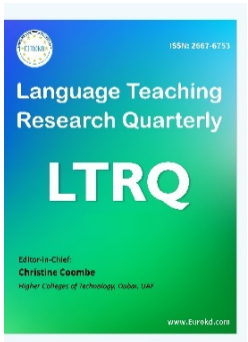


Language Teaching Research Quarterly

2023, Vol. 37, 144–160

Achieving Technical Economy: A Modification of Cloze Procedure

Albert Weideman[1*], **Tobie van Dyk**[2]

[1]University of the Free State, South Africa
[2]North-West University, South Africa

Received *22 April 2023* ***Accepted*** *28 September 2023*

Abstract

This contribution investigates gains in technical economy in measuring language ability by considering one recurrent interest of JD Brown: cloze tests. In the various versions of the Test of Academic Literacy Levels (TALL), its Sesotho and Afrikaans (Toets van Akademiese Geletterdheidsvlakke – TAG) counterparts, as well as related other tests used in South Africa, the test designers have used a modification of this procedure to very good effect. This paper reports on the steady evolution of its format over many years, how it is currently used, what its outstanding empirical properties are, and how the kind of technical economy it brings to the measurement of the ability to handle the demands of academic language at the level of tertiary education can be further applied. The modification involves the conventional, more or less systematic mutilation of a selected text, with two multiple choice questions about every gap in it: where the gap is, and which word has been omitted. We have not seen anywhere else analyses of this format, which in itself may be of interest to test designers. We proceed by defining technical economy, and then develop an argument on the basis of the empirical properties of TALL on how that idea can be applied, in particular to the design and task selection of such tests, before giving illustrations of how such choices may contribute to further and other productive and responsible designs and test formats.

Keywords: *Technical Economy, Cloze Tests, Test of Academic Literacy Levels*

Enduring Interest in Cloze Procedure

Though he also contributed more widely to discussions on language testing (e.g., Brown, 2014), and specifically to improving quantitative measures in applied linguistics (e.g. Brown, 2021; Purpura, Brown & Schoonen, 2015), cloze procedure has been an enduring specific interest in the work of J. D. Brown. Since the early analyses (Brown, 1980) through to the Brown and Gruter (2022) study, cloze tests have been scrutinized with persistent and growing sophistication. This paper will not seek to review all that has been written about cloze, and with increasing deliberation (see, for example, Trace, Brown, Janssen & Kozhevnikova, 2017; Trace, 2020). Its focus is rather on how a modification of cloze procedure has enabled one team

* Corresponding author.
E-mail address: albert.weideman@ufs.ac.za https://doi.org/10.32038/ltrq.2023.37.07

of test designers to apply it imaginatively, underlining in the process its efficiency and utility (Brown, 1980, pp. 312, 315), and doing so in a theoretically justifiable manner. All the while, we shall be mindful of the implication of the title and content of Purpura, Brown and Schoonen (2015), that language testing is a sub-field of applied linguistics (Weideman 2017a). Applied linguistics will therefore indeed be the background and context of this paper.

Furthermore, the analyses presented in this paper 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 (Read, 2010, p. 292; for a critical discussion: Weideman, Patterson & Pot, 2016), we 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, 2024).

In what follows, we shall address the issue of responsible test design first, setting out what the various principles are that we used as requirements for test development. One of these criteria concerns the achievement of technical economy, which we define against the emergent theory of applied linguistics described below. The interaction of that condition for responsible test design with other design principles will be noted and illustrated. We shall then briefly describe the evolution of the modified format for cloze, before examining its empirical properties in an administration of the Test of Academic Literacy Levels (TALL) in 2023, and their meaning for the achievement of technical economy. Finally, we shall link the discussion to how this new modified cloze format can fit, with other kinds of subtest, into a larger design of a useful and efficient test.

Methodological Starting Points and Definitions

Though many other definitions are indeed possible, we have adopted as a potentially productive starting point for this paper the notion that applied linguistics is a discipline of design (Van Dyk, 2022; Weideman, 2017b, 2022a). This means that its concepts involve making sense of a modality that we might term the technical mode or aspect of experience, which guides designed or planned language interventions (including language tests). ‘Technical’ in this respect does not indicate complexity, in the everyday sense of how that term is sometimes used (as in the statement: “It’s quite technical” – and hence possibly beyond casual understanding) but rather that test development is characterized by design, and that design (or planning, shaping, devising or fashioning) is an essentially technical act. When we conceptualize applied linguistic ideas and constructs, the defining characteristic of design constitutes the nuclear moment of technically qualified actions and roles (Schuurman, 2009, p. 384; Van Riessen, 1949, pp. 623, 625). Responsible test design may then be interpreted as the adherence to a number of principles for the development of applied linguistic artefacts that obtain across three major designed interventions: language policies, language tests, and language courses (Weideman, 2017b, p. 225). Selecting from these principles the ones which are most relevant for the purpose, we shall utilize the following criteria in our examination of the technical usefulness of cloze procedure in the design of language tests: technical homogeneity, technical

consistency, technical differentiation, theoretical defensibility of the design, and technical economy.

