Tobie van Dyk and Albert Weideman
Unit for Language Skills Development
University of Pretoria

Finding the right measure: from blueprint to specification to item type

ABSTRACT One of the important challenges of test design and construction is to align the blueprint of a test and the specifications that flow from it with the task types that are selected to measure the language ability described in the blueprint. This article examines a number of such task types and their alignment with the blueprint of a particular test of academic literacy. In particular, we consider a reconceptualisation of one traditional task type that has been utilised in some pilot tests. Its modification, problems, and potential value and future application are examined.

Keywords: language testing; test construction; test specifications; task types

Context

In a previous paper (Van Dyk & Weideman 2003), we have set out the rationale for employing a new construct for the test of academic literacy levels (TALL) that we are designing for the University of Pretoria. In that paper, we not only describe the evolution of the new construct, but also argue for the use of a streamlined version of the blueprint that would both facilitate its design and overcome the various logistical constraints that are associated with the administration of such a test to large numbers of students within a limited time. The purpose of the test is not to exclude students from study opportunities — they have already gained entry to the university by the time the test is taken — but to determine whether their level of academic literacy puts them at risk in their studies. Should the test indicate that they have even moderate risk, they are required to follow a prescribed set of courses aimed at improving their academic literacy. The test is taken by more than 6000 students on our campus, and by proportionately similar numbers at some medium-sized institutions of higher education that also use it, and there is great pressure for its marks to be available within 24 to 48 hours of being written.

The size of the testing populations and the urgency of making the results of the test known in themselves point to the necessity of employing a multiple choice format (that can be marked electronically), and of exploiting to the full other economical means that may be available.

The necessity of using multiple choice questions has forced us, incidentally, to abandon the knee-jerk, normally negative reaction that language educators have with regard to this kind of testing. In being forced to conceive of ways of employing this format, we have become more inventive and creative than we would otherwise have been, if we had simply succumbed to the prejudice that "one cannot test (this or that aspect of) language in this way". The ingenuity of some of the questions has demonstrated to us that this bias is often exactly that: a prejudice that prevents us from seeking creative solutions. Take, as an example, the following question (based on a reading passage) that was employed to test understanding of metaphor:

> We should understand the phrase "milk in their blood" in the first sentence to mean that both men
> (a) have rare blood diseases inherited from their parents
> (b) are soft-spoken, mild-mannered young farmers
> (c) don't like to make profit at the expense of others
> (d) are descended from a long line of dairy farmers

The same applies to the following question, that tests one's understanding of idiom:

> Paragraph 2 speaks of 'hatching a plan'. Normally, we would think of the thing that is hatched as
> (a) a door
> (b) a loft
> (c) a car
> (d) an egg

Or consider this one, which is designed to test the knowledge of the candidate regarding what counts as evidence:

> In the second paragraph, we read that "milk farms have been the backbone of this country" for centuries. Which sentence in the fourth paragraph provides evidence of the claim that it has been so 'for centuries'?
> (a) The first sentence
> (b) The second sentence
> (c) The third sentence
> (d) None of these

As may already be evident from these few examples, the tasks that the test takers are being asked to perform belong to a set of abilities or task types that are much broader in scope than that of a test that defines academic literacy in terms of skills, or reduces it to the mastery of sound, form and meaning.

There are several reasons, also dealt with in the previous article, for adopting a less restrictive, more open view of language as the basis for the current test. One is that these tasks are closer to the kinds of language performance that are required of students in academic institutions. In fact, in presenting the following framework of the new test to colleagues in seminars and conferences, we have been impressed with the degree to which it resonates with the experience of other academics. The construct (Weideman, 2003: xi) attempts to describe the components, or abilities, of being academically literate, and requires that students should be able to:

- understand a range of academic vocabulary in context;

- interpret and use metaphor and idiom, and perceive connotation, word play and ambiguity;
- understand relations between different parts of a text, be aware of the logical development of (an academic) text, via introductions to conclusions, and know how to use language that serves to make the different parts of a text hang together;
- interpret different kinds of text type (genre), and show sensitivity for the meaning that they convey, and the audience that they are aimed at;
- interpret, use and produce information presented in graphic or visual format;
- make distinctions between essential and non-essential information, fact and opinion, propositions and arguments; distinguish between cause and effect, classify, categorise and handle data that make comparisons;
- see sequence and order, do simple numerical estimations and computations that are relevant to academic information, that allow comparisons to be made, and can be applied for the purposes of an argument;
- know what counts as evidence for an argument, extrapolate from information by making inferences, and apply the information or its implications to other cases than the one at hand;
- understand the communicative function of various ways of expression in academic language (such as defining, providing examples, arguing); and
- make meaning (e.g. of an academic text) beyond the level of the sentence.

