (13.0%) *63-74 (.31)*	246 (14.2%) *63-74 (.41)*
Correlation between observed and 'true' scores	308 (9.3%) *65-72 (.21)*	176 (10.2%) *66-72 (.27)*
GLB based:		
Correlation between test and hypothetical parallel test	360 (10.9%) *64-73 (.26)*	213 (12.3%) *66-72 (.27)*
Correlation between observed and 'true' scores	256 (7.7%) *66-71 (.15)*	152 (8.8%) **67-71 (.21)**

The numbers in Table 5 show that in the case of institution 1, as few as 7.7% of students who took the test, or as many as 13.0% of them, could potentially have been misclassified. What is more, these potential misclassifications lie for test takers in this institution between 0.15 or 0.31 standard deviations around the cut-off point. This is an important yardstick, and can be used in subsequent calculations to assist in determining who qualifies for a second chance test.

What this means is that the test administrators can be advised to give a second chance test to at least half of the potentially misclassified students. The assumption behind this is that the test may have evenly placed students at a disadvantage, by giving them fewer marks than they might have scored, or may have advantaged them by giving them more. In concrete terms, for institution 1, the advice would be to give another chance to between half of 256, i.e. 128 students, which is the lowest number of those potentially affected by a misclassification, and half of 432, i.e. 216 students, which is the highest number potentially affected. To err on the side of caution, just fewer than 220 students might have been given a second opportunity, and they would be the first 220 lying below the cut-off point.

Disclosure of the meaning of design

The examples given at the end of the previous section provide a first illustration of how, in a responsible test design and administration, care is exercised not to place the humans who are affected by tests at a disadvantage. Responsible designs for measuring instruments such as language tests demonstrate love and care for others, and do not place them under duress. If a test has been responsibly designed, its implementation will be a disclosure of the meaning of design. In disclosing or opening up the meaning of design in this manner, the leading technical function of the design anticipates the social dimension of experience (cf. McNamara & Roever, 2006).

In the framework being employed here, the technical nature of a test needs to be related first, for example, to what can be called constitutive concepts such as technical consistency and technical validity, that were both discussed above as conventional criteria for test design. Weideman (2009a: 246) sets it out graphically:

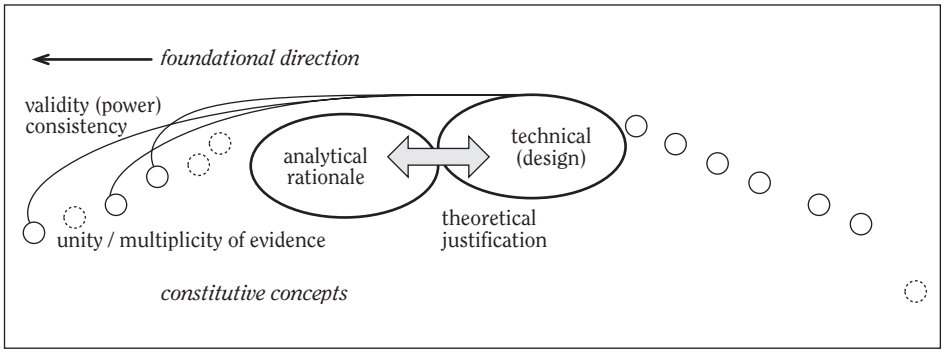

Figure 4: Foundational concepts of applied linguistic designs

The technical design function of a language test therefore crucially depends on such foundational principles as systematicity (an echo of the analogical numerical notion of a technical unity within a multiplicity of sets of evidence that support the test design), as well as on technical reliability or consistency, and validity (as an instrument, or technical object, the test has the technical power to yield technical effects, or scores). In the technical rationale or theoretical defensibility of a test, a reflection of the relation between the leading technical function of the test and its analytical basis, we find, in turn, the conceptual basis for the widely held notion of construct validity. It is as if, in this concept, we look at validity, an originally physical concept, from the vantage point of the analytical dimension of reality, and this is the basis for much of the misunderstanding between Messick and his detractors (e.g. Borsboom, Mellenbergh & Van Heerden 2004). They look at test validity from a purely analogical physical relation (between objective technical cause and effect), while Messick and those who follow in his paradigmatic footsteps take a view that provides not only for theoretical defence of a construct, but its continuing subjective validation through amassing evidence for the quality of the technical instrument that has been developed. The foundational side of the test design framework that has formed the background of this discussion was the topic of a recent thesis (Van Dyk 2010), and is set out there in more detail.

