Per Linguam 2020 36(2):57-75
http://dx.doi.org/10.5785/36-2-970

Journal Home Page:
http://perlinguam.journals.ac.za

Per Linguam

A Journal for Language Learning
Tydskrif vir Taalaanleer

COMPLEMENTARY EVIDENCE IN THE EARLY-STAGE VALIDATION OF LANGUAGE TESTS: CLASSICAL TEST THEORY AND RASCH ANALYSES

Albert Weideman
University of the Free State

ABSTRACT

Test validation may more aptly be conceived of as the process of designing language tests responsibly. While a good test gains in reputation as it is administered over time, the early stages of its validation are perhaps the most critical. There is now general agreement that the validation process should be reported in the form of an argument that brings together multiple sets of evidence to justify the design and implementation of the measurement instrument, the language test. The format of such integration is, however, still contestable ground. Referring to an example of language test design and development, this paper seeks to demonstrate how a framework for responsible test design may be employed to achieve such an integrated argument, as well as how two of the methodological tools most frequently employed to muster empirical evidence for validating test design, namely classical test theory (CTT) and Rasch analyses, complement each other in designing tests responsibly. While most language tests designed in South Africa have used CTT, the employment of Rasch analyses has been more limited. A secondary aim of the paper is therefore to provide applied linguists who work in the subfield of language testing with an example of how the latter kind of analysis can complement the former. In all, however, these disparate approaches must be integrated into the theoretical justification for the development of language tests, in order to satisfy a number of conditions for their responsible design.

Keywords: validity, validation, theory of applied linguistics, design principles, language assessment

VALIDATION OF LANGUAGE TESTS AND RESPONSIBLE DESIGN

A test of good quality begins its life in a deliberate design. The various arguments for considering not only the objective validity of a language test, but also the broader process of its ongoing validation, will not be repeated here. These two – validity and validation – are, respectively, the objective and subjective dimensions of justifying test designs that are regularly disputed in what is known as 'validity theory' (Weideman, 2012). However, despite ongoing contestations about conceptualising validity, or even whether tests may be considered to possess validity, some language tests do gain in reputation over time, and those reputations depend in good measure on their quality and technical adequacy (Davies & Elder, 2005), both of which are conceptually synonymous with 'validity' (cf. e.g. Messick, 1980). Indeed, as has been argued in a number of further papers, there is much to be gained by reconceptualising the validation of language tests as the process of ensuring that they are responsibly designed

(Weideman, 2017a, 2017b, 2019a, 2019b). The idea of responsible design both broadens and systematises the process of validation of a language assessment.

In the first instance, the notion of responsible design potentially broadens the scope of language test development and implementation by introducing a wide-ranging set of constitutive and regulative principles for their design (Weideman, 2009), instead of merely lumping all conditions together under one umbrella, and so conceptually overstretching the concept of validity. Second, the idea that language tests must be designed responsibly systematises and structures the process of validation by making each of the more than a dozen principles that are enumerated below, as well as their derivatives, a condition that belongs to a coherent conceptual framework (as opposed to their random employment in an eclectic conceptual bog). That framework is described in the next section. The further premise is that this conceptual framework applies in general to all applied linguistic artefacts, including those in the subfield of language test design. The framework not only provides the systematic basis for their integration, but also offers a checklist of conditions that must be fulfilled, for example, by tests of language ability, when claims are to be made in the process of validation about whether such assessments have been responsibly designed.

Though other language tests may also be referred to obliquely, this paper attempts to demonstrate how the idea of responsibly designed language tests can be applied primarily to the design, development and refinement of one particular language assessment. The test in question is a measure of the language ability of prospective or early-career employees in the banking sector. The analysis pays specific attention to the critical early-stage validation of that test. Not every conceivable principle of language test design is applicable with equal force in each stage of the design; the paper therefore focuses on a selection of principles that are usually assumed to be more critical to ensuring measurement quality in the early stages of test development. In the following sections, I turn to an elaboration of those principles and a rationale for their selection. In doing so, the paper follows the conventionally accepted procedure of viewing the process as an argument for which various sets of evidence must be produced (Kane, 1992, 2001, 2010, 2011; Van der Walt, 2012), though the format in which this procedure must be accomplished remains contestable. In what follows, a systematic framework for bringing together the 'warrants' sought to justify language test designs offers one suggestion of such a format. The analyses offered here take their methodological tools for securing empirical evidence for test quality from both classical test theory (CTT) and from Rasch analysis, showing that they are complementary.

PRINCIPLES OF RESPONSIBLE LANGUAGE TEST DESIGN

The starting point of the procedure outlined in this paper lies in the acknowledgement that the guiding or defining dimension of a test of language ability may be found in its technical aspect: the measuring instrument is thus viewed as one that is planned, formed, shaped, designed, developed and brought into being deliberately as an intentional design. The leading technical aspect of the instruments known as language tests has as its nuclear moment the idea of design, which itself is not further definable (Schuurman, 2009: 417; Strauss, 2009: 127, 157, 339). The three key sets of designed applied linguistic interventions, namely language plans and policies, language curricula and courses, and language tests and assessments, all carry the stamp of the formative, technical dimension of experience: all are designed interventions that are imagined, planned, formed and shaped to solve or address a crucial,

widespread or pervasive language problem or issue. To theorists, the technical aspect of experience therefore defines the scope applied linguistics: it is a discipline of design.

By considering how other dimensions of experience relate to that qualifying technical aspect of the design of language tests, the particular kind of applied linguistic artefact taken as example in this paper, a number of principles can be conceptualised. Since that conceptualisation has already been dealt with in detail in other publications (Weideman, 2017c: chapter 11; 2020a), only the relevant formulations of these principles are offered here. Because space does not allow a full explanation, the following exposition omits a number of potentially important arguments; to avoid misunderstanding, reference to these analyses is therefore advisable. Mainly, however, it needs to be noted that the conditions to which responsible language test design is subject range more widely than the conventionally identified issues of validity, reliability and fairness, crucial as these may be. Though they remain provisional, a more complete set of design principles (taken, and substantially adapted, from Weideman, 2017c: 225) formulated as injunctions to which language test design must conform, may look like this:

1. Integrate the multiplicity of components of the language test so that it is a unity within that multiplicity of components, which are integrated in orderly fashion to measure a unique ability or various different, but related, sub-abilities.
2. Specify clearly to the users of the test, and where possible to the public, its circumscribed and limited scope. Exercise humility by neither overestimating, nor making inappropriate claims about, what the measurement proposed can in fact accomplish.
3. Ensure that the measurements obtained are consistent, and obtain empirical evidence for the reliability of the instrument that has been designed.
4. Ensure effective measurement by using defensibly adequate test instruments or assessment material.
5. Have an appropriately and adequately differentiated test, in which each component is organised in such a way in relation to others that it provides insight into a functionally different sub-ability, but nevertheless works together organically as a viable whole.
6. Make the test intuitively appealing, acceptable and attractive to those who take it, who use its results and who are affected by it.
7. Mount a theoretical defence of how the language ability that is being tested can best be defined, in the most current terms, or at least in terms of clearly articulated and plausible alternative theoretical paradigms or perspectives.
8. Make sure that the test yields interpretable and meaningful results; that it is intelligible and clear in all respects.
9. Ensure that the test fits the level of ability of the candidates who will take it, so that it is appropriate and has relevance for the social sphere it is intended for. Make accessible to as many as are affected by it not only the test, but also additional information about it prior to its administration, through as many and diverse media as are appropriate and feasible.
10. Ensure utility by making the test an efficient and frugal measure, and obtaining the test results efficiently to ensure that they are useful both to test takers and those who will use the results.
11. Mutually align the test with language development interventions and policies, for example with the language instruction that will either follow or precede the test, and

harmonise the policy, test and instruction as closely as possible with the learning or language development foreseen in their design, and with the social environment.
12. Be prepared to give account to the users as well as to the public of how the test has been or will be used, whether its design is justifiable, and what may reasonably and legally be implied by its results.
13. Value the integrity of the test; make no compromises of quality that will undermine its status as an instrument that is fair and compassionate to everyone, and that has been designed with care and love, with the interests of the end-users in mind.
14. Spare no effort to make the test significantly trustworthy and reputable.

As can be seen, when the principles for language test design are formulated thus, they include the well-known notions of reliability (principle 3), validity (principle 4), face validity (principle 6), construct validity (7) and fairness (13). In addition, they ask for attention to appropriateness and domain relevance (9), the interpretability and meaningfulness of results (8), usefulness (10) and justice (12). None of these conditions for test design is unknown; what is different here is that they are systematically arranged and integrated into a single theoretical framework (Weideman, 2020b). What is more, when they are used to gauge the quality of a language test, their meaning has to be argued for, and interpreted. All are therefore open to further interpretation, and to argument. Principles are not realised in the positive shape of a design in any fixed or immutable way (Strauss, 2009: 291). As conditions for design, these principles characteristically leave open the possibility of being given positive form in a particular test design that differs from their fulfilment in other test formats.

As has been noted before, not all of these principles apply with equal strength at each of the five stages of test design (Weideman, 2019c). I therefore turn below to the specification of this general theoretical framework in the early-stage validation of a particular language test design. Before doing so, however, a short description is provided of the language test that serves as the main illustration, as well as of the population on which it was piloted.

TEST AND POPULATION

The language test whose early-stage validation is being used as the main example is an assessment of the ability of prospective or entry-level employees in the banking sector. The test is called the Assessment of Language for Economics and Finance (ALEF). Its purpose is to determine, for the training agency that has been contracted by the banks to do this initial training, whether the language ability of those undergoing the training is at the required level for entry into post-school, further tertiary or vocational training (NQF level 5).

In the period between November 2018 and November 2019, the language ability of a total of 458 prospective or early-career bank employees was assessed. The results of the assessment were reported to the training agency in several risk bands, where ‘risk’ is associated with the level of language ability needed for the successful completion of a training course at NQF level 5. Data from early piloting, as well as extensive experience in setting cut-off points for such tests at this level, were used to set the parameters for each risk band, as outlined in Table 1.

Table 1: ALEF score interpretations: levels and associated risk

Risk band	Interpretation	Range
4	Little to no risk of level of language ability interfering with academic performance	75+
3	Less risk of level of language ability interfering with academic performance	49-74
2	Some risk of level of language ability interfering with academic performance	40-48
1	High risk of level of language ability interfering with academic performance	0-39

Of the original 458 who wrote the test, the results of 446 were adjudged to be sufficiently complete to use in the statistical and other analyses of the empirical properties of the test.

The test as a whole is theme-based. In conforming to the condition of relevance (principle 9) as well as in an attempt to make the assessment intuitively appealing (principle 6), only texts, graphs or vocabulary relevant to economic life and financial issues or topics were considered for use. Moreover, texts were selected with reference to their Flesch Reading Ease indices and their Flesch-Kincaid Grade Levels, as calculated by Microsoft Word, so that, at least by these conventional estimates, their level of difficulty never exceeds or falls too far below the appropriate reading ease (around 50%) or school level (Grade 12 or slightly above). The format of the items is multiple choice, and answers are usually recorded on optical reader sheets, which are then scanned and marked. That goes some way towards fulfilling the criterion for responsible design implied by the norm of technical utility and efficiency (principle 10). ALEF is an 83-item, 80-mark test, and takes 90 minutes to complete. It consists of seven sections or subtests, shown in Table 2.

Table 2: Sections and subtests of ALEF

Section/Subtest	Description	Mark
Pre-test	Pre-test skim reading task	Unscored
1	Vocabulary in context; definitions	18
2	Text comprehension and making sense of numbers	20
3	Interpreting graphic and visual information	12
4	Register: matching text types	5
5	Making a whole from scrambled text	5
6	Grammar and text relations: word choice, function	20
Total		**80**

In the first section, intended as a lead-in to the test, test takers are required to skim through the text of subtest 2 (text comprehension) before answering three True/False questions about its overall meaning. This introductory section is followed by subtest 1, which tests the candidates' knowledge of vocabulary (especially that of phrasal verbs) and definitions, once again taken from the text of subtest 2, which was skimmed in the pre-test section. The arrangement of skimming a text that will feature later and testing vocabulary found in that same text is a deliberate design feature intended to introduce and organise the material (see principle 5) with which test takers are to engage in a piecemeal fashion and with incremental difficulty, in adherence also to principle 9, relating to the technical accessibility of the test. This is sometimes referred to as 'scaffolding', a technique that makes what may be experienced as highly challenging material somewhat easier by introducing elements

gradually and organising the components of a test so that they progress from the more accessible to the more challenging.

A more substantial challenge then follows: subtest 2 assesses insight about not only the meaning and implications of what is stated in the text, but also simple numerical computations that either relate to or are implied by quantifiable information in the text. In subtest 3, interpreting graphic and visual information, test takers are required to interpret a graph by discerning trends and recognising proportions and fractions, again doing simple numerical calculations, and making extrapolations from the numerical data in the graph. Subtest 4 assesses the ability of test takers to recognise matching stretches of text, usually of sentence or similar length, taken from five different text types, in order to assess genre sensitivity. In subtest 5, by unscrambling a paragraph with five sentences, candidates show that they can restore the wholeness and integrity of a text. In the final subtest (6), they consider a more or less systematically mutilated text (a modification of a cloze procedure), and have to say where the word is missing and which word would fit into that space. This final subtest assesses their ability to understand the functional purposes of words like verbs, nouns, adjectives, prepositions, articles and conjunctions. As in subtest 5, the task also mimics what happens during text editing and making lexical choices.

