Language Teaching Research Quarterly **EUROKD**

2025 (50), 101–119
https://doi.org/10.32038/ltrq.2025.50.08

Underestimated Utility: Recording Design Decisions in Language Test Development

Albert Weideman
University of the Free State, South Africa

Correspondence
Email: albert.weideman@ufs.ac.za

Abstract
Our designs of language tests understandably focus on the endpoint: a useful test with interpretable results. We should not neglect what precedes that endpoint: the critical phases in the development of the test. We do not reflect enough on the development process, perhaps because we think that this is not worthy of research. Awareness of the stages of test design is useful, however: it pays handsome dividends when the quality of the end product is examined. The wide-ranging scholarly work of Carol Chapelle has focussed, among other issues, on two critically important components of testing language ability: assessing vocabulary knowledge and computer-assisted platforms for taking language tests. Their interaction will be highlighted here. Using a five-phase approach to language test design, this contribution sets out how the intuitive initial solution becomes the prompt to greater deliberation; how the articulated specification of what is being tested leads to greater theoretical defensibility; and how considerations of economy can be accommodated, leading to justifiable technical compromises in devising a quick test of reading levels. The initial solution associated reading ability closely with vocabulary knowledge only. The more deliberate subsequent design sought to overcome some of the limitations of that premise, by enriching the test with the addition of several further task types. The record of these design decisions indicates the rise in language assessment literacy (LAL) among the design team members, and how useful such a design record can be to enhance the quality of the eventual measure.

ARTICLE HISTORY
Received: 04 August 2024
Revised: 11 November 2025
Accepted: 02 December 2025

KEYWORDS
Language Testing, Test Design, Language Test Development

How to cite this article (APA 7th Edition):
Weideman, A. (2025). Underestimated utility: Recording design decisions in language test development. *Language Teaching Research Quarterly*, *50*, 101–119. https://doi.org/10.32038/ltrq.2025.50.08

©The Author(s), 2025. Published by EUROKD. This article is licensed under a Creative Commons Attribution 4.0 International License, which grants permission to use, share, adapt, distribute and reproduce in any medium or format provided that proper credit is given to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if any changes were made. www.EUROKD.com

[1]A Focus Not Only on the Endpoint, But on the Process

Among all the significant contributions that she has made to applied linguistics in general, and in particular to language assessment as a subfield of that discipline, Carol Chapelle has made three that are particularly relevant for the discussion that follows. The first two components of her scholarly work that will be referred to here have focused on critically important elements of testing language ability, the interaction of which will be highlighted in what follows. The first of these, assessing vocabulary knowledge (Read & Chapelle, 2001), concerns a dimension of that ability, while the second, computer-assisted platforms for taking the tests, focuses on their mode of delivery (as in Chapelle & Douglas, 2006). The third is her contribution to the validation of language assessments (e.g., Chapelle, 2021). Let me refer to that first.

The current orthodoxy in validating language assessments has focused our design and development of tests on their endpoint: a language test that can be useful because its results are interpretable. In fact, the technical meaningfulness of a language test is now accorded prime importance. Its quality is adjudged in terms of how the interpretation of its results can be theoretically defended. Such a defence is usually presented as an argument referring to its empirical properties, and offered in the form of a set of claims about its quality, as well as warrants to substantiate these claims (Drennan et al., 2024; Weideman, 2020).

Using a theory of applied linguistics which conceptualizes language assessment as a subfield of applied linguistics (Weideman, 2023a), this contribution will identify the interpretability of our measurement as a conceptual connection between two modes of our experience, the technical dimension of design, and the lingual function of expression. The theory sees the connection between the shaping (design, planning) of a language intervention like a test of language ability and the articulation and signification of this design as a fundamental concept of applied linguistics (Weideman, 2024a, pp. 161-177). This connection can be examined as one of technical meaningfulness, an idea which can be explored in a number of further concepts which tap into the analogical link between the technical and the lingual modalities. The first of these ideas is already evident in an argument-based approach to the validation of language tests: the requirement that the scores, the result of the designed measurement, be interpretable. Technical means and meaningful ends are paired in this requirement, and they refer to the connection between these two modes of experience, viz. the technical modality of designing, and the lingual mode of expression and signification. There are many conceptual connections besides technical interpretability related to this reflection of the lingual mode within the technical sphere. There is, for example, the expression of technical (design) specifications in the blueprint of a test. There are also the signification of intent in the declaration of test purpose, the technical articulation of the test construct, and the technical information

[1] This paper is part of a special issue (2025, 50-51) entitled: In honour of Carol A. Chapelle's contributions to language assessment and learning (edited by Christine Coombe, Tony Clark, and Hassan Mohebbi).

about the test made available publicly and accessibly to both test takers (Rambiritch, 2012) and test users. Of particular relevance to our current discussion, there is also the notion of recording the technical process of how the test is developed. All these conceptual variations of the idea of technical meaningfulness constitute fundamental, in the sense of basic, regulative ideas that steer the process of test development. In this perspective, technical meaningfulness can be articulated as a primitive in the conceptualization of fundamental applied linguistic notions.