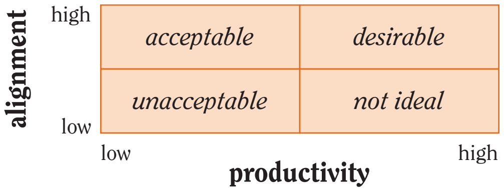
Test specifications

A first major challenge to test developers is aligning the construct of a test with the specifications that flow from this framework. Securing this alignment is a significant initial step in ensuring that the test indeed measures what it sets out to measure, i.e. is valid. Specifications can be drawn up in terms of a predetermined format, conceived of in terms that satisfy a range of formal requirements. For example, one may wish to specify, with a view to articulating the construct validity of a test, first the construct or blueprint from which the item type or the item itself is derived. One may add to this the logistical conditions under which the task type or individual item must be administered (pencil and paper mode; electronic format; the need for a sound system or time measurement device), the format of the kind of item, or an exemplification of this format, and so forth. For examples of such specification formats, there is a detailed exposition in Davidson & Lynch's (2002: 14 et passim) *Testcraft*.

For the sake of brevity, however, and since this is not the main focus of the present discussion, we shall attend here to one aspect only of such item type specifications: that of how they align with a feature or component of the blueprint of our test of academic literacy levels. We therefore tabulate below the sub-components of the test in such a way that they can immediately be related (or not) to the item types that test each language ability specification.

Test task types

The eventual selection of task types for a particular test depends not only on their alignment with the blueprint of the test and the specifications that flow from it, but also on the judgment of the test developer as to the productivity of the item type. Thus, the test developer has to make a judgment, either in terms of experience and expectation, or based on some quantitative measure, or on both of these, in respect of each test task type. Broadly, test task types will be adjudged to fall into one of four categories, *acceptable* (a high degree of alignment with the



test construct, but apparently not productive), *unacceptable* (low productivity coupled with small degree of alignment with blueprint), *desirable* (high alignment with construct, as well as productive), or *not ideal* (potentially productive, but not quite aligned with framework):

The judgment that a task type is not ideal does not mean that it should immediately be rejected; there may be subsequent opportunities of reconceptualising and redesigning these kinds of task to align them more closely with a test construct. Similarly, a task type that is evaluated as acceptable may simply need to be made more productive through progressive trialling in order to make it desirable, i.e. both aligned with the construct and productive.

In view of these distinctions, one may now consider the task types that we have already experimented with in developing our test of academic literacy:

- Scrambled text (a paragraph whose sentences have been jumbled)
- Register and text type (a task that requires the candidate to match the sentences in one list with those in another; usually from different registers)
- Vocabulary knowledge (within the context of a short sentence)
- Dictionary definitions (a derivative of the vocabulary knowledge type, but based on real dictionary definitions of terms taken from an academic word list such as that of Coxhead [2000])
- Error identification (the identification of especially South Africanisms)
- Interpreting and understanding visual and graphic information (questions on graphs; simple numerical computations)
- Longer reading passages (testing a wide variety of academic language abilities)
- Academic writing tasks (particularly ones requiring a measure of classification, and that comparisons be made)
- Cloze procedure (the restoration of a text that has been systematically mutilated)
- C-procedure (a derivative of cloze, in which half, or just more than half, of every second word is deleted)

Of these, only one, error identification, falls into the 'unacceptable' quadrant, since South Africanisms are probably more problematic in the context of general language proficiency than specifically in academic language ability. The others all have some measure of positive alignment with the test construct, and at least possess the potential to become productive items through a process of trialling.