As a measure of the homogeneity of the tests to be discussed below, we shall take a coherent construct as potential evidence, along with empirical measures such as factor analyses and measures of item infit mean square. Consistency will be conventionally defined as reliability, using coefficient alpha as initial, and, if available, Greatest Lower Bound (GLB) (CITO, 2013, pp. 19, 31, 37, 38). Technical differentiation will be considered by scrutiny of the degree of subtest-test correlations, as well as subtest-intercorrelations. The theoretical defensibility of the test design will to a great measure depend on the theoretical defensibility of its construct, so a discussion of the construct will be undertaken. Finally, technical economy is defined as a measure of technical utility, the requirement to "obtain the test results efficiently and ensure that all are useful" (Weideman, 2017, p. 225). It is worthwhile considering Brown's (1980, p. 312) observation in this connection, specifically with the use of cloze:

> *Usability* is the practicality of a test. This concept includes such considerations as how easy a test is to develop, administer and score, and how much it costs to do so. If all other considerations (i.e., reliability, validity, mean item facility and mean item discrimination) are about equal for two tests, usability is certainly a justifiable basis for deciding which one to use.

By referring to reliability and validity, and using them in that early analysis, Brown's observation also emphasizes that the criteria mentioned above will be in interaction: fulfilling one requirement may depend on, influence, or contribute to another. Before we turn to the specific research questions, in the form of hypotheses (claims) and evidence (warrants), we first describe the evolution of the subtest of TALL that we are specifically focusing on to illustrate the productivity of this particular modification of cloze. The test task described in the next section features not only in TALL, but also in its equivalent counterparts in other languages, notably the Toets van Akademiese Geletterdheidsvlakke (TAG), for Afrikaans, and another, for Sesotho.

The Evolution of a Potentially Useful Format of Cloze

Our interest as test designers in a more efficient test began when we switched from a commercially available test to one with a more defensible theoretical construct (Van Dyk & Weideman, 2004a). Theoretical defensibility was for us what is conventionally labelled construct validity (Weideman, 2022b). What we gained in the changeover was a test in which the multiple-choice format was exploited to the full (Van Dyk & Weideman, 2004b; Patterson & Weideman 2013b), so that logistically, a few thousand tests could be administered and marked, and the results published, within 36 hours. Such a gain was already a clear instance of technical frugality (Schuurman, 2022, p. 81), a massive saving of resources, by cutting the time it took to hand-mark its predecessor from a full week to less than two days. It was a remarkable gain in technical economy. What we missed in making this change, however, was a shorter test that could be used for quick screening. Such a test is useful, as we shall note again below, since it can be employed to sift those who are already accomplished or skilled enough, and for whom writing a longer test would be a waste of effort and time. So we began to explore, as Geldenhuys

(2007) has recorded, the possibility of devising shorter tests of academic literacy without abandoning the new construct for TALL. The modified cloze whose evolution will be described in this section stood out as a prime candidate for inclusion in such a shorter test.

The subtest under scrutiny started its life in a format which had to be hand-marked. As an example, we take an adaptation of a subtest from a workbook of practice tests (Weideman, 2018), which is often used by candidates to prepare for academic literacy tests similar to TALL. The first example is how this particular subtest and the items it contained worked; the snippet of text that was systematically mutilated by deleting every seventh word (where possible). The frequency of deletion was arrived at through a process of informal experimentation, and was confirmed to be appropriate for texts at this level during piloting. The text was taken from a Wikipedia entry on 'Vulcanization', accessed in 2010.

> **Place an omission mark (/) in the space where a word has been omitted:**
> Charles Goodyear (1800–1860) invented the vulcanization of rubber when he was experimenting by heating a mixture of rubber and sulphur. The Goodyear story is one of either pure luck or careful research, but both are debatable. Goodyear insisted that it was the, though many contemporaneous accounts indicate the latter.

The correct answer is: after the word 'the', in the phrase "… insisted that it was the, though many…" However, it might just as profitably have been phrased thus:

> **Which word has been omitted in the place marked / ?**
> Goodyear insisted that it was the /, though many contemporaneous accounts indicate the latter.

Here the likely answer would have been 'former'. But what if one combined the two formats, as follows:

> **In the following, two words (A and B) have been omitted. Mark the space where the word is missing in the text, and in the column beside that write the word that has been left out:**
> Charles Goodyear (1800–1860) invented the vulcanization of rubber when he was experimenting by heating a mixture of rubber and sulphur. The Goodyear story is one of either pure luck or careful research, but both are debatable. Goodyear insisted that it was the, though many contemporaneous accounts indicate the.
>
> A. _____
> B. _____

The item is now much more productive: it measures four times. But the difficulty is that it needs to be hand-marked, and markers would be faced with the choices of how to score, which Brown (1980) had already identified. As he observes (Brown, 1980: 315), for ease of scoring nothing at that time came close to multiple choice. Skipping some of its further evolutions, one

of which we shall return to in the discussion below, let us show how the test developers of TALL came up with the current format:

In the following, you have to indicate the possible *place* where a word may have been deleted, and which *word* belongs there. Here are two examples:

Charles Goodyear (1800–1860) invented the vulcanization of rubber when he was experimenting by heating a mixture of rubber and sulphur. The Goodyear story is one of either pure luck or careful research, but both are debatable. Goodyear insisted that it was [i] the [ii], though [iii] many [iv] contemporaneous [i] accounts [ii] indicate [iii] the [iv].

Where has the word been deleted?
A. At position (i).
B. At position (ii).
C. At position (iii).
D. At position (iv).

Which word has been left out here?
A. indeed
B. very
C. former
D. historically

Where has the word been deleted?
A. At position (i).
B. At position (ii).
C. At position (iii).
D. At position (iv).