If interpretability is the endpoint which language test designers and developers focus on, there is a risk that they may take for granted that this applied linguistic artefact is the outcome of an involved process of development, piloting, refinement and commissioning. They rightly have that goal in mind, since they wish to create a test that can eventually be defended in a diligent validation. This contribution will argue that we should not neglect what precedes that endpoint. The argument will demonstrate that we do not reflect often enough on or analyse the development process. As test designers we are indeed aware, in the process of test development, that there are critical phases, but at times we accept these as givens. Yet the professional growth of test designers in becoming more literate in language assessment (Weideman, 2024b) will be greatly aided by reflection on the process of design, which in turn will be supported by an adequate technical record of that activity. Where the test design and development team has inexperienced members, their professional growth in becoming more assessment literate will be an important consideration in keeping technical records of design decisions. What is more, it may become part of a drive towards greater transparency and accountability.

The necessary recording of test making may have been neglected as a result of an opinion that tracking the test development process is not important enough to be considered worthy of research (Read, 2010, p. 292). We may then remain unaware of the utility of such articulation of our work as test designers and developers. Greater awareness of the stages of test design leads to their being treated with the seriousness they deserve. That in turn, this contribution will demonstrate, pays handsome dividends when the quality of the end product is examined. Using a five-phase approach to language test design, the analysis will set out how the intuitive initial solution may – and should – become the prompt to designing language tests with greater deliberation, and, by stimulating the technical imagination of the test designers, contribute to their professional growth, or what is referred to as their language assessment literacy.

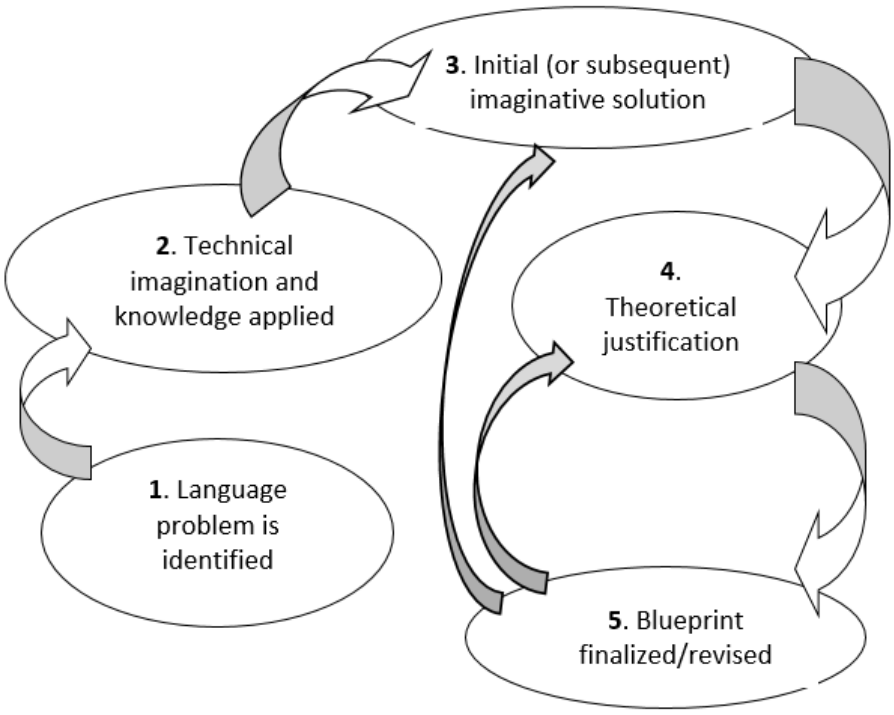
In what follows, I shall first outline the five-phase approach to test development (Green et al., 2024) which will be used as the theoretical framework for understanding the process. In the subsequent sections I shall then describe the test and its development process that will be used to illustrate the vacillations in design and their resolutions in each of the phases. Finally, the meaning of this analysis for the language assessment literacy of test designers will be discussed.

A Five-Phase Model for Test Development

Several conceptualizations of the process of designing language tests, as well as other applied linguistic artefacts, have been proposed. The most well-known probably is the eight-phase model proposed by Fulcher (2010, p. 94). Its start is the identification of the test purpose, and the articulation of the construct, before specifications are drawn up, and piloting undertaken. Finally, inferences are made from the test results, and the decisions reached on the basis of these may be fed back into the first stage. Similarly, Read (2015, p. 176-177) has proposed a five-phase process of planning, design, operationalization of the design, field testing, and implementation. Schuurman (2009, p. 43-44) has a more general three-phase proposal for all kinds of technical designs. All models proposed raise questions about their relation to actual actions and events, but their potential shortcomings or benefits will not be the focus of this discussion. In any event, all models have the advantage of being idealized renderings of design events, and are as a result flexible in their application. Instead of discussing in more detail their disadvantages and value, I shall use a proposed five-phase adaptation of the language test design derived from Schuurman's (2009) model. This adaptation, taken from Weideman (2024, p. 27), is schematically presented in Figure 1.