Though it is not further argued here, it is necessary to note that, in limiting the technical scope of the test to measuring language ability within a particular domain (economics and finance), by attending diligently to pitching the test at the right level (NQF 5) and by stripping away, in early piloting, those items, tasks or even subtests that did not perform, the design of ALEF already sought to conform at an early stage with the condition noted above as principle 2: that the technical range of the test be limited, and acknowledged as such. The further adherence to this principle would be demonstrated in how, in light of the interpretation of the test results offered to those who will use them (see principle 8 and Table 1), the expectations of what the results can be used for will be managed and technically contained.

EVALUATING THE FULFILMENT OF CONDITIONS OF LANGUAGE TEST DESIGN

In addition to the first attempts to fulfil the seven requirements which have been referred to in the previous section, namely technical or instrumental relevance, appeal, utility, organisation, accessibility, scope and meaningfulness, several other principles of test design were prominent in the early-stage validation of the test. In order to select the further principles that were most relevant from the framework of principles for test design, the test designer focused on what is conventionally considered to be the most prominent or critical in the early stages of test validation. Not surprisingly, many of these rely heavily, though not exclusively, on the interpretation of quantitative data and the outcome of technical analyses of the empirical, factual properties of the test. In this section, the selected principles are enumerated, as well as the subsidiary questions that might be asked to demonstrate the fulfilment (or breach) of those principles. Following the methodology suggested by Van der Walt and Steyn (2007, see also 2008 and Van der Walt, 2012), one or more claims are made for evaluating whether the principles have been met. In Van der Walt and Steyn's (2007) seminal study, 10 claims were made in order to test the design of an Afrikaans undergraduate test of academic literacy. Nonetheless, their methodology is more generally usable, and a number of claims were therefore also investigated for ALEF, some similar to those in their original investigation. The results of the empirical measures employed to investigate whether the claims are warranted

are reported on in the next section (results and discussion); here, only the principle, the question flowing from that (with number, e.g. [1], indicated), the related claim [n^c], together with the various analyses (numbered 1.1, 1.2, 1.3, 2.1, and so on) to be undertaken, are provided.

For the first principle selected, that of technical integrity (principle 1), the question was whether the test performs as a technical **unit within a multiplicity** of components. As a measure of fulfilling this condition, the following subsidiary question was asked about the test results:

[1] What empirical measure(s) of homogeneity or heterogeneity can be offered to demonstrate an acceptable level of homogeneity for the test?

In light of this question, the following claim was investigated:

[1^c] ALEF exhibits a sufficient degree of homogeneity, for which there are several warrants.

In order to answer question [1] and investigate the claim derived from it [1^c], the test results were subjected to a factor analysis as first analytical measure (1.1). This is a statistical criterion used in CTT that indicates the degree of homogeneity of the test. The results were then subjected to a second analysis (1.2), belonging to the Rasch model (Linacre, 2018), that concerns the degree of fit of each item with others in the test. In short, misfitting 'items degrade the quality of our measurement' (McNamara, Knoch & Fan, 2019: 47), threatening its integrity. Researchers may therefore examine values either of infit or outfit, depending on whether the variation is more or less than predictable. The generally agreed measure to be used here is the Infit mean square (Infit MNSQ), an average calculation of fit across all items.

A second principle relates to the technical **reliability** or consistency of the test (principle 3 in section 2, above). The question in this case was:

[2] Which measures will demonstrate that the test has acceptable levels of reliability, in that it consistently measures the ability being assessed?

The related claim to be investigated was:

[2^c] ALEF shows that it is a reliable measure of language ability on several counts: at test level as well as at item level.

In order to test this claim, the results were subjected to five analyses. First, investigations of two measures of test level reliability, namely coefficient (colloquially: Cronbach's) alpha – usually identified as a more conservative index, giving lower readings – and greatest lower bound (GLB), a less strict measure, were done (analysis 2.1). Next, what is known as 'person reliability' in the Rasch model, which is related to the former two CTT measures (2.2), was drawn. Third, another Rasch-derived measure, called item reliability, which is an estimate of item reliability across the test as a whole, was examined (2.3). Finally, two measures of the average item-total correlation were utilised, expressed as a point-biserial correlation calculated in Rasch analyses together with the infit statistics (as in analysis 1.2) for every item, and in CTT both for items and at test level as a total point-biserial correlation (*Rpbis*), i.e., as a correlation of the score with the total score, or as the *Rit* (for item-rest correlation). *Rit* is a Pearson coefficient correlation of the item scores and the test total scores, which

provides a measure of the discriminating power of items (CITO, 2005: 29). For the latter, only the mean *Rpbis* and the *Rit* are reported in analysis 2.4, as calculated by Iteman 4.4 (Assessment Systems Corporation, 2017) and TiaPlus.

A further principle concerns the ability of the test to function as a technically **differentiated** but whole assessment, as noted in principle 5: the functionally different components (subtests) of the test must each contribute potentially unique information about a sub-ability, while still working together with the other subtests and the complete test as an organised, viable whole. The principle relates to the technical organisation of the test, and the following question, deriving from it, was formulated:

[3] What empirical evidence is there that the test is organised as a differentiated but technical whole, with functional parts working together, each contributing to the viability of its measurement?

The claim made by the test designer in response to this was:

[3^c] ALEF is organised as a differentiated whole, with each subtest functioning both uniquely and together with others in contributing to the viability of the measurement.

The way in which this can be tested is by considering both the correlations between the various parts of ALEF (its subtests) and the correlation of each of them with the overall test. The analysis (3.1) was done both in Iteman 4.4 and TiaPlus (CITO, 2005); the latter, being more comprehensive, is reported on in this paper.

A fourth principle, already addressed in part, concerns the further fulfilment of the criterion of the technically stamped **appropriateness** and **relevance** of ALEF. Thus, the following question was formulated:

[4] What empirical evidence, apart from the considerations of relevance evident in the selection of theme and materials, is there that the test is technically appropriate?

The related claim was:

[4^c] ALEF exhibits an adequate degree of fit by distributing candidates normally as regards language ability, while it simultaneously has an acceptable degree of difficulty; moreover, it can be demonstrated that the test fits the ability of candidates, in that there is a likelihood of minimal misfit either of items or persons in its measurements.

The first two analyses to provide empirical evidence for claim [4^c] could be done through the conventional CTT-derived measures of plotting the distribution of candidates and the deviations from the norm in that respect, as well as by considering the mean *P*-value (percentage correct) or facility of the test (as already shown in analysis 2.4). To show the appropriateness of items to the ability of persons, both the kind of matching shown by item-to-person fit (4.1) and person-to-item fit (4.2) as done in a Rasch analysis are reported in this paper.