Which word has been left out here?
A. historical
B. latter
C. now
D. incontrovertibly

Now answer the following questions in the same way:

Goodyear claimed that he **1&2** [i] discovered [ii] vulcanization [iii] 1839 [iv] but did **3&4** [i] not [ii] patent [iii] the [iv] until June 15, 1844, and …

1. Where has the word been deleted?
A. At position (i).
B. At position (ii).
C. At position (iii).
D. At position (iv).

2. Which word has been left out here?
A. first
B. rubber
C. year
D. in

3. Where has the word been deleted?
A. At position (i).
B. At position (ii).
C. At position (iii).
D. At position (iv).

4. Which word has been left out here?
A. then
B. apparently
C. invention
D. fully

In its latest incarnation, following discussions among test developers about what is being tested (Trace et al., 2017), the subtest now has been marked as one measuring "Grammar and text relations". Their argument was that it measured not only vocabulary, but also insight into

lexico-grammatical position and meaning, including relations between different parts of a text, or, in some instances, an ability to deal with advanced grammatical features in a text, and even of being able to deal with communicative function. What is more, an argument was made that the challenge it presented (it was usually the most difficult subtest, and therefore placed last) indicated a level not only of so-called receptive language skill (Weideman, 2021), but an indication of being able to measure, if not writing, then pre-writing abilities.

Research Questions and Hypotheses

The interaction of the various principles referred to above will be gauged by how the realization of each in the test used as illustration here (TALL) may contribute to the selection of subtests that might be used in a potentially more economical, shorter test. We examine below several hypotheses in the form of claims, and the associated evidence for their substantiation, as 'warrants'.

In order to demonstrate technical homogeneity, we investigate the claim that

Claim 1: The use of modified cloze will contribute to the design of TALL, which consists of a multiplicity of components that are unified.

Warrant 1A: A factor analysis of the test may provide evidence for Claim 1.

Warrant 1B: An examination of the elements of the construct that have been operationalized into subtests will show whether the ability being tested is a theoretically unified idea of a particular language ability.

Warrant 1C: A Rasch analysis of infit mean square may reveal the technical integrity or lack of integrity of the items making up the test.

In order to show whether the test and its subtests are reliable, the claim to be investigated is

Claim 2: The reliability indices of TALL and its subtests will be an indication that a shorter test utilizing the most satisfactory of these may be constructed.

Warrant 2: The coefficient alpha and GLB (where available) measures will be determined to check whether they are at or above 0.9; the estimated coefficient alpha if the subtest had a standard norm length of 40 items, as calculated using the Spearman-Brown formula in TiaPlus (CITO, 2013, pp. 32, 40), should also conform to those requirements. In addition, a Rasch analysis can be employed to supplement this evidence in calculations of "person reliability" and "item reliability" for the test as a whole, as well as for the subtest whose contribution is the focus of this analysis.

For the potential demonstration of technical differentiation, the claim is:

Claim 3: TALL is a sufficiently differentiated test, in which each subtest contributes to the test overall, yet measures a different sub-ability. In combination with the possible warrants for Claim 2, this would enable an even more defensible selection of subtests for a shorter test.

Warrant 3A: Correlations of the subtests with the test will be investigated to check whether they are above 0.6 (see the re-examination of these by Brown, 2021), and

Warrant 3B: Subtest-intercorrelations will be calculated to see whether they lie between the parameters of 0.2 and 0.5.

The next hypothesis is intended to provide grounds for the construct of the test being theoretically defensible:

Claim 4: TALL is a designed operationalization of a theoretically defensible construct.

Warrant 4: Evidence from Warrant 1B, as well as discussions and analyses of the construct of the test, will be brought together to demonstrate this.

In addition to the foregoing, a final claim must be made:

Claim 5: There is sufficient evidence that a combination of subtests selected from the current set of subtests employed in TALL may be used for a shorter test, for the purposes of prior screening or other considerations of technical economy, and that the "Grammar and text relations" subtest will be a prime component of such a shorter test.

Warrant 5A: Determine whether the evidence from the warrants above, but especially for Claim 3, is adequate to enable the test designers to identify subtests that can be thus combined.

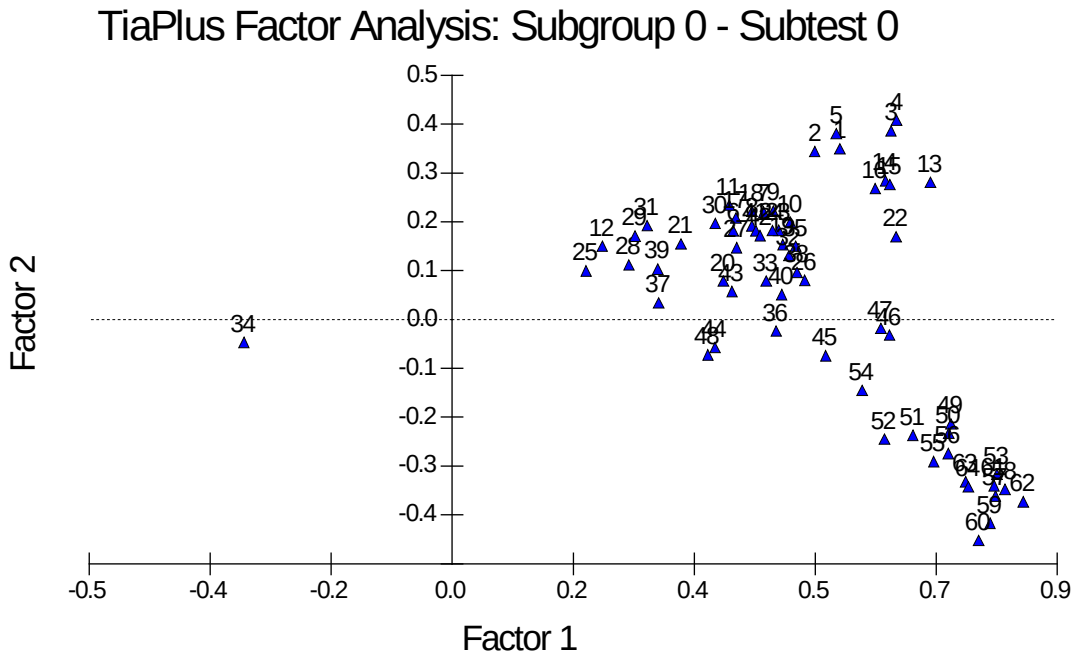
Warrant 5B: Do a correlational analysis of the marks obtained by candidates who wrote TALL on the test and its relevant subtests.