Figure 1

Five Phases of Applied Linguistic Designs



The model has already been applied in an analysis of test development and use at an Australian university (Green et al., 2024), as well as to analyse the design and production of a whole set of academic literacy tests (Weideman, 2021). The present discussion will therefore indicate whether it is even more broadly applicable. While the five phases outlined in Figure 1 are more or less self-explanatory, the further discussion below will

articulate them in greater detail. What must be pointed out, however, is that each may have several sub-stages, and that there may be a looping back either to seek another imaginative solution (a reiteration of Phase 3), or to secure a theoretical defence (Phase 4) for the final test and its blueprint (Phase 5).

In the discussion that follows the development of the test being used as illustration will be analysed with reference to each of the phases identified in Figure 1. Both the entity that commissioned the test, the names of the test development team, and the identity of those schools or test takers who participated in the pilot tests have been anonymized in the discussion, as well as in the log of design decisions which was used to record the process, an extract of which is to be found in Appendix A.

Phase 1: Challenge Identified

The test developer was approached by a private education provider in April 2022. The education services provider serves schools in several locations in Asia. It already had a test to be used for placing pupils in their schools at the appropriate reading level, which it intended to align with the gradings of the Common European Framework of Reference (CEFR) (Council of Europe, 2001, 2018). The test prototype, designated for the purposes of the analysis QARLp1 (Quick Assessment of Reading Level, prototype 1) was in multiple choice format, and was to be presented on a computer adaptive language testing (CALT) platform.

The intervention design problem was framed by the intended purpose of the test: to give teachers in the schools a quick measure of placing their pupils on a set of graded reading tasks according to the Common European Framework of Reference (CEFR) levels (Pre-A1 through to C2) (Council of Europe, 2001, 2018). The test developers agreed that the eventual test (QARL) should be long enough to reach a reliable estimate, and short enough not to take up too much time. It was evident from the start that a trade-off between reliability, economy and utility would have to be found, which was to be achieved without compromising test quality.

Phase 2: Trouble in Paradise

The client had initially wanted merely to have a validation of the test undertaken by experts. It had mustered local expertise both in test item development and in adaptive testing, and the items had been uploaded onto a CALT platform provided by another institution. In order to determine the difficulty of the items, the client had apparently appointed a panel of experts to judge it on a home-made scale. The existing knowledge and the technical resources at their disposal had been utilized, and applied to the solution of the language intervention design problem (see Phase 2, Figure 1).

The first dozen or so entries in the log of design decisions (see Appendix A) indicate that, after the project milestones had been set and the contract entered, a request was made in February 2023 for the data on the 427 candidates who had completed the current test

to be made available. Entry 15 on the log summarizes the problem: the data turned out to be unusable for a number of reasons, the most significant of which was that the estimate of difficulty had not been empirically determined beforehand, and the judgements of difficulty of items were purely subjective.

Phase 3: An Uncertain Beginning

Upon scrutiny, the test (QARLp1) was found to be inadequate in several further respects. Firstly, it consisted only of a number of vocabulary items, with some items measuring grammar. This was done on the basis of an argument that there is a strong correlation between vocabulary knowledge and reading ability. Though there are many claims about a higher correlation between these, a recent study in the same context of use which this test was being aimed at in fact indicated that this correlation is rather moderate, than large (Manihuruk, 2020). Thus, the test in its current prototypical format would not qualify as a test of reading ability. Secondly, there were no data available that could objectively verify the facility (levels of difficulty) of the various items. These had been subjectively judged by experts.

How does one overcome this obstacle? As has been pointed out in an earlier discussion of the five-phase model,

> *Phase 3 of applied linguistic designs is often characterized by experimentation and exploration … In the first two phases, there is little room for experimentation and trialing: all attention is focused on … preparing a preliminary design. It is only when that preparatory intervention is put to the test in Phase 3 that one can in fact begin to develop an initial, more informed opinion about its potential efficacy. (Weideman, 2024, p. 34)*

In determining whether a theoretical justification (Phase 4) and a final blueprint (Phase 5) are achievable, empirical data are required. Hence the decisions, recorded in entries 17 through to 21, to discard those items in the current databank of QARLp1 which, upon close scrutiny, had been identified as problematic. The entry on the log for 25 May 2023 records an initial purge of the database of some 68 problematic items, which left 532. It was decided to select two sets of items from the balance that remained for subsequent experimentation. The first was a selection of 40 items that tested vocabulary at mid-level ability, the second a mixed selection, also of 40 items, over all ranges of ability. The items selected for retrialing were placed on an online platform made available by the Inter-Institutional Centre for Language Development and Assessment (ICELDA, 2025).