The final two questions asked about the **interpretability** of the results (principle 8) and whether the test treats those to whom it is administered fairly. Thus, the following question was formulated with regard to interpretability:

[5] Are the results obtained through ALEF interpretable and meaningful?

This question led to the initial claim:

[5^{C}] ALEF yields scores that are clear, meaningful and intelligible.

The last question, relating to principle 13, that of **fairness** and beneficence (Kunnan 2000), was:

[6] Does the test treat those measured fairly? Does it foresee a way of mitigating potentially unfair results?

This question yielded the claim:

[6^{C}] ALEF measures so consistently that the number of potential misclassifications it produces is smaller than 5% of the total test population, and the test developers have a way of identifying such misclassifications in order to give those potentially misclassified a fair chance of taking a similar test.

This last claim could be investigated through re-examination of the Rasch analyses of item reliability (2.3) and item fit, while four different scenarios in CTT that compare two measures of test reliability, namely coefficient ('Cronbach's') alpha and GLB, in relation to a same test or a similar test (analysis 6.1) were used to supplement these data. Claim [5^{C}] is discussed in the next section, along with the report on other claims.

RESULTS AND DISCUSSION

For each of the above six claims, one or more sets of analysis can be offered as evidence of whether it is justified.

In order to consider, in examining claim [1^{C}], whether the test is an integral whole, one may look at the results of analysis 1.1, which can be graphically presented (Figure 1) in a factor analysis, generated by TiaPlus (CITO, 2005). In a factor analysis, the usual answer that test developers are seeking in relation to the technical unity of the assessment is 'whether all items are measuring the same trait (one factor)' (CITO, 2005: 19). Given the richness of many constructs of language ability, this is not always wholly possible. In the current case, however, only one item (number 42) falls outside of the desired single factor (factor 1). This item may therefore require re-examination and either modification or replacement in subsequent versions of the test. Nonetheless, the degree of homogeneity is more than adequate: to have one item out of 80 not performing as expected is more than acceptable. What is more, when the factor analysis of the particular subtest (on the interpretation of graphic and visual information) is drawn, this item measures, like the others, the single factor that is assessed by the particular group of items in that subtest. Despite being a slight outlier in the test as a whole, item 42 is therefore technically aligned with items in a similar task.

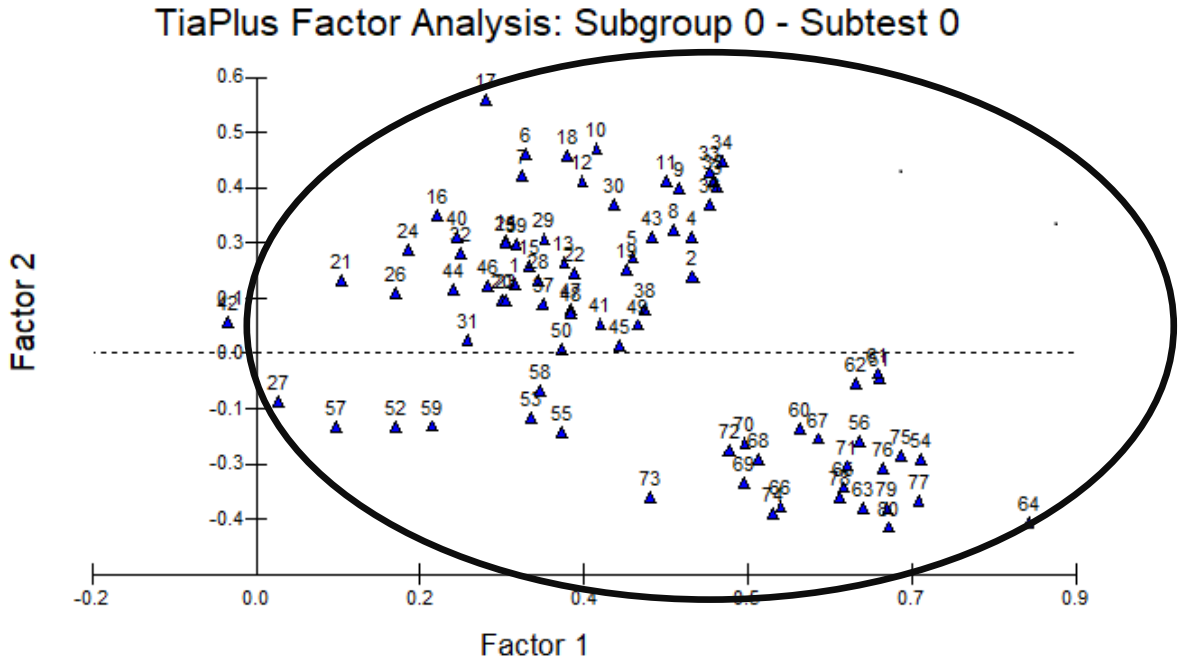

Figure 1: Factor analysis of ALEF (2018-2019)

The claim that ALEF is a technical unity therefore finds evidence in this first analysis, associated with CTT. How does it fare in this respect with the Rasch analyses? Consider first the infit readings generated by Winsteps. While Linacre (2018: 341, 354) notes that the benchmark for the measure of average fit (as in the calculated Infit MNSQ) is 1.0, with only the 'expected values … greater than 1.5 [being] problematic', others suggest even more conservative limits, in the range of between 0.75 and 1.3 (McNamara, Knoch & Fan, 2019: 45; Van der Walt & Steyn, 2007). In the case of ALEF, the stricter parameters were adopted, as is shown in the discussion of Table 3. For now, the relevant column indicating whether there is a second warrant for claim [1^{c}] is the one showing infit mean square, though I shall again refer to the column Ptmeasure-Al Corr below. The latter is defined by Linacre (2018: 244) as the Pearson point-biserial correlation 'for all observations including the current observation in the raw score', computing the 'correlation between the total … scores including all responses and the responses to the targeted item and person.' Since all 80 items in this analysis (1.2) show the measure of fit required by the more conservative parameters, only a truncated version of the full Rasch analysis is given in Table 3; the statistics associated with the better fitting items are omitted.

Table 3: Misfit order: items in ALEF

Item	Total count (n)	Infit MNSQ	Ptmeasure-Al Corr	Expected
42	446	1.07	0.03	0.21
31	446	1.09	0.20	0.33
27	446	1.23	0.07	0.34
57	446	1.14	0.08	0.27
21	446	1.18	0.13	0.35
24	446	1.10	0.19	0.33
16	446	1.06	0.19	0.28
	Further better fitting items not shown			
75	446	0.86	0.48	0.32
76	446	0.86	0.48	0.33
77	446	0.86	0.50	0.34
51	446	0.85	0.51	0.34
54	446	0.85	0.50	0.32
61	446	0.85	0.49	0.33
64	446	0.82	0.54	0.32

Table 3 gives the results for analysis 1.2. In this analysis, the two terminal values of infit mean square, 0.82 and 1.23 (shaded) fall well within the stricter limits of 0.75 and 1.3. The two analyses (1.1 and 1.2) therefore both confirm that claim [1^c] is warranted: ALEF exhibits the expected technical integrity, as its items show both homogeneity and overall fit.