Results

We turn first to the evidence for Claim 1, that TALL has a level of technical homogeneity, by considering the factor analysis of its 2023 administration, generated as a graph (Figure 1) by a TiaPlus analysis (CITO, 2013). The test was administered to a population of 2137 first-year students in four different faculties at a large South African residential university; each participant was fully informed about the aim of the test, and agreed to their marks being used anonymously in empirical analyses.

Figure 1

Factor Analysis: 2023 Administration of TALL



It is evident from Figure 1 that TALL has a sufficient degree of homogeneity; only one item (number 34) is an outlier, and had already been identified as unsuitable by its negative discrimination value (*Rit*: Item-test correlation, a Pearson Product Moment Correlation coefficient – see CITO, 2013) of -20, where the selection criterion for TALL is a positive Rit value of 30. All the other items show a fit with a single factor. In measurement of academic language ability, TALL thus is a technical unity.

A second piece of evidence can be derived from the way that the test designers have justified their use of the "Grammar and text relations" subtest employing a modified cloze procedure, with reference to an articulation of the construct. The construct itself has been formulated and discussed frequently (Patterson & Weideman, 2013a, 2013b; Weideman, 2021, pp. 34-36, 2022b). In Table 1, the elements of this construct are listed, in summary form, in the left-hand column, while the subtests that operationalize the construct are listed on the right. For examples of the tasks and activities involved, see the sample test at LCaT (2023). It is clear that the modification of cloze described above serves to test several components of the construct: vocabulary, metaphorical expression, complex grammar, text relations, and even communicative function. One may conclude, thus, that the "Grammar and text relations" subtest fulfils a vital function in measuring several distinguishable sub-abilities of the construct of academic literacy. We return to this below, in the argument in support of Claim 4.

Table 1

Operationalizing the Construct of TALL

Understand / interpret / have knowledge of	Task type / Subtest measuring this
vocabulary and metaphor	Academic vocabulary
	Text comprehension (in larger context)
	Grammar & text relations (modified cloze)
complex grammar, and text relations	Grammar & text relations (cloze)
	Scrambled text / organisation of text
communicative function	Text comprehension
	Text type / Register awareness
	Grammar & text relations
	Scrambled text / organisation of text
text type, including visually presented information	Text type / Register awareness
	Text comprehension
	Interpreting graphic & visual information
essential/non-essential information, sequence and numerical distinctions, identifying relevant information and evidence	Text comprehension
	Interpreting graphic & visual information
inference, extrapolation, synthesis of information, and constructing an argument	Text comprehension
	Scrambled text / organisation of text

A third element of evidence for the technical homogeneity of TALL can be found in a Rasch analysis (Linacre, 2018) that determines the degree of fit of each item with others in the test. Misfitting "items degrade the quality of our measurement" (McNamara, Knoch & Fan, 2019, p. 47), threatening its integrity. So we may therefore consider values either of infit or outfit, depending on whether there is more than predictable variation, or less. A generally agreed measure in this regard is the Infit mean square (Infit MNSQ), an average calculation of fit across all items. In Table 2, a truncated version (not showing all the items that fit) is given. Ptmeasure-Al corr 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." It is similar to the Rit measures already referred to above.

Table 2

Misfit Order: Items in TALL 2023

Item	Total count (n)	Infit MNSQ	Ptmeasure-Al Corr	Expected
34	2125	**1.74**	-.20	.44
37	2125	1.08	.21	.31
12	2125	**1.26**	.20	.42
25	2125	**1.30**	.20	.45
39	2125	1.18	.26	.42
28	2125	**1.25**	.24	.45
29	2125	1.23	.25	.44
Further better fitting items not shown				
60	2125	.80	.60	.44
53	2125	.79	.59	.41
59	2125	.79	.61	.44
61	2125	.79	.63	.45
57	2125	.78	.62	.44
58	2125	.77	.63	.44
62	2125	**.74**	.65	.44

The mean infit MNSQ of all items in the test is 1.00. The two terminal values for Infit MNSQ are at 1.74 and 0.74. Linacre (2018, pp. 341, 354) indicates that when the benchmark for the measure of average fit (as in the calculated Infit mean square value or Infit MNSQ) is 1.0, as it is here, for individual items only the "expected values … greater than 1.5 [may be] problematic". McNamara, Knoch and Fan (2019, p. 45) and Van der Walt and Steyn (2007), however, suggest even more conservative limits, in the range of between 0.75 and 1.3. If this is adopted, then the two terminal items in terms of Infit MNSQ, 34 and 62, as well as the others in bold in Table 2 (12, 25, 28), may be identified. Item 34 has already been flagged in the discussion of the first warrant above. However, these very few potentially problematic items in a test of 64 items do not seriously detract from the technical integrity of the test, as is also shown in other analyses. As regards the modified cloze procedure, the last subtest, only two items (6 and 14) are outside the more conservative parameters, at Infit MNSQ respectively of 1.42 and 0.73. Once again, they are close enough not to be problematic.