The leading design function of a test not only reaches out, as it were, in a foundational direction to the conventional building blocks of test design, such as consistency and validity. By anticipating the lingual dimension of experience, in the articulation of the blueprint of the test, and in the responsible interpretation of test scores – the third important test design principle discussed above – its leading technical design function takes the first step towards the disclosure of the meaning of design. In addition, the economic analogy within the technical yields the principle of technical utility or frugality, as clarified by Bachman and Palmer in their influential (1996) work on language testing, that was referred to above. In anticipating the aesthetic dimension of reality, the technical connects to the harmonisation of the designed instrument with educational, social and societal issues, such as aligning the test with teaching and learning (to overcome or promote what is known as 'washback'), test accessibility and transparency. Finally, in the test design principles of technical accountability and fairness, we see the connection of the technical with the juridical and ethical modes, in what may be called the regulative ideas that spring from those connections. This is set out in Figure 5 below (Weideman 2009a: 248):

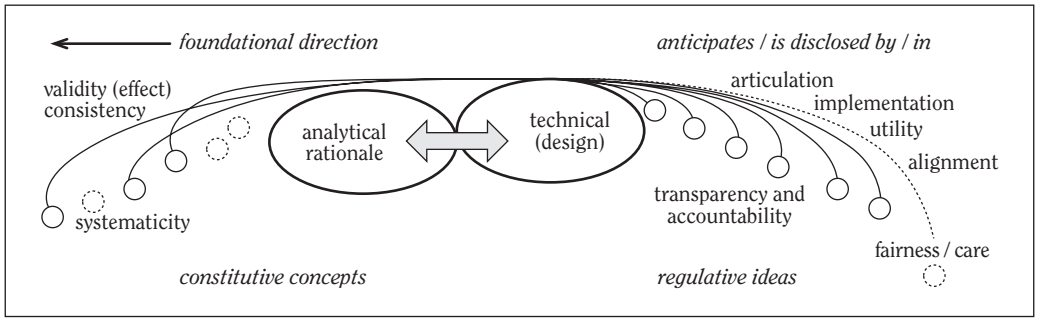

Figure 5: Constitutive concepts and regulative ideas in applied linguistic designs

It is in order to be able to defend the social impact of a test, to become accountable in the public domain for the consequences of one's test design, that test developers use various analyses to locate and identify possible test biases. Such biases may result in certain language or racial groups being privileged over others (Van der Slik & Weideman 2010), or members of a certain gender being advantaged as compared with members of another (Van der Slik 2008). These regulative notions are currently the subject of a doctoral thesis (Rambiritch, in preparation), in which the regulative ideas of technical transparency, accessibility, accountability and fairness are conceptually examined in a systematic way that has not yet been accomplished. We hope soon, therefore, to understand more of this complex topic.

Conclusion

This contribution has sought to clarify at least three critically important design principles for language tests. There are a number of implications that flow from this particular, and admittedly partial, clarification. In doing so, it has attempted to illustrate that our current obsession with trying to subsume everything under the notion of validity may be ill-founded. Of course a test should be validated, just as its consistency must be scrutinised, and its construct subjected to critical examination. But if we accept that it is an instrument designed

to measure language ability, then its technical nature must compel us to investigate a much broader range of technically stamped design principles than the conventional. Given what we know today about principles for responsible test design, we must ensure that tests also conform to regulative conditions for the technical design process. Tests derive their integrity from their theoretical defensibility, as well as from their social transparency or public accountability. They must not only be consistent, but their scores must also be responsibly interpreted. No matter how much evidence is amassed to demonstrate their validity, they must also be administered in such a way that their care and concern for those whom they measure are evident. In short, an approach to the design and development of language tests that takes as its point of departure a comprehensive set of design principles common to a number of applied linguistic artefacts is more likely to provide us with a responsible, defensible instrument.

ACKNOWLEDGEMENTS

I would like to thank Colleen du Plessis, my research assistant, for helping to put the final touches to this contribution, as well as the guest editor, Bertus van Rooy, and two anonymous reviewers for helpful suggestions for improving it.