The log records that when the accuracy of the estimates of item difficulty was tested, the item analyses revealed, as expected, that the judgements were unreliable, and indisputably so. The levels of difficulty assigned by the experts had almost no correlation with the actual facility values of the items which were tested. Note 31 on the log refers to a Tiaplus analysis (CITO, 2013) of the mixed-level test from which only one conclusion

could be reached: "There is thus very little empirical ground for the judgement of facility value". Furthermore, no specifications for developing items were available, or forthcoming (entry 17). This seriously detracted from the quality of the initial database, as did the observation that the items in the database contained numerous errors, or measured other than vocabulary knowledge, e.g., geographical knowledge. Some items could be salvaged and retained in building a new database for the CALT platform, but these needed empirical evidence of productivity (i.e., falling within the set parameters for their facility and discrimination values; see entry 48).

By May 2023, it was clear that the test as a whole had to be reconceived, and redeveloped. There was no way in which QARLp1 would successfully be validated, as its owners had initially thought possible.

Phase 4: Seeking Theoretical Justification

Though what is done in this phase may come earlier during the initial stages and planning, Phase 4 of language test design is, more often than not, focused on the elimination of the obstacles identified in Phase 3. A rethinking of the construct (of reading ability, as defined by the CEFR, with which alignment was being sought) was necessary, as well as an operationalization of that construct, with specifications and item types clearly articulated. Table 1 gives an indication of the initial design of what became QARLp2, the new version of the test to be developed. Eight of the nine subtests in it are (a) adaptations of the QARLp1 (subtest 1, and to a lesser degree subtest 2); (b) an augmentation of the assessment of vocabulary knowledge by testing both 2-word items (subtest 2) and collocations and phrases (subtest 3); (c) a further extension to conform to the prescriptions of the CEFR (2018) by testing for reading ability in terms of a number of purposes (subtests 4-7); and (d) the addition, finally, of an adaptation of cloze procedure (subtest 8) which has been shown to correlate highly with reading ability and academic literacy (Weideman & Van Dyk, 2023) and, if needs be, (e) a task type called Scrambled text in which a paragraph of five sentences needs to be restored to its original sequence.

Table 1

Initial Specifications of Subtests/Tasks: QARL

Number	Subtest
(1)	Vocabulary knowledge (one-word)
(2)	Vocabulary knowledge (two-word)
(3)	Knowledge of collocations and phrases
(4)	Reading for correspondence
(5)	Reading for information
(6)	Reading for orientation
(7)	Reading instructions
(8)	Grammar and text relations
[9]	Scrambled text

Several entries in the log of design decisions made in May 2023 record how the specifications for items were drawn up and communicated to the rest of the development team. The specifications for vocabulary items, for example, set out the steps to be followed:

1. *Select word to be tested from an acceptable word list (the New General Service List, the New Academic Word List, or reputable alternatives).*
2. *Select entries on the basis of frequency in the wordlist, referring to a defensible criterion for the selection, e.g., Nation's note on vocabulary size and the CEFR.*
3. *Write the item in the form of a sentence, and devise four distractors (A, B, C, D) based on either semantic, morphological, phonological or other defensible similarity (but preferably not a mix of these considerations in the same item).*
4. *Record the reason for formulating the item in this way, if it is not obvious or superfluous.*

In order to begin to align the items to be developed with the CEFR, Nation's (2023a) categories for the first 120 (for the A1 level) to the final 9000 (C2) were adopted. For this, Nation's (2023b) *First 10000 words headwords* list was used, as well as reference to the 31000 words in the New General Service List (2023) and its academic wordlist. The test developers knew that taking words from these lists for frequencies that have been related to the CEFR does not guarantee that the actual items will have similarly progressive difficulty levels (facility values, or *P*-values, for percentage correct), once tested. Another set of dynamics is then introduced, of the ability of the population, of the contextual reality of the question, and so on, but it is a first step towards achieving some kind of alignment, based at least on frequency of use. For the collocations, Pearson's *Academic collocation list* (2024) was used.

The same applies to the development of another two types of subtest which were considered to be potentially useful in measuring reading level, viz. the scrambled text task mentioned before, and the modified cloze procedure task that has been termed "Grammar and text relations". In developing these subtests, the measures of Flesch-Kincaid Grade Level and Flesch Reading Ease index were used, with the easiest texts intended for use on the lower (A levels), and the more difficult ones on the higher levels (C1 and C2) of the CEFR. Once again, the complexity or simplicity of the text in terms of such measures does not guarantee the facility of the items or subtests in which they will be used, but at least one has some starting point and indication of how to arrange progression from easy to difficult.