When one considers evidence for Claim 2, several observations are relevant. In keeping with its demonstrated reliability over many years and administrations of its various versions, TALL and its subtests are also highly reliable. In the Rasch analysis of the data, three kinds of reliability are calculated: a person reliability, a test reliability and an item reliability. Two of these are related to the Classical Test Theory (CTT) analyses discussed in what follows, for which McNamara, Knoch and Fan (2019, p. 52) consider 0.8 as an acceptable level. In respect of test reliability, TALL expectably scored the same 0.93 as it did in the coefficient alpha measure of CTT, which is similar to that (McNamara, Knoch & Fan, 2019, p. 51). As regards the estimate of item reliability across the test as a whole, that has no equivalent in CTT, and whose value should also be higher than 0.8, TALL achieved a remarkable 1.00, and on person

reliability 0.91. For the subtest Grammar and text relations, Rasch readings of 0.93 for reliability, 0.8 on person reliability, and a very high 0.99 on item reliability are given.

In Table 3, the coefficient alpha (Cronbach alpha) of the 2023 administration (n = 2137) of the test and each of its subtests are set out. Overall, the technical consistency measured by this index is 0.93, indicating a high level of reliability. As regards the subtests, the best performing ones in respect of technical reliability are Scrambled text (where five sentences of a paragraph have to be arranged in their initial, unscrambled sequence) at 0.90; Text comprehension at 0.77; Register and text type (where two groups of five sentences - 10 sentences taken from five different text types - need to be paired) at 0.75; and, remarkably, Grammar and text relations (the modified cloze) at the same reliability level (0.93) as the test overall. The relatively high figure for the technical reliability of the Text comprehension subtest can be partly explained by its length (24 items). For a shorter test, one would of course need a set of subtests with fewer rather than more items. It needs to be noted that in other versions of TALL, Academic vocabulary regularly fares much better than in this administration of TALL, but its shortness, along with its adjusted coefficient alpha value of 0.86 (for a 40-item instead of a 9-item length test) shows that even in this administration, it demonstrates potential for inclusion in a shorter version of the test.

Table 3

Test-Subtest Correlations, and Subtest Inter-Correlations: TALL 2023

Subtest		**Test**	**1**	**2**	**3**	**4**	**5**	**6**
Scrambled text	*1*	0.60						
Interpreting graphic information	*2*	0.49	0.27					
Text comprehension	*3*	0.83	0.44	0.43				
Academic vocabulary	*4*	0.68	0.38	0.31	0.54			
Register & text type	*5*	0.61	0.25	0.21	0.41	0.38		
Grammar & text relations	*6*	0.85	0.36	0.31	0.52	0.45	0.46	
Number of testees		2137	2137	2137	2137	2137	2137	2137
Number of items		64	5	5	24	9	5	16
Average test score		41.84	2.91	4.39	15.58	6.65	2.89	9.41
Average P-value		65.4	58.1	87.9	64.9	73.9	57.8	58.9
Standard deviation		12.20	2.04	0.92	4.32	1.79	1.70	5.52
SEM		3.30	0.51	0.60	1.84	1.15	0.73	1.42
Coefficient Alpha		0.93	0.90	0.53	0.77	0.59	0.75	0.93
Adjusted 40-item length alpha		0.89	0.99	0.90	0.85	0.86	0.96	0.97
Greatest Lower Bound (if available)		-	0.94	0.57	0.83	-	0.83	-

If one discards the possibility of including the longest subtest (Text comprehension), then, as candidates for a shorter test emerge Grammar and text relations, Scrambled text, Register and text type, and perhaps Academic vocabulary. Grammar and text relations stands out in all respects: not only does it have the same technical consistency (0.93) as the test overall, but its 40-item length consistency (0.97) is even higher than that of the test as a whole. It is also almost 7 points more difficult than the test on average, which gives a further indication of its potential

as an initial screening test, intended to deselect from the test population those who are already capable.

When one examines the possible warrants for Claim 3, concerning the extent of technical differentiation of the test, the answers can again be found in the descriptive statistics summarized in Table 3. Except for subtest 2 (Interpreting graphic information) all of the test-subtest correlations are above the desired 0.6. Even more remarkably, all of the subtest intercorrelations are also within the parameters set, 0.2 for the lowest, and 0.5 for the highest. The same candidates for inclusion in a shorter test as were identified in examining Claim 2 are again prominent, except that now Academic vocabulary, based on words from Coxhead's (2000) Academic Word List, and at a subtest-test correlation of 0.68, is clearly higher than Register and text type.