The second claim, claim [2^c], involves the reliability of ALEF. To demonstrate this, analysis 2.1 yields the coefficient (Cronbach) alpha and GLB generated in CTT by programs such as TiaPlus (CITO, 2005). The ranges for coefficient alpha and GLB are normally, in these kinds of tests of language ability, above 0.85 and 0.9, respectively. Table 4, extracted from the relevant TiaPlus analysis, shows that, at test level, ALEF has a more than adequate technical consistency, with respective values of 0.9 and 0.97.

Table 4: Reliability (Cronbach's alpha and GLB) and related indicators: ALEF

Number of persons	446	Number of items	80
Average test score	40.45	Standard deviation	12.34
Average *P*-value	50.56	Standard error of measurement	2.35
Average *Rit*	0.34		
Coefficient alpha	0.90	SE coefficient alpha	0.01
GLB	0.97	Asymptotic GLB coefficient	0.96

Further evidence of reliability comes from analyses 2.2. and 2.3. Analysis 2.2 is known as 'person reliability' in Rasch, which is related to the former two CTT measures, and where an acceptable level is above 0.8 (McNamara, Knoch & Fan, 2019: 52). Here, ALEF expectably scored the same 0.9 as it did in the coefficient alpha measure of CTT, which is similar to that (McNamara, Knoch & Fan, 2019: 51), and so is not in this instance as strongly independent a measure only because it originates in a Rasch analysis. Nonetheless, it is well above the benchmark of 0.8 that was set. The other Rasch-derived measure used for this was item reliability, which is an estimate of item reliability across the test as a whole and which has no equivalent in CTT. The value for item reliability should also be higher than 0.8 (analysis 2.3), and ALEF scored a highly satisfactory 0.99.

In examining the fourth set of warrants for reliability, we may again consider the evidence in Table 3, and, for individual items, its last two columns. However, an equally useful and interpretable evaluation can be derived from analysis 2.4 on the overall values for these measures of reliability (*Rit* and *Rpbis*). In this respect, we note an average *Rit* (see Table 4) in TiaPlus of 0.34, and a mean *Rpbis* of 0.3 in the Iteman analysis. Depending on the strictness of the item reliability measure, test designers may be satisfied with values of 0.2 for individual items, whereas average measures, such as these, indicate a high level of reliability when they go beyond 0.3. On both of these counts, as on the other pair of warrants, ALEF satisfies the principle of technical consistency. A further observation that must be made in this regard is that the version of ALEF being reported on here is a refined pilot: even though the results being analysed here are for an early-stage test, the initial pilot (involving 217 second year students of accounting at a medium-sized residential university) has allowed the test developers to adjust or discard items, and even to add sections that function more effectively. The improvement in ALEF from reliability levels of 0.79 and 0.81 (on Cronbach's alpha) in its initial pilots to these highly satisfactory levels in the current test is due in good part to heeding the lessons of the initial pilot in this refined version.

The technical reliability of a test is not the final measure of its quality, however. Claim [3^c] allows us to examine a further requirement: the degree to which the different subtests work together as functional parts in an organised whole. Since the pre-test, introductory section of ALEF was not scored, it cannot yet be quantitatively evaluated in this respect. For that, it would perhaps need a post-test reception questionnaire such as those used by Van der Walt and Steyn (2007) in their earlier study. It is clear, though, that its placement and function is intended to make what follows a more viable assessment. For the scored data, however, we may consider as measures of technical functionality and organised differentiation the subtest-test correlations (which should preferably be higher, usually above 0.6), and the subtest inter-correlations, which must ideally be neither too high (with 0.5 as a possible upper limit) nor too low (with 0.2 as a lower limit), according to Van der Walt and Steyn. The idea is that each subtest must contribute a unique measure of the functional sub-ability being assessed (thus showing low correlations with other subtests that are intended to measure different sub-abilities), while continuing to function viably with the others overall (yielding higher subtest-test correlations). We may examine these values in the results of analysis 3.1, taken from the TiaPlus calculations and presented in Table 5.

Table 5: Test-subtest correlations, and subtest inter-correlations: ALEF

Subtest		Test	1	2	3	4	5	6
Vocabulary in context	1	0.68						
Text comprehension	2	0.72	0.48					
Interpreting graphic & visual information	3	0.63	0.34	0.47				
Register & text type	4	0.50	0.18	0.20	0.25			
Scrambled text	5	0.47	0.14	0.21	0.29	0.31		
Grammar & text relations	6	0.78	0.30	0.33	0.30	0.38	0.37	
Number of testees		446	446	446	446	446	446	446
Number of items		80	18	20	12	5	5	20
Average test score		40.5	11.5	10.8	5.5	2.5	1.5	5.6
Average P-value		50.6	63.8	54.0	46.0	49.9	37.8	41.4
Standard deviation		12.34	3.77	3.61	2.29	1.54	1.31	5.55
SEM		2.35	1.32	1.74	1.48	0.75	0.71	1.22
Coefficient alpha		0.90	0.75	0.69	0.60	0.58	0.51	0.90
GLB (if available)		0.97	0.88	0.81	-	0.80	0.67	0.96
Asymptotic GLB		0.96	0.88	0.77	-	0.77	0.71	0.95

The *P*-values for the test as a whole (51%), and for the subtests, indicate that the progression from easy to more difficult has almost been accomplished: the first subtest average facility was 64%, while the average score for the last two subtests was 38% and 41%, respectively. ALEF therefore conforms to the requirements of technical organisation (principle 5) of the various subtests and the technical accessibility (principle 9) of the test as a whole, since its facility decreases incrementally. The more important numbers, however, concern the test-subtest correlations and subtest intercorrelations. In respect of the first, only two of the six subtests, 4 and 5, fall outside the desired parameters, with values lower than 0.6 (shaded). But they are at the same time the shortest two subtests: each has only five items. Moreover, the TiaPlus calculation of the estimated coefficient alpha if subtest 4 had a standard norm length of 40 items shows that its technical consistency would then have been at a very satisfactory 0.95. In the case of subtest 5, the same estimated alpha for a potentially longer version stands at 0.92. So, while not conforming to the original conservative parameters, it would be premature to discard these two subtests. Rather, as has been suggested for these kinds of tests (Weideman, 2019d), one may consider adjusting the requirement downward to 0.5, in which case they would not have been flagged. As regards the second measure, the subtest intercorrelations, only one out of the 15, between the vocabulary test and the scrambled text task, is too low. On the whole, therefore, though not fully, the test satisfies the condition of technical viability, with its differentiated parts functioning together.