When we compare the remaining item types here (i.e., the ones we have been experimenting with minus the unacceptable) with the original sub-components of the blueprint, the following picture emerges:

SpecificationTask type(s)	
Vocabulary comprehension	Vocabulary knowledge Dictionary definitions Cloze C-procedure

Understanding metaphor & idiom	Longer reading passages
Textuality (cohesion and grammar)	Scrambled text Cloze C-procedure (perhaps) Register and text type Longer reading passages Academic writing tasks
Understanding text type (genre)	Register and text type Interpreting and understanding visual & graphic information Scrambled text Cloze procedure Longer reading passages Academic writing tasks (possibly also) C-procedure
Understanding visual & graphic information	Interpreting and understanding visual & graphic information (potentially:) Longer reading passages
Distinguishing essential/non-essential	Longer reading passages Interpreting and understanding visual & graphic information Academic writing tasks
Numerical computation	Interpreting and understanding visual & graphic information Longer reading passages
Extrapolation and application	Longer reading passages Academic writing tasks (Interpreting and understanding visual & graphic information)
Communicative function	Longer reading passages (possibly also:) Cloze, Scrambled text
Making meaning beyond the sentence	Longer reading passages Register and text type Scrambled text Interpreting and understanding visual & graphic information

It is evident already from this rudimentary classification and comparison of specifications with task types that some task types at least potentially measure academic literacy more productively than others. Why should one then rather not employ those that appear to satisfy more specifications than others? The answer is complex, and opens up an old debate that concerns the quest for a single measure, or at least a very few measures, of language ability, such as in the historical debate on so-called integrative forms of testing (such as cloze, and possibly dictation) versus tests of discrete abilities. The stock answer to this, no doubt, is that while a limited number of sub-tests may be ideal from a test developer's point of view, it is generally so that the larger the number of sub-tests, the greater the chances are that the instrument will test reliably. For us, however, the reasons for not focussing exclusively on the most productive tasks type(s) are more practical: while it is evident, for example, that a longer reading passage might yield more, and a greater range of information on a candidateís level of academic literacy, it is also the case that it takes longer for candidates to complete questions that are based on an extended text. In a test that is severely constrained by time, one has to explore other, less time consuming types of task in addition to longer reading passages. Similarly, practical and logistical constraints dictate that, while we may include an academic writing task in a test, its marking must be handled creatively. For example, if we find that the results for the various task types that can be adequately tested in multiple-choice format consistently correlate well (i.e. above 0.85) with the total marks of a test that contains both multiple choice items and others, we may have grounds for deciding to leave the marking of the academic writing task(s) for later, when decisions on borderline cases need to be made.

What we also leave undiscussed for the moment in answering the question relating to the use of more productive types of task, is the relative weight of each specification and, by implication, the test task types that give expression to a specification. We may, for example, wish to assume that in academic language proficiency the understanding of vocabulary generally deserves much greater weight in a test than, say, the understanding of how metaphor and idiom are used by academic authors. Or we might wish to emphasise the importance of distinction-making, which lies at the heart of the academic enterprise, at the expense of a demonstration of the ability to extrapolate, on the assumption that the former ability precedes the latter in the development and socialisation of a new entrant into the academic world (and new entrants indeed provide the level at which our academic literacy test is pitched).

What is clear from the provisional alignment above of test task type and test specification is that we need to explore a number of further test task types to satisfy all of the specifications, and that we certainly need to consider making some of the existing test task types more productive. The remainder of this article will focus on how we have developed a format for making one of these task types, cloze procedure, more productive.

A closer examination of a specific task type: text editing

Our interest in this section is on cloze procedure. According to the task type-specification alignment table in the previous section, cloze is a potentially productive task type because it may test vocabulary, grammar and cohesion, an understanding of text type, and possibly also communicative function. Initial trials that we did with multiple choice cloze (i.e. with a given choice of words that might fill the gaps), however, were disappointing. For example, of the five questions in this task type in a pilot test we ran for 1111 candidates in March 2003, only one made the cut, i.e. had a discrimination index of greater than 0.3, and fitted into the acceptable facility value range we had set, of between 0.3 and 0.7. Only when we relaxed the acceptability criteria to a discrimination index upwards of 0.26 and a facility value of between 0.26 and 0.84, did three of the five items qualify.