The question set in Claim 4, whether the test has a theoretically defensible construct, has to some extent already been answered, in the discussion of the second set of evidence for Claim 1. The construct is based on ideas of communicative competence dating back more than 50 years (Habermas, 1970; Hymes 1971; Halliday, 1978; Halliday & Hassan, 1976), but that have stood the test of time (Halliday & Webster, 2002, 2003). First adapted for language assessment in South Africa by Yeld (2001; cf. too Cliff, Yeld & Hanslo, 2006; Cliff, 2014, 2015), its definition of academic language ability as an educational and scholarly language interaction has been refined (Van Dyk & Weideman, 2004a), and reconsidered, re-examined and augmented (Patterson & Weideman 2013a). This latest definition has been operationalized for use in at least two tests of academic literacy for postgraduate students. One, developed by Keyser (2017) is called TAGNaS (Toets van Akademiese Geletterdheid vir Nagraadse Studente, the Afrikaans version of TALPS, the Test of Academic Literacy for Postgraduate Students; see Pot, 2013; Pot & Weideman, 2015). The second, the Assessment of Preparedness Present Multimodal Information (APPMI) was developed by Drennan (2019). More than 100 scholarly articles in accredited journals, dissertations, theses and book chapters are listed in the Bibliography of the Network of Expertise in Language Assessment (NExLA) (2023), dealing mainly with TALL and similar tests, and attesting to their robustness and the diligence with which they have been scrutinized. In fact, Read (2016) judges that the careful attention to construct definition and refinement is what makes TALL and its associated tests noteworthy internationally (Weideman, Patterson & Pot, 2016, and the papers collected in Weideman, Read & Du Plessis, 2021).

As to Claim 5, the analyses above give an adequate basis for selection, as we shall note again in the next section, where the productive applications of these analyses are discussed. Not only are the various shorter subtests noted above eminent candidates for inclusion in a shorter format of the test, but their prime representative, the modified cloze procedure subtest, has a correlation of 0.85 with the test overall.

Discussion and Application

From the analyses presented above, it is clear that the modified cloze procedure employed in TALL, its Grammar and text relations subtest, has a high level of reliability for a 16-item test (a coefficient alpha of 0.93; a person reliability of 0.8, and an item reliability of 0.99), an outstanding correlation (0.85) with the test overall, and, from the point of view of being used as part of an initial screening test, an equally desirable difficulty level (at 59%), well below the

P-value (Table 3) of 65% of the whole test. Its high degree of correlation with the test total is not the only indication of technical quality: that is further enhanced by its low to moderate intercorrelation with the other subtests (ranging from 0.31 to 0.52). It appears, therefore, both to be contributing to the test as a whole, and to be functioning as an important individual component.

Similar considerations indicate that other subtests may also become candidates for inclusion in a shorter test, notably (if one disregards the longer, 24-item Text comprehension subtest) Scrambled text, Academic vocabulary, and Register and text type. The fact that Academic vocabulary correlates higher than the other two with the whole test (0.68) may indicate that it would not be wrong to consider it for inclusion in a shorter test; its adjusted (for 40-item length) coefficient alpha of 0.86 confirms this.

How one might apply this knowledge in test design depends on one's aims. Either way, the result is likely, in light of the analyses presented in the previous section, to be a shorter test, and thus the realization of the principle of efficiency, frugality or technical economy. The objective on which this contribution is focused is primarily the role that such a modification of cloze may be able to play as a component of an initial test for the purposes of screening. There are two examples of where it has already been successfully incorporated into the design of a test of academic literacy.

The first example is ALLT, the Academic Literacy Levels Test developed for the University of Southern Queensland (Green, Davis, Judith, Harmes & Weideman, in press). Here, the test designers have settled, for similar reasons to those enumerated above, on a three-tier test. At the first level (Tier 1), students are allowed to volunteer to complete a shorter, 30-item test, which has three types of subtests as components (Table 4).

Table 4

Specifications for ALLT: Tier 1

Tier 1	
Subtest	*Marks*
Scrambled text 1	4
Scrambled text 2	4
Vocabulary (one word)	6
Vocabulary (two word)	4
Grammar & text relations	12
Total	30
Duration	30 minutes
Results in two categories:	
Below 80% - Take Tier 2 test	
Above 80% - No need to take any further test	

Should students choose not to do the Tier 1 test, or if they do, but score less than 80% (a cut-off point that must still be adjusted on the basis of empirical evidence once the ALLT has been administered more regularly), they are required to proceed to do the Tier 2 test (Table 5).

Table 5

Specifications for ALLT: Tier 2

Tier 2	
Subtest	*Marks*
Scrambled text	5
Vocabulary (one word)	8
Vocabulary (two word)	6
Interpreting graphic & visual information	10
Register & text type	5
Text comprehension	30
Grammar & text relations (2+2+12)	16
	80
Duration	75 minutes
Results in risk bands:	
Category 1: Very high risk	
Category 2: High risk	
Category 3: Borderline - Write Tier 3 test	
Category 4: Less risk	
Category 5: Little to no risk	

The 2+2+12 specification for the modified cloze signals a scaffolding that was part of the evolution of this subtest, not discussed above, but re-employed here. Where the latest incarnation of the subtest described in the discussion of its evolution above merely has two examples, the designers of ALLT have adopted a format in which, apart from examples, in the first two questions test takers are required to indicate only the place where a word is missing. In the next two, they have to choose from among four distractors only the word that is missing in the places marked. Finally, in the 12 remaining questions of this 16-mark subtest, they have to indicate both the location of the gap and select the missing word.