In investigating claim [4^c], about the degree of fit with and appropriateness for the ability of candidates, we may first look at the mean *P*-value, which measures the overall facility of the test. As both CTT analyses reported in Table 4 and 5 show, this is at 51%, and therefore as close to the desired 50% as can be. We may also examine the results of two Rasch analyses, envisaged as 4.1 (item-person map) and 4.2 (person-item map), as in Figures 2 and 3.

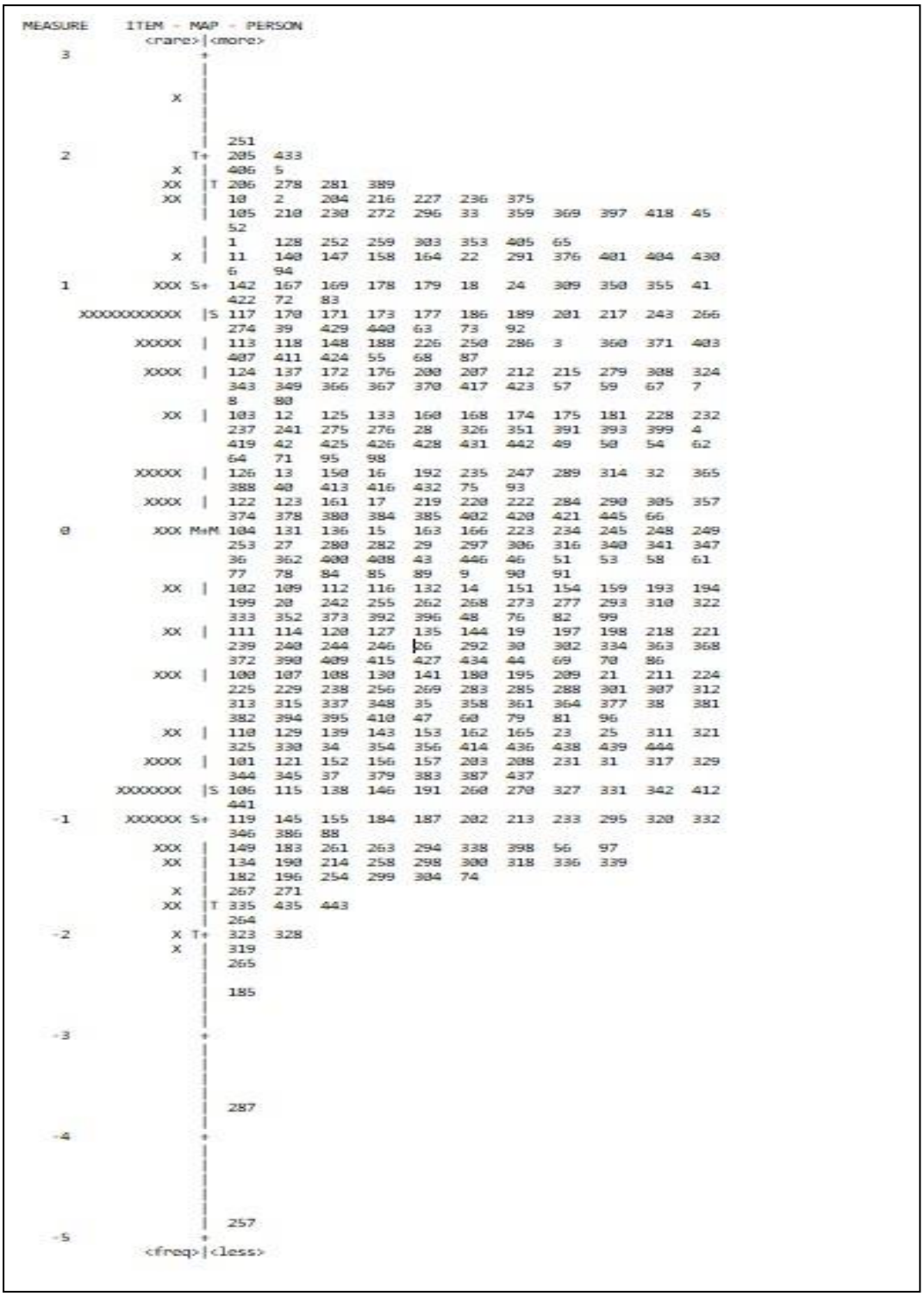

Figure 2: Wright map: item-person distribution map

A number of more sophisticated inferences can be made from these representations (see McNamara, Knoch & Fan, 2019: 34ff.), but for the present, we may focus only on the warrants for claim [4^{c}] to be found in these maps. In the item-person map in Figure 2, the extremes vary between log odds units, or logits, usually between -5 for a probability measure of a lack of success by candidates (numbered on the right from 1-466) in getting correct the items (indicated with 'X' on the left), and +5 to indicate their probability of success. The

parameters suggested by Van der Walt and Steyn (2007) indicate that we should be wary of a measurement that falls outside the -3 and +3 logit parameters. In Figure 2, we observe only two candidates (out of 466), candidates 257 and 287, who fall outside of the parameters. There is thus a significant degree of person fit.

The same parameters apply to the person-item distribution, represented in Figure 3. Once again, there are no items (on the right) that fall outside the -3 and +3 logits parameters. What is more, the distribution of candidates (on the left) indicates a fairly normal curve. Though item 42 again becomes noticeable as an outlier (it is the most difficult item), two other items, 45 and 16, also show up at the other extreme (they are too easy, as is confirmed by the CTT analysis).