REFERENCES

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

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

Blanton, L.L. 1994. Discourse, artefacts and the Ozarks: understanding academic literacy. *Journal of second language writing* 3(1): 1-16. Reprinted (as Chapter 17: 219-235) in V. Zamel & Spack, R. (Eds.), 1998. *Negotiating academic literacies: teaching and learning across languages and cultures.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

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

Butler, G. 2009. The design of a postgraduate test of academic literacy: accommodating student and supervisor expectations. *Southern African linguistics and applied language studies* Special issue: Assessing and developing academic literacy (ed.: J. Geldenhuys) 27(3): 291-300.

CITO. 2005. *TiaPlus, Classical Test and Item Analysis*©. Arnhem: Cito M. & R. Department.

Geldenhuys, J. 2007. Test efficiency and utility: Longer and shorter tests. *Ensovoort* 11 (2): 71-82

Jackson, P.W. & Agunwamba, C.C.1977. Lower bounds for the reliability of the total score on a test composed of nonhomogeneous items: I. Algebraic lower bounds. *Psychometrica* 42: 567-578.

Le, Phuong Loan. 2011. *Assessing academic literacy of first year Vietnamese students: How appropriate is the TALL?* Unpublished master's dissertation. Groningen: Rijksuniversiteit Groningen.

Marais, F. & Van Dyk, T. 2010. Put listening to the test: An aid to decision making in language placement. *Per linguam* 26(2): 34-51.

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 measuremen*t. Third edition. New York: American Council on Education/Collier Macmillan, pp 13-103.

Rambiritch, A. In preparation. *Accessibility, transparency and accountability as regulative conditions for a post-graduate test of academic literacy.* Unpublished doctoral thesis. Bloemfontein: University of the Free State.

Van der Slik, F. 2008. Gender bias and gender differences in two tests of academic literacy. *Southern African linguistics and applied language studies* Special issue: Assessing and developing academic literacy (ed.: J. Geldenhuys) 27(3): 277-290.

Van der Slik, F. & Weideman, A. 2005. The refinement of a test of academic literacy. *Per linguam* 21(1):23-35.

Van der Slik, F. & Weideman, A. 2007. Testing academic literacy over time: Is the academic literacy of first year students deteriorating? *Ensovoort* 11(2): 126-137.

Van der Slik, F. & Weideman, A. 2008. Measures of improvement in academic literacy. *Southern African linguistics and applied language studies* 26(3): 363-378.

Van der Slik, F. & Weideman, A. 2009. Revisiting test stability: further evidence relating to the measurement of difference in performance on a test of academic literacy. *Southern African linguistics and applied language studies* Special issue: Assessing and developing academic literacy (ed.: J. Geldenhuys) 27(3): 253-263.

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? *SAALT Journal for language teaching* 44(2): 106-118.

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

Van Dyk, T. 2010. *Konstitutiewe voorwaardes vir die ontwerp en ontwikkeling van 'n toets vir akademiese geletterdheid.* Unpublished Ph.D. thesis. Bloemfontein: University of the Free State.

Van Dyk, T. & Weideman, A. 2004a. Switching constructs: on the selection of an appropriate blueprint for academic literacy assessment. *SAALT Journal for language teaching* 38 (1): 1-13.

Van Dyk, T. & Weideman, A. 2004b. Finding the right measure: from blueprint to specification to item type. *SAALT Journal for language teaching* 38 (1): 15-24.

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

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

Weideman, A. 2006b. Assessing academic literacy in a task-based approach. *Language matters* 37(1): 81-101.

Weideman, A. 2007. *Academic literacy: Prepare to learn.* 2nd edition. Pretoria: Van Schaik.

Weideman, A. 2009a. Constitutive and regulative conditions for the assessment of academic literacy. *Southern African linguistics and applied language studies* Special issue: Assessing and developing academic literacy (ed.: J. Geldenhuys) 27(3): 235-251.

Weideman, A. 2009b. *Beyond expression: A systematic study of the foundations of linguistics.* Grand Rapids, Michigan: Paideia Press.

Weideman, A. 2011. Straddling three disciplines: Foundational questions for a language department. 30th DF Malherbe Memorial Lecture. *Acta varia*. Bloemfontein: University of the Free State.

Weideman, A. & Van der Slik, F. 2008. The stability of test design: Measuring difference in performance across several administrations of a test of academic literacy. *Acta academica* 40(1): 161-182.

Weideman, A. & Van Rensburg, C. 2002. Language proficiency: current strategies, future remedies. *SAALT Journal for language teaching* 36 (1 & 2): 152-164.

ABOUT THE AUTHOR

Albert Weideman
Department of English
University of the Free State
Email: albert.weideman@ufs.ac.za