The risk categories in Table 5 relate to risk associated with the ability to meet the demands of academic language at university level. Like the 80% cut-off point for the Tier 1 test, they need to be decided upon by considering empirical evidence gathered from further analyses done on subsequent administrations of ALLT. The third tier of this test is designed as a second-chance test for those who have been potentially misclassified as a result of measurement error. These calculations are a standard part of descriptive statistics yielded by programs like TiaPlus. For the administration of TALL which is being used as an example here, the misclassifications were as set out in Table 6; Rxx refers to correlating test scores with parallel test scores, and Rxt to correlating (observed) test scores with true scores (CITO, 2013, p. 19). As an illustration, should one allow for a second-chance test in this instance, one would calculate that in the worst case, 34 of the 2137 candidates were potentially misclassified. Since we may assume that there is an even chance of being misclassified to one's detriment or to one's advantage, another test opportunity can be offered to the first 17 test takers immediately below the cut-off point for Category 4 ("Less risk associated with language") in Table 5.

Table 6

Misclassifications in TALL (2023)

	Alpha based		**GLB based**	
Rxx case	Percentage	0.7	Percentage	1.6
	Number	14	Number	34
Rxt case				
	Percentage	0.5	Percentage	1.6
	Number	10	Number	34
90% Confidence limits for Coefficient Alpha: (0.92 =< 0.93 =< 0.93)				

For those who scored in this calculated borderline range, ALLT thereupon provides a second-chance test in its third tier, specified in Table 7. As is indicated under the notes to the specifications, such a test could also be offered to those who were ill or otherwise incapacitated when the Tier 2 test was administered. The test consists of a writing assignment, requiring an argument to be made to answer a question related to the topic of the (theme-based) test, using the by now familiar texts of the Tier 2 test, and adding one more.

Table 7

Specifications for ALLT: Tier 3

Tier 3	
Subtest	*Marks*
Making academic arguments (450 words, three paragraphs) [Add additional text to prior texts]	
	20
Duration	60 minutes

For second/third chance attempts for those with Category 3 results or incapacitated test takers of Tier 1 & 2 tests

Since the number of test takers should be much smaller for Tier 3, the writing assignment can be hand-marked. Not much is lost in terms of technical economy, since the first two tiers have acted as major selection instruments.

The second example of a test that incorporates the modification of cloze task under discussion is TAGNaS, the Afrikaans postgraduate-level test of academic literacy (Keyser, 2017). Its design echoes the considerations already articulated above, though it is more likely, once it has moved into a further phase of piloting and refinement, that it will become a two-tier test, since, it is argued, writing proficiency cannot be omitted from the test that forms the centerpiece at this level without losing face validity.

In sum, both the analyses given above, involving the substantiation of the claims by the warrants presented, and experience in various applications of TALL and associated tests lead us to conclude that a modified cloze procedure of the format discussed here is a viable and potentially productive component of a test that conforms to the norm of technical economy.

Conclusion

The productivity and robustness of the multiple-choice format cloze procedure discussed in this contribution is a clear illustration not only of technical frugality, but also of how various design principles interact and are interdependent. The notion of technical economy serves as an idea, a lodestar, setting the direction for and anticipating a saving of time and resources

through a design. To achieve such technical economy, it needs a second set of design principles, technical building blocks, as it were, to be realized and given shape. Without technical homogeneity, technical consistency, technical effectiveness, technical differentiation, technical appeal (or "face validity), or the theoretical defensibility of the design ("construct validity"), we shall not achieve the saving in resources and technical effort we desire.

One of the benefits of cloze procedure mentioned by Brown (1980) is that it is easy to produce. The format discussed here may require a little more effort, but its effects show the real range of its benefits: ease of scoring, high reliability, good correlation with other tests and the test as a whole, and a productive component of a longer or a shorter test. This constitutes a whole series of gains. No doubt, the production of this kind of subtest will become even easier when we are able to fully exploit the technological means that are now becoming available, in the form of ChatGPT and similar algorithms. Once we add in the savings of time and effort when computer adaptive language testing platforms become more accessible, the technical gains might be even more substantial. In the meantime, however, even if and where it still needs to be produced manually, it will serve well for exactly the purposes we have described in this paper.

ORCID

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

https://orcid.org/0000-0002-0303-5669

Acknowledgements

Not applicable.

Funding

Not applicable.

Ethics Declarations

Competing Interests

No, there are no conflicting interests.

References

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *The Modern Language Journal 64*(3), 311–317. https://10.1111/j.1540-4781.1980.tb05198.x

Brown, J. D. (2014). The future of World Englishes in language testing. *Language Assessment Quarterly 11*(1), 5–26. https://10.1080/15434303.2013.869817

Brown, J. D. (2021). Problems caused by ignoring descriptive statistics in language testing. In B. Lanteigne, C. Coombe, & J. D. Brown. (Eds.), *Challenges in language testing around the world* (pp. 15–24). Springer. https://doi.org/10.1007/978-981-33-4232-3_2

Brown, J. D. & Gruter, T. (2022).The same cloze for all occasions? Using the Brown (1980) cloze test for measuring proficiency in SLA research. *International Review of Applied Linguistics 60*(3), 599–624. https://doi.org/10.1515/iral-2019-0026

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

Cliff, A. (2014). Entry-level students' reading abilities and what these abilities might mean for academic readiness. *Language Matters*, *45*(3), 313–324.

Cliff, A. (2015). The national benchmark test in academic literacy: How might it be used to support teaching in higher education? *Language Matters*, *46*(1), 3–21.