By November 2023, as the log indicates, hundreds of new one-word vocabulary items, split into six pilot tests, were ready to be tested out, and in January 2024 their results became available for analysis. Entry number 48 notes (Table 2):

Table 2

Entry #48 - Log of Design Decisions

Action/Decision	Notes
All Tiaplus analyses completed and commented on	Analyses with items to be discarded sent to QARLp2 administrator and Development team manager

In a critical moment, a design and development team meeting on 31 January 2024 faced a tough decision: the initial specifications of the QARLp2 as set out in Table 1 had encountered objections from the client related to both its technical implementability and potential length. As the log indicates, there was agreement that QARL should be both long enough to reach a reliable estimate, and short enough not to take up too much time. A trade-off between reliability, economy and utility was desired, which was to be achieved without compromising test quality.

It was therefore acknowledged that the test will, like many other language assessments, embody a technical compromise, in which, for the sake of implementability, a rougher measure is utilized as a trade-off. The greater concern then is reliability. Since no test is perfectly reliable, some misclassifications are possible, as in any other. Should there be indications of improper classification, a second-chance test for borderline cases will be needed. In order to accommodate all these concerns, a new set of specifications (as in Table 3) was drawn up.

Table 3

Eventual Specifications of Subtests/Tasks: QARL

	Subtest / Category	Marks	Requirement
(1)	Vocabulary knowledge (one-word)		
(2)	Vocabulary knowledge (two-word)	Not applicable: level assigned	Minimum number of answers as specified: (7) for each of the three categories
(3)	Knowledge of collocations and stock phrases		
(4)	Scrambled text subtest(s)	5 marks for each of a maximum of 5 subtests	Advance if score is > 3/5; when score drops to 2/5 or lower, assign previous level.
In doubt? Then administer a second-chance test:			
(5)	Grammar and text relations subtest(s)	20 marks for each of a maximum of 6 subtests	Advance if score is > 11/20; when score drops to below 10/20, assign previous level

The benefit of this technical compromise is that in borderline cases, or in cases where there is doubt after administering subtests (1) to (4), subtest (5) can be used as a second-chance opportunity to demonstrate reading level. The inclusion of the Scrambled text (subtest 4) and the Grammar and text relations (subtest 5) components was motivated by the consideration that vocabulary knowledge alone could not yield an adequate measure of reading level.

What followed in the busy process of subtest and item development is also adequately recorded in the log of design decisions. The record attests to the painstaking attention to conforming to specifications for item development within subtests (see entries 60-63), to the learning that takes place among new test item developers, and to how all kinds of objections can be handled (see entries 93 and 94 on the log). The Classical Test Theory (CTT) analyses of the pilots (done with Tiaplus: see CITO, 2013) and the Rasch analyses handled with Winsteps (Linacre, 2018) all contributed to the refinement of both items and subtests, and will be reported on separately. What was accomplished during this development and refinement process was, as is indicated in Figure 1, (a) finding a theoretical rationale to measure more than vocabulary knowledge, even when a quick assessment of reading level was required and (b) employing the analytical resources which enable us to understand the empirical qualities of tests better (see e.g. entries 48, 68-70, 72, 96 on the log). The test had become more theoretically defensible.

Phase 5: A Blueprint Emerges

While some piloting still remains to be done (see entry 96 on the log, attached as Appendix A), the final blueprint for QARL will probably be close to that of its second prototype, set out in Table 3. What still needs to be finalised, in other words, is the second-chance test, how it will be administered, and how its results will be interpreted.

On the CALT platform which is envisaged for QARL, there are some smaller issues of technical capacity which still need answers. But it is highly likely that the overall interpretation of the results of QARL, once it is operational, will be done in a manner related to the ranges suggested in Table 4. These were applied to a non-adaptive version of QARL, which was compiled from a selection of productive items, and then placed on Testportal as an interim assessment until the adaptive version was operational. This nonadaptive version and its results also acted as a further experimental version of the test, whose data were utilized in an initial validation by an independent validation panel. However, especially when the adaptive version becomes operational, this will need further experimentation and, where possible and feasible, calibration against other measures of reading ability.

Table 4

Proposed Interpretation of Marks: QARL

Level	Pre-A1	A1	A2	B1	B2	C1	C2
Cut Point	0	30	50	70	80	90	95
Range	0-29	30-49	50-69	70-79	80-89	90-94	95-100

Conclusion: The Usefulness of Keeping Records

The analyses presented in this contribution are offered from the point of view of language test designers and developers. They serve in this respect not only as illustrations of test design decisions, but also of how such decisions are recorded, of why such a record is

sound practice, and how these records may contribute to responsible design. Contrary to the opinion that test development itself is not intrinsically a noteworthy focus of research, I believe that the articulation of the test design in the form of a blueprint, and the subsequent record of how those technical requirements – most often including detailed test, subtest and item specifications – are met, are critically important disclosures of the meaning of language test design (Weideman, 2023b, 2024b). Such records may even be recognized by test users and policy makers as the first signification of professional, technical expertise on the part of the test developers.