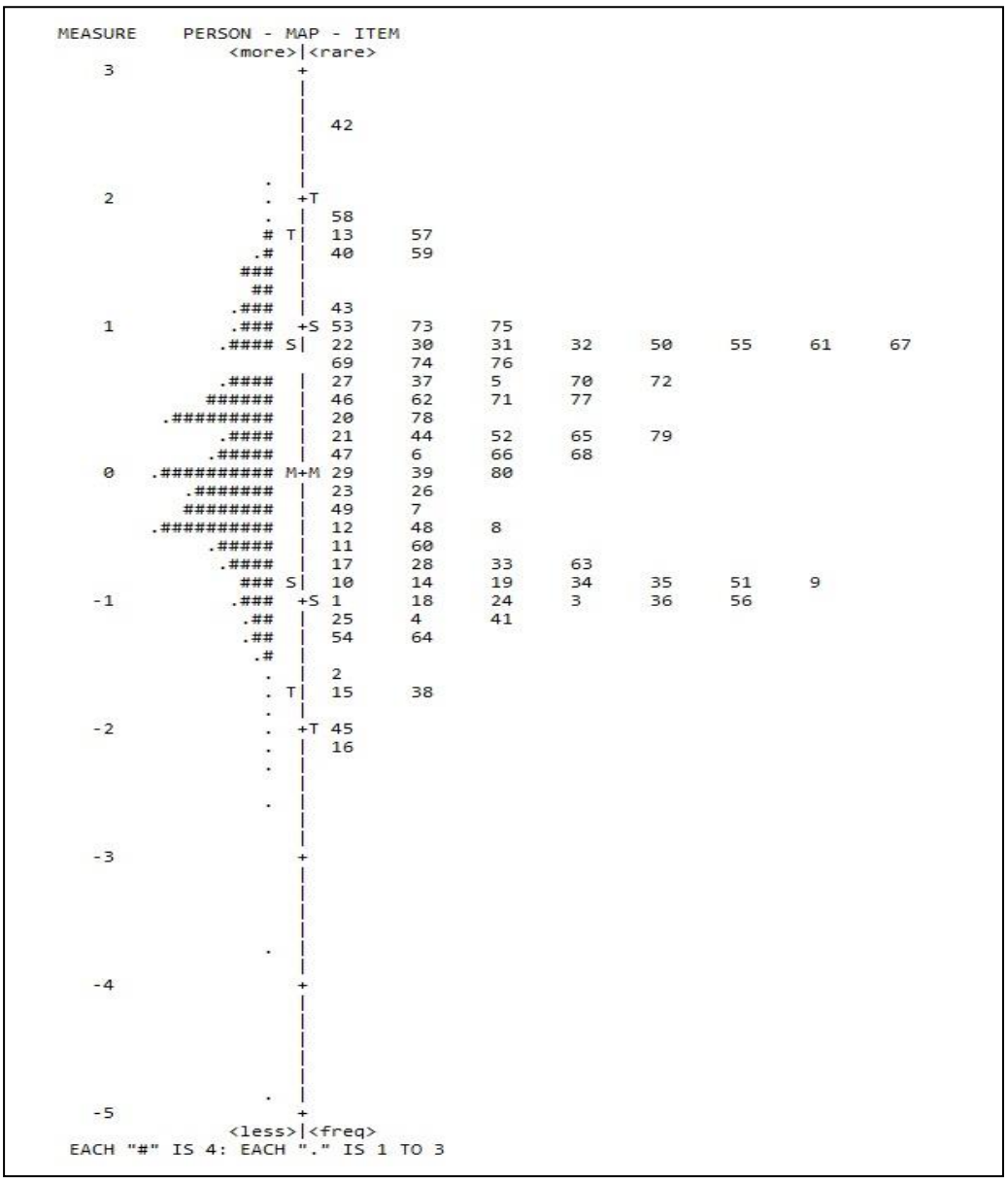

Figure 3: Wright map: person-item distribution map

We may therefore conclude that ALEF has an adequate degree of fit between the ability of the persons taking it, and a more than adequate likelihood that test takers will be able to manage

both the easy and more difficult items. Once again, we might question whether the parameters employed are appropriate. This time, however, they may seem to be too lenient. In fact, Keyser (2017) has suggested that in high-stakes tests, one might consider setting them more conservatively at between -2 and +2 instead.

Claim [5^c] has already been demonstrated in the interpretation of test scores in Table 1 above. Further investigation of just how these technical interpretations of the score assist users, such as the training agency whose students sat for ALEF, must still be undertaken. But the kind of interpretation given in Table 1 has worked well in the publication of the results of more than 30 000 students on undergraduate and postgraduate tests of academic literacy at several South African universities over the last 15 years (Weideman, 2020b).

The final claim, [6^c], concerns fairness. There are many possible analyses that may be done here, including testing for differential item functioning (DIF) to see whether the test places one group of students (e.g., as defined by gender or language) at a disadvantage or advantage over students belonging to another group. In the current case, this may be possible, but there has been no indication yet that there is a sufficient degree of dissimilarity that it needs to be specially attended to. Nothing precludes doing DIF analyses the moment that this indeed appears to have become relevant. One early indication that the test treats candidates fairly is the result of calculating the number of candidates that might potentially have been misclassified by the test by using four scenarios (alpha- or GLB-based; same test or parallel test [Rxx or Rxt case] – CITO, 2005: 17-18). Technical reliability has already been demonstrated in analyses 2.3 and 2.4 above. Analysis 6.1, generated by TiaPlus and reported in Table 6 below, shows that of the 466 who took ALEF, a maximum of 34 candidates might have been misclassified in the worst outcome, and a minimum of 14 at best. On the basis of there being an even chance of being misclassified above or below the cut-off point, that means that at worst 3.75%, or 17 candidates, may have been unfairly treated, which is well below the 5% mark of claim [6^c]. The calculation enables the test designers and users to offer a second chance test to the first 17 candidates below whatever is chosen as the cut-off score.

Table 6: Potential misclassifications in the administration of ALEF (2018-2019)

Misclassifications				
Alpha-based			**GLB-based**	
- Rxx' case:	Percentage	7.5%	Percentage	4.5%
	Number	34	Number	20
- Rxt case:	Percentage	5.4%	Percentage	3.2%
	Number	24	Number	14

All claims made in relation to six different principles of test design are therefore warranted, and we may provisionally conclude that the present refined pilot of ALEF needs little further attention. Since this is an early-stage validation, however, other principles than those selected here still need to be brought into play. I return to a discussion of those in the final section below.

CONCLUSION: RASCH AND CTT PROVIDE COMPLEMENTARY EVIDENCE

In all of the reports in the previous section, it is evident that CTT and Rasch analyses sometimes echo, and at other times complement one another. Especially in the case of the Rasch analyses, they offer additional evidence, from another point of view, about the

probability of how a test will perform. In that sense, Rasch analyses show their main difference from the descriptive and inferential data yielded by CTT analyses, usually only for one particular set of test takers: Rasch analyses focus instead on the likelihood that any test taker of a certain ability, independent of the group whose results were analysed, will perform in a predictable way on the items in the test. This kind of complementary information serves to refine a test further, but has been used in South Africa only to a limited degree, by researchers such as Van der Walt (2012) and Van der Walt and Steyn (2007, 2008; also Weideman, 2019d), as well as by Keyser (2017), the latter in developing an Afrikaans test of academic literacy for postgraduate students. The conclusion of this paper is that Rasch analysis deserves more attention, though CTT will still be employed, especially for smaller groups, since Rasch needs larger numbers. Where numbers allow, there is no reason for not employing both, and for making increasingly sophisticated inferences about the technical strength of both items and tests.