Cliff, A., Yeld, N. & Hanslo, M. (2006). *Assessing the academic literacy skills of entry-level students, using the Placement Test in English for Educational Purposes (PTEEP).* Mimeographed MS.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly 34*(2), 213–238. https://doi.org/10.2307/3587951

Drennan, L. (2019). *Defensibility and accountability: Developing a theoretically justifiable academic writing intervention for students at tertiary level* [Doctoral dissertation, University of the Free State]. KovsieScholar. https://hdl.handle.net/11660/10888

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. (in press). Using a five-phase applied linguistics design to develop a contextualized academic literacy placement test for pre-university pathway students. *Literacy Research and Instruction*.

Habermas, J. (1970). Toward a theory of communicative competence. In H. P. Dreitzel, (Ed.). *Recent sociology 2 (*pp. 41–58). Collier-Macmillan.

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning.* Edward Arnold.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Halliday, M. A. K., & Webster, J. J. (Ed.). (2002). *Linguistic studies of text and discourse* (Collected works of M. A. K. Halliday, Vol. 2). Continuum.

Halliday, M. A. K. & Webster, J. J. (Ed.) (2003). *On language and linguistics* (Collected works of M. A. K. Halliday, Vol. 3). Continuum.

Hymes, D. (1971). On communicative competence. In J. B. Pride, & J. Holmes (Eds.), (1972), *Sociolinguistics: Selected readings* (pp. 269–293). Penguin.

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]. KovsieScholar. http://hdl.handle.net/11660/7704.

Language Courses and Tests (LCaT). (2023). Sample test. https://lcat.design/academic-literacy/sample-test-of-academic-literacy-levels/

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

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice and language assessment: The role of measurement*. Oxford University Press.

Network of Expertise in Language Assessment (NExLA). (2023). Bibliography. https://nexla.org.za/research-on-language-assessment/

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

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

Pot, A. (2013). Diagnosing academic language ability: An analysis of TALPS [Unpublished MA dissertation]. Rijksuniversiteit Groningen.

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

Purpura, J, Brown J. D., & Schoonen, R. (2015). Improving the validity of quantitative measures in applied linguistics research. *Language Learning*, *65*(Suppl. 1), 37–75. https://doi.org/10.1111/lang.12112

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

Read, J. (Ed.). (2016). *Post-admission language assessment of university students*. Springer Publishing International.

Schuurman, E. (2009). *Technology and the future: A philosophica1 challenge* (H. D. Morton, translator). Paideia Press. (Original work published 1972 as *Techniek en Toekomst: Confrontatie met Wijsgerige Beschouwingen*, Van Gorcum).

Schuurman, E. (2022). *Transformation of the technological society*, Dordt Press.

Trace, J. (2020). Clozing the gap: How far do cloze items measure? *Language Testing 37*(2), 235–253. https://10.1177/0265532219888617

Trace, J., Brown, J. D., Janssen, G., & Kozhevnikoval, L. (2017). Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. *Language Testing 34*(2), 151–174. https://10.1177/0265532215623581

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

Van Dyk, T. (2022). Die toegepaste taalkunde: 'n Oorsig. In W. A. M. Carstens & T. van Dyk (Eds.), *Toegepaste taalkunde in Afrikaans* (pp. 3–20). Van Schaik.

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

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

Van Riessen, H. (1949). *Filosofie en techniek.* J.H. Kok.

Vulcanization (2010). In *Wikipedia*. Retrieved July 26, 2010 from http://en.wikipedia.org/wiki/Vulcanization#Goodyear.27s_contribution.

Weideman, A. (2017a). Does responsibility encompass ethicality and accountability in language assessment? In C. D. Leymarie, & S. B. Makoni (Eds.), *Breaking down barriers in applied linguistics: Studies in honour of Alan Davies (1931–2015)* [Special issue]. *Language & Communication 57*, 5–13. http://dx.doi.org/10.1016/j.langcom.2016.12.004

Weideman, A. (2017b). *Responsible design in applied linguistics: Theory and practice*, Springer International Publishing. https://doi.org/10.1007/978-3-319-41731-8

Weideman, A. (2018). *Academic literacy: Five new tests.* Geronimo Distribution.

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. (2022a). Is die toegepaste taalkunde 'n onderdeel van die linguistiek? In W. A. M. Carstens & T. Van Dyk (Eds.), *Toegepaste taalkunde in Afrikaans* (Hoofstuk 2, pp. 21–36). Van Schaik.

Weideman, A. (2022b). Context, construct, and validation: A perspective from South Africa. *Language Assessment Quarterly 19*(2), 124–141. https://doi.org/10.1080/15434303.2020.1860991

Weideman, A. (2024). 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* (pp. 220–234). Routledge. https://doi.org/10.4324/9781003384922 -12

Weideman, A., Patterson, R., & Pot, A. (2016). Construct refinement in tests of academic literacy. In J. Read (Ed.), *Post-admission language assessment of university students* (pp. 179–196). Springer Publishing International. https://10.1007/978-3-319-39192-2

Weideman, A., Read, J., & Du Plessis, T. (Eds.). (2021). *Assessing academic literacy in a multilingual society: Transition and transformation* (New perspectives on language and education, 84). Multilingual Matters. https://doi.org/10.21832/WEIDEM6201

Yeld, N. 2001. Equity, assessment, and language of learning: key issues for higher education selection and access in South Africa [Unpublished Doctoral dissertation]. University of Cape Town.