One not entirely unexpected effect of the record keeping that accompanied the development of QARL is the usefulness of the information in the final column on the log (Appendix A). Here, the file references to which the actions or decisions referred, was noted. When a nonadaptive version of QARL was prepared for further experimentation, the client commissioned an initial validation of the test. Among other things, the log of design decisions was made available to the validation panel. In a private communication with the test developer, the coordinator of the validation panel expressed appreciation for this summary and index of the documentation needed. It made their work, which will be reported on separately, much easier.

I have playfully labelled the discussion of Phase 2 "Trouble in Paradise". In fact, the discussion of the second and third phases demonstrates a more realistic view of what happens in actual test design than the somewhat idealized processes that start neatly with the identification of test purpose, identify a construct, operationalize that in the form of specifications, and smoothly proceeds to trialing, refinement and finalization. The five-phase design process discussed here probably presents a more realistic perspective of what really happens in test design.

Not all language tests are designed with the support of the copious technical resources of the large institutional providers that make commercial language tests available globally. Their procedures are well established. Those who wish to design their own language assessments, however, and are often inexperienced in doing so, may not necessarily have unfettered access to the design decisions taken in the development of large commercial tests, or able to take advantage of what is publicly available, so as to learn from them. Those who intend to design their own, locally developed language assessments may well start, as in the case we have been considering here, with their own technical expertise and resources (Phase 2, Figure 1).

Upon closer examination, such efforts are quite likely to be shown to fall short. That needs acknowledgment: language intervention designs often start with initial solutions that muster existing institutional resources. Yet the initial imaginative solution is usually not the best. Instead, anticipating objections to the quality of the test, we loop back to ensure

that the theoretical defence of our design will stand up to scrutiny, and that the assessment will be appropriate, useful, and efficient.

That reciprocity between technical imagination and theoretical defensibility is yet another hallmark of applied linguistics, a conceptual primitive which allows us to understand our design work better. Theoretical analysis becomes the servant of technical imagination, and the corrections we make as a result of that interaction serve us well. We should record them, for reference. Their technical utility (Geldenhuys, 2007) will make a world of difference in a context where language assessment literacy levels are rising (Weideman, 2024b).

ORCID

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

Publisher's Note

The claims, arguments, and counter-arguments made in this article are exclusively those of the contributing authors. Hence, they do not necessarily represent the viewpoints of the authors' affiliated institutions, or EUROKD as the publisher, the editors and the reviewers of the article.

Acknowledgements

I wish to express my sincere appreciation for the Inter-Institutional Centre for Language Development and Assessment (ICELDA), which made trialing items easier by making its online testing platform available to the test developers, and also to the teachers and their pupils at various schools who have agreed to have their test results analysed for the further refinement of the test. I am especially grateful to the reviewers, who have made helpful suggestions for improving this manuscript. The final version remains my responsibility.

Funding

The test is a commercially developed test, for an client that for this reason remains anonymous.

CRediT Authorship Contribution Statement

Albert Weideman is the first and only author.

Generative AI Use Disclosure Statement

No use has been made of Generative AI.

Ethics Declarations

World Medical Association (WMA) Declaration of Helsinki–Ethical Principles for Medical Research Involving Human Participants

Participation in and data generated by participants were properly cleared with the client beforehand, as well as by the client themselves with their respective schools, and the data

were studiously anonymised, along with the exact settings in which the data were gathered. The research did not involve any human for experimental medical research.

Competing Interests

There are no interests to declare in this respect.

Data Availability

Enquiries with the author: albert@lcat.design.

References

Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. Sage.

Chapelle, C. A. & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press.

CITO. (2013). *TiaPlus users manual*. Arnhem: M & R Department.

Council of Europe. (2001). *Common European framework of reference for language learning and teaching*. Cambridge University Press. Retrieved August 20, 2022, from https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=09000 01680459f97

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment*. Companion volume with new descriptors. Language Policy Programme, Education Policy Division, Education Department, Council of Europe, Education.

Drennan, L., Joubert, M. & Weideman, A. (2024). Institutional responses to academic literacy challenges: An in-house test as an alternative for assessing academic literacy levels. *Journal for Language Teaching 58*(2), Article 6269. https://doi.org/10.56285/jltVol58iss2a6269

Fulcher, G. (2010). *Practical language testing*. Hodder Education.

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

Green, J., Davis, C., Judith, K., Harmes, M., & Weideman, A. (2024). Using a five-phase applied linguistics design to develop a contextualized academic literacy placement test for pre-university pathway students. *Literacy Research and Instruction*, 1–27. https://doi.org/10.1080/19388071.2024.2340031

ICELDA (Inter-Institutional Centre for Language Development and Assessment). (2025). Home page. https://icelda.com/

Linacre, M. (2018). *A user's guide to Winsteps Ministep Rasch-model computer program. Program manual 4.3.0*. s.l.