The further conclusion is that these methodological tools can be put to good use at least in the early stages of what is conventionally called the validation of a test, a process that has here been framed as one of demonstrating responsible design. In fact, the Rasch and CTT analyses employed in this instance gain more prominence, and are more clearly expressed, when placed into a systematic framework for responsible test design. The framework allows the identification of principles for responsible test design, of which a good number have been investigated above. Necessary as their fulfilment may be early on, they are, however, not sufficient. Not every principle in the framework offered in the second section above has been investigated. The test in question, like all others, still needs to be justified, for example, at least as regards its construct (principle 7) and face validity (principle 6), its alignment with language policies and instruction (principle 11), and its reputability (principle 14). Hence, this early start and the early indications that ALEF is potentially a test of good quality anticipate a further justification that may have to employ various additional methodologies and may well result in its further refinement.

The third and final conclusion is that the process of validation, when viewed as the justification of the design of an instrument for measuring language ability, can be given systematic form when viewed against a framework that is embedded in a theory of applied linguistics. Some (Rambiritch, 2012; Van Dyk, 2010; Keyser, 2017) have already experimented with this framework in the subfield of language testing. Others (Pretorius, 2015) have done so equally productively in another subfield of applied linguistics, language course design. That, too, may therefore be worth taking further in future investigations.

REFERENCES

ASSESSMENT SYSTEMS CORPORATION (2017). *User manual for Iteman 4.4.* Minneapolis, MN: Assessment Systems Corporation.

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

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

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

KANE, MT. 2001. Current concerns in validity theory. *Journal of Educational Measurement*, 38(4):319-342. DOI: https://doi.org/10.1111/j.1745-3984.2001.tb01130.x

KANE, MT. 2010. Validity and fairness. *Language Testing*, 27(2): 177-182. DOI: https://doi.org/10.1177/0265532209349467

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

KEYSER, G. 2017. Die teoretiese begronding vir die ontwerp van 'n nagraadse toets van akademiese geletterdheid in Afrikaans. MA dissertation, University of the Free State. URI: http://hdl.handle.net/11660/7704

KUNNAN, AJ. 2000. Fairness and justice for all. In AJ Kunnan (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. 1-14.

LINACRE, M. 2018. *A user's guide to Winsteps Ministep Rasch-model computer program. Program manual 4.3.0*. S.n.: s.l.

MCNAMARA, T, U KNOCH, & J FAN. 2019. *Fairness, justice and language assessment: The role of measurement*. Oxford: Oxford University Press.

MESSICK S. 1980. Test validity and the ethics of assessment. *American Psychologist,* 35(11):1012-1027. DOI: https://doi.org/10.1002/j.2333-8504.1979.tb01178.x

PRETORIUS, M. 2015. The theoretical justification for the design of a communicative course for nurses: Nurses on the Move. MA dissertation, University of the Free State. URI: http://hdl.handle.net/11660/683

RAMBIRITCH, A. 2012. Transparency, accessibility and accountability as regulative conditions for a postgraduate test of academic literacy. Doctoral thesis, University of the Free State. URI: http://hdl.handle.net/11660/1571

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

STRAUSS, DFM. 2009. *Philosophy: Discipline of the disciplines*. Grand Rapids, MI: Paideia Press.

VAN DER SLIK, F & WEIDEMAN. 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. DOI: https://doi.org/10.2989/SALALS.2009.27.3.3.937

VAN DER WALT, JL. 2012. The meaning and uses of test scores: An argument-based approach to validation. *Journal for Language Teaching,* 46(2):141-155. DOI: http://dx.doi.org/10.4314/jlt.v46I2.9

VAN DER WALT, JL & STEYN, HS. 2007. Pragmatic validation of a test of academic literacy at tertiary level. *Ensovoort,* 11(2):138-153.

VAN DER WALT, JL & STEYN, F. 2008. The validation of language tests. *Stellenbosch Papers in Linguistics,* 38:191-204.

VAN DYK, T. 2010. Konstitutiewe voorwaardes vir die ontwerp en ontwikkeling van 'n toets vir akademiese geletterdheid. [Constitutive conditions for the design and development of a test of academic literacy]. Doctoral thesis, University of the Free State. http://hdl.handle.net/11660/1918

WEIDEMAN. 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. DOI: https://doi.org/10.2989/SALALS.2009.27.3.3.937

WEIDEMAN. 2012. Validation and validity beyond Messick. *Per Linguam*, 28(2):1-14. DOI: https://doi.org/10.5785/28-2-526

WEIDEMAN. 2017a. Does responsibility encompass ethicality and accountability in language assessment? *Language & Communication*, (57):5-13. Breaking down barriers in applied linguistics: Studies in honour of Alan Davies (1931-2015); edited by CD Leymarie and SB Makoni. DOI: http://dx.doi.org/10.1016/j.langcom.2016.12.004

WEIDEMAN. 2017b. The refinement of the idea of consequential validity within an alternative framework for responsible test design. In Allan, J & AJ Artiles (Eds.), *Assessment inequalities: Routledge world yearbook of education*. London: Routledge. 218-236. DOI: https://doi.org/10.4324/9781315517377

WEIDEMAN. 2017c. *Responsible design in applied linguistics: Theory and practice*. Cham, Switzerland: Springer. DOI: https://doi.org/10.1007/978-3-319-41731-8

WEIDEMAN. 2019a. Degrees of adequacy: The disclosure of levels of validity in language assessment. *Koers*, 84(1). DOI: https://doi.org/10.19108/KOERS.84.1.2451

WEIDEMAN. 2019b. Validation and the further disclosures of language test design. *Koers*, 84(1). https://doi.org/10.19108/KOERS.84.1.2452

WEIDEMAN. 2019c. Definition and design: aligning language interventions in education. *Stellenbosch Papers in Linguistics Plus*, 56:33-48. DOI: https://doi.org/10.5842/56-0-782

WEIDEMAN. 2019d. Assessment literacy and the good language teacher: Four principles and their applications. *Journal for Language Teaching*, 53(1):103-121. DOI: https://doi.org/10.4314/jlt.v53i1.5

WEIDEMAN. 2021a. A skills-neutral approach to academic literacy assessment. In Weideman, J Read & LT du Plessis (Eds.), *Assessing academic literacy in a multilingual society: Transition and transformation*. *New Perspectives on Language and Education: 84*. Bristol: Multilingual Matters. 22-51. DOI: https://doi.org/10.21832/WEIDEM6201

WEIDEMAN. 2020b. Context, construct, and validation: A perspective from South Africa. *Language Assessment Quarterly*. https://doi.org/10.1080/15434303.2020.1860991.

BIOGRAPHICAL NOTE

Albert Weideman is Professor of Applied Language Studies and Research Fellow at the University of the Free State, and Extraordinary Professor in Language Education at the University of the Western Cape. His *Responsible design in applied linguistics: Theory and practice* (2017, Springer) has been followed by the jointly edited *Assessing academic literacy in a multilingual society: Transition and transformation* (2021, Multilingual Matters). He focuses on how language assessment relates to a theory of applied linguistics.

E-mail address: albert.weideman@ufs.ac.za
Professional website: https://albertweideman.com/