Manihuruk, D. H. (2020). The correlation between EFL students' vocabulary knowledge and reading comprehension: A case study at the English Education Department of Universitas Kristen Indonesia. *Journal of English Teaching, 6*(1): 96 –95. https://doi.org/10.33541/jet.v6i1.1264

Nation, I. S. P. (2023a). Vocabulary, the CEFR levels, and word family size. Retrieved May 11, 2023, from https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists

Nation, I. S. P. (2023b). The first 10000 words headwords [zipfile]. Retrieved May 11, 2023, from https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists

New General Service List Project. (2023). New General Service List and New Academic Word List. Retrieved May 11, 2023, from https://www.newgeneralservicelist.com/

Pearson PTE Academic. (2024). *The academic collocation list*. Retrieved February 8 2024, from https://www.pearsonpte.com/teachers/academic-collocation

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

Read, J. (2010). Researching language testing and assessment. In B. Paltridge & A. Phakiti (Eds.), *Continuum companion to research methods in applied linguistics* (pp. 286–300). Continuum.

Read, J. (2015). *Assessing English proficiency for university study*. Palgrave Macmillan.

Read, J. & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing 18*(1): 1–32. https://doi.org/10.1177/026553220101800101

Schuurman, E. (2009). *Technology and the future: A philosophica1 challenge* (Translated by H.D. Morton). Paideia Press. (Original work published 1972 as *Techniek en toekomst: Confrontatie met wijsgerige beschouwingen)*

Weideman, A. (2020). Complementary evidence in the early stage validation of language tests: Classical Test Theory and Rasch analyses. *Per Linguam 36*(2), 57–75. https://doi.org/10.5785/36-2-970

Weideman, A. (2021). A skills-neutral approach to academic literacy assessment. In A. Weideman, J. Read, & T. du Plessis (Eds.), *Assessing academic literacy in a multilingual society: Transformation and transition* (pp. 22–51). Multilingual Matters.

Weideman, A. (2023a). The practicality of theory: Reciprocity, assessment and applied linguistics. *Stellenbosch Papers in Linguistics Plus 66*, 177–196. https://doi.org/10.5842/66-1-935

Weideman, A. (2023b). Yardsticks for the future of language assessment: Disclosing the meaning of measurement. In M. R. Salaberry, W-L. Hsu, & A. Weideman (Eds.), *Ethics and context in second language testing: Rethinking validity in theory and practice (*pp. 220–234). Routledge. https://doi.org/10.4324/9781003384922-12.

Weideman, A. (2024a). *A theory of applied linguistics: Imagining and disclosing the meaning of design*. Springer. (Educational Linguistics, 65). https://doi.org/10.1007/9783031675591

Weideman, A. (2024b). Advancing professionalisation: The achievement of language assessment literacy. In B. Baker & L. Taylor (Eds.). (2024) *Language assessment literacy and competence Volume 1: Research and reflections from the field* (pp. 239-249). Cambridge University Press. (Studies in Language Testing Vol. 55).

Weideman, A. & Van Dyk, T. (2023). Achieving technical economy: A modification of cloze procedure. *Language Teaching Research Quarterly* (Special issue in honour of James Dean Brown's five-decade contribution to language testing and assessment), *37*, 144-160. https://doi.org/10.32038/ltrq.2023.37.07

Appendix A

Extract from Log of Design Decisions

	Date	Action/Decision	Notes	Deadline/ Completed	File Reference
2	6 April 2022	First quote	to agent		lct01305
6	25 August 2022	Milestones proposed	and dispatched to client		lct01369
11	15 January 2023	Work order and contracts signed, and dispatched			lct01388; lct01389
13	10 February 2023	Request for data on 427 candidates on test pilot	to QARL administrator, plus alternative options		lct01407
15	10 February 2023	Received test data: unusable, since randomly answered	Subjective determination of difficulty level flagged as problem		lct01409
17	10 February 2023	Lack of specifications for items identified	Alerted QARL administrator	23 May 2023	lct01414
18	10 February 2023	Lack of theoretical basis/word lists (e.g. Nation's) for choices	Requested documentation on word lists used	23 May 2023	lct01414
19	10 February 2023	Decision to repilot a 40 + 40 item sample	and to determine facility value (P-value) of items		lct01414
20	13 February 2023	Made Testportal available to QARL administrator for new pilots	through administrator of Testportal		lct01416
21	14 February 2023	Various problematic items identified (errors included)	Alerted QARL administrator about inadequacy of item bank; removal or correction of items requested; again noted lack of specifications, and lack clarity of		lct01417; lct01418

			what is being tested (construct lacking)		
31	9 May 2023	Tiaplus analyses done of mid-level test	Added notes to both mixed level and mid-level test analyses	9 May 2023	lct01477, lct01478
37	19 May 2023	Formulated draft specifications for vocabulary items	for discussion with Development team manager, and reformulation; dispatched with Nation's note on vocabulary size and the CEFR (lct01485).	19 May 2023	lct01489, lct01485
45	6 November 2023	Vocabulary pilots uploaded on Testportal	and linked to open platform		lct01539
46	1 January 2024	Vocabulary pilot test (6) results obtained	Cleaned, marked and analysed		lct01553 through to lct01598
48	24 January 2024	All Tiaplus analyses completed and commented on	Analyses with items to be discarded sent to QARL administrator and Development team manager		lct01559; lct01566; lct01573; lct01580; 1587; 1598
49	31 January 2024	Decisions on future steps in developing test	… It turns out the test has two purposes... In order to develop a shorter test for purpose 1, the number of subtests in the eventual test will be enlarged to include a section on Scrambled text. Albert will revise the specifications, identifying five tests for inclusion in the … test (purpose 1), namely subtests (1) to (3), as well as (8) and (9); … Albert will start to develop items for the subtests (8) and (9), based on texts with different		lct01485, lct01461, lct01318

			Flesch Reading Ease indices and Flesch-Kincaid Grade Level estimates, from about 60% reading ease (for easier texts) to more difficult ones (with a reading ease index of 45% or less).		
54	8 February 2024	Guidelines on developing items testing collocational knowledge	Dispatched Pearson's Academic Collocation List (ACL) to Development team manager and QARL administrator for comment and action		lct01607; lct01608
60	26 February 2024	Explained deletion procedure for Grammar and text relations subtests	and noted specifications, as well as numbers of subtests needed should my proposed promotion procedure be feasible.		lct01640
61	18 March 2024	Corrected Scrambled text questions moderated and dispatched	All texts now have unambivalent sequences of sentences		lct01658
63	18 April 2024	Curating of Grammar & text relations subtests completed	And dispatched to QARL administrator for uploading on Testportal in order to pilot new subtests	First week of May 2024	
64	18 April 2024	Report by Development team manager	to her manager about work planned from April 2024 to December 2024: from piloting to test refinement to validation		lct01691
65	10 May 2024	Three sets of pilot tests placed on Testportal	Pilot tests for measuring two-word vocabulary knowledge; knowledge of collocations and stock phrases; and scrambled text subtests have been curated,	10 May 2024	links to the various tests provided

			proofread, and links made available to schools and professional networks		
66	13 June 2024	Piloting of three sets of pilot tests completed	Data to be cleaned and analysed		
67	13 June 2024	Subtest 5: (Grammar & text relations) to be uploaded to Testportal	QARL administrator will upload		
68	8 July 2024	CTT analyses of pilot tests completed: Scrambled text	12 pilot tests analysed with Tiaplus (CTT); outcomes as intended, with desirable results.	8 July 2024	The analyses: all records between lct01764 and lct01795.
69	8 July 2024	Report to test development team on CTT analyses of Scrambled text pilots	Though numbers were small, results are usable, and indications are that except for pilot test 5 (on plastic) the others are all of good quality.	8 July 2024	Report on analyses: lct01882.
70	18 July 2024	CTT and Rasch analyses of pilot tests: Two-word vocabulary knowledge	6 pilot tests, with between 12 and 14 items each, analysed. Both CTT and Rasch analyses show highly desirable results.	18 July 2024	lct01796 to lct01842
72	20 July 2024	CTT and Rasch analyses of pilot tests: Knowledge of collocations and stock phrases	6 pilot tests analysed, with 10 items for tests 1-5, and 9 items for test 6. Both CTT and Rasch analyses indicate a high degree of quality in each subtest.	20 July 2024	Analyses lct01844 to lct01880

93	3 September 2024	Feedback that nonadaptive QARL test was too difficult noted	Feedback from teachers using the test that it is too difficult for their pupils, and that vocabulary appropriate for B level on CEFR, like collocations, is not appropriate for them. Response: if their pupils get in the 20% to 30% range, they are indeed, according to the scheme for the interpretation of marks, at an A level.		
94	4 September 2024	Feedback that the nonadaptive QARL test is too easy noted	Results of the nonadaptive QARL on Testportal analysed. Results indicate that for that population of 200+ the test was indeed a little too easy.	4 September 2024	lct01986, lct01987, lct01988
96	15 October 2024	First two pilot tests of Subtest 5 (Grammar & text relations) analysed	Only two of the eight subtests have been completed by a sufficient number of test takers to make analyses possible for what is intended as a second-chance test. The two that do have sufficient numbers have excellent reliability (0.85 and 0.99) are already usable, though the second one can be further refined.		lct02006; lct02011