Obviously, these values could potentially be improved by, for example, modifying the distractors. But in the meantime, through experience that we had gained in constructing another test, we began to wonder whether we could not also adjust the format of the question. We subsequently designed another task for piloting that not only omitted every 7th word, but also asked of candidates to identify the place where the word has been omitted. Candidates are therefore required to select, from the list given, the word that has been omitted, and to indicate its original place. Here is an example:

Text	Word	Place
In $_{1\&2\ (a)}$/ this $_{(b)}$/ act, $_{(c)}$/ the $_{(d)}$/ context otherwise indicates, $_{3\&4(a)}$ / 'candidate $_{(b)}$/ means any $_{(c)}$/ person $_{(d)}$ bound to serve $_{5\&6(a)}$/articles $_{(b)}$/of $_{(c)}$/ clerkship $_{(d)}$/or to $_{7\&8\ (a)}$/ perform $_{(b)}$/community $_{(c)}$ / under a contract $_{(d)}$ / of service.	1. (a) if (b) when (c) that **(d) unless** 3. (a) lawyer **(b) attorney** (c) conveyancer (d) paralegal 5. (a) with **(b) under** (c) his (d) regular 7. (a) regularly (b) serious **(c) service** (d) work	2. (a)/ (b)/ **(c)**/ (d)/ 4. (a)/ **(b)**/ (c)/ (d)/ 6. **(a)**/ (b)/ (c)/ (d)/ 8. (a)/ (b)/ **(c)**/ (d)/
It $_{9\&10(a)}$/ that $_{(b)}$ / irregular $_{(c)}$/service $_{(d)}$/ as a $_{11\&12(a)}$/ candidate $_{(b)}$/, within $_{(c)}$/ the $_{(d)}$/ meaning of $_{13\&14(a)}$/ the $_{(b)}$/above, $_{(c)}$/ is $_{(d)}$/ irregular service $_{15\&16(a)}$/ under articles $_{(b)}$/ or a $_{(c)}$/ of $_{(d)}$/ service as defined.	9. (a) is (b) concludes (c) regulates (d) follows 11. (a) paralegal (b) conveyancer (c) lawyer (d) attorney 13. (a) mentioned (b) definition (c) construction (d) regulation 15. (a) contract (b) definition (c) way (d) mode	10. (a)/ (b)/ (c)/ (d)/ 12. (a)/ (b)/ (c)/ (d)/ 14. (a)/ (b)/ (c)/ (d)/ 16. (a)/ (b)/ (c)/ (d)/

... The correct answers to the first four (questions 1–8) have been done for you, and are marked in bold in the second and third columns:

The results of this 50-item test were encouraging from a number of angles. In the first place, the test developer is always on the lookout for sub-tests that have a reliability measure (*alpha*) that is higher than that of a whole test, since this is an indication that, in general, this particular kind of test task discriminates not only more reliably, but better and more efficiently than other types of task. This original 50-item pilot on a test group of above average language ability first year students (n = 43) indeed proved slightly (0.8%) more reliable than a previous pilot that also contained 50 items, but in that case spread over seven item or task types.

By eliminating some of the items within the text that did not measure within the desired parameters (i.e. either too low a discrimination value, or too easy or difficult), and by modifying some of the distractors among the remaining items that did not appear to work efficiently, we then came up with a second version of the pilot. Though it used essentially the same text, this version contained only 30 items. It was again trialled on two groups of first year students, the one (n = 21) with lower levels of academic literacy according to another test of academic language proficiency that they had written at the beginning of their studies, the other (n = 23) with higher levels of academic language proficiency according to the same test.

These results were even more encouraging. Here is a comparison of the mean score, reliability (*alpha*), and average facility value of items (mean percentage correct) of this second version of the test for the two groups. The last column (mean item-total correlation) gives an indication of how, on average, items in the tests discriminate:

Version	Ability	Mean	Alpha	Mean % correct	Mean item-total
First	above average	64.418%	0.749	64	0.232
Second	above average	66.956%	0.817	67	0.361
Second	below average	58.570%	0.810	59	0.325

As is evident, the improvement of the second version on the first is marked (just in terms of reliability: 9% and 8%, respectively). These results therefore certainly seem to indicate that this particular modification of cloze procedure has promise, and that its productivity and efficiency needs to be exploited further. In fact, in subsequent tests of this type, that were embedded within a longer test of academic literacy for a large group of below average ability students, the results remained equally encouraging:

Version	Ability	Mean	Alpha	Mean % correct	Mean item-total
Third (English)	below average	48.341%	0.86	448	0.353
Fourth (Afrikaans)	below average	70.191%	0.82	870	0.331

When the test was administered for the first time at the beginning of 2004, the results once again were excellent. This time the test was on a group of 6427 students of average ability:

Version	Ability	Mean	Alpha	Mean % correct	Mean item-total
English	average	61.943%	0.92	462	0.656
Afrikaans	average	58.193%	0.87	658	0.532

There are, however, some problems that prevent us from exploiting this format fully. We turn to these in the concluding section.

Remaining problems

As with any experimental format, this modified form of cloze procedure still has a number of difficulties that need ironing out before one can recommend it for wider use. The description of these difficulties here is, however, an invitation to others to participate in helping to find solutions to the difficulties that we still encounter with this productive item type. Two unanswered questions that remain, simply because we have not yet had the time to experiment more widely, are how to select the right text for our audience, and how to mutilate the text (deleting every 7th word, or less or more frequently). More frequent deletion would, in our judgement, probably make a text too difficult to restore for new entrants into the academic world. Less frequent deletion would solve a number of sheer practical problems (fitting both text and questions into three narrow columns). We would welcome advice on this.

The most vexing problem, however, is how to write the rubric so that it is clear and intelligible. Acting on advice from some of the invigilators in our first two rounds of pilots, we have now, in a third pilot with another (probably easier) text, changed the second and third columns of the original around, and re-written the instruction, as follows:

> ***Text editing***
>
> Some words are missing from the text below. In this task, you have to restore the text in the first column to its original form. To do so, you first select a **place** [marked (a)/, (b)/, (c)/ or (d)/ in the text, with the question numbers in the second column] where you think the word is missing in the text. Then you select a **word** [(a), (b), (c) or (d)] from those in the third column to complete the text. Mark the correct answers on your answer sheet. The *unevenly* numbered questions (65, 67, 69, 71, etc.) are for the **place**, the *evenly* numbered ones (66, 68, 70, 72, etc.) for the **word** you have selected to fill in.
>
> The correct answers to the first three have been done for you, and are marked in bold in the second and third columns.

We are not entirely convinced that the advice we took (of switching the columns) was the best: candidates still find this too difficult, and too many in the third round of piloting asked for an explanation. If they were more familiar with the format, they would probably not have needed an additional explanation. In the meantime, however, we still need to devise a way, either by rewriting the rubric in more intelligible language, or by making another plan altogether, of really making a promising task type work. In that way, a type of task that started out, in terms of our matrix above, as 'not ideal', can gravitate first towards ‘acceptable’ and, finally, the ‘desirable’ part of the quadrant.

Conclusion

Very little is written in South Africa about test production. We believe that this is not a healthy situation: in being accountable as test constructors, we need more discussion and narrative (cf. Shohamy 2001) about how tests are made, and what kind of decisions are taken, especially when we move from blueprint to specification to test type. This article constitutes one attempt to stimulate such discussion.

REFERENCES

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

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

Brindley, G. 2002. Issues in language assessment. Pp. 459–470 in R.B. Kaplan (ed.), 2002. *The Oxford handbook of applied linguistics*. Oxford: Oxford University Press.

Coxhead, A. 2000. A new academic word list. *TESOL Quarterly* 34(2): 213–238.

Davidson, F. & Lynch, B.K. 2002. *Testcraft*. New Haven: Yale University Press.

Shohamy, E. 2001. *The power of tests: a critical perspective on the uses of language tests*. Harlow: Pearson Education.

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

Weideman, A.J. 2003. *Academic literacy: prepare to learn*. Pretoria: Van Schaik.

Albert Weideman is director of the Unit for Language Skills Development of the University of Pretoria. His primary responsibility lies in developing courses, and his *Academic literacy: prepare to learn* was published in 2003. Other published courses include Starting English (for beginners) and *Making Certain* (for advanced learners).

Prof. Albert Weideman
Director
Unit for Language Skills Development
University of Pretoria
0002 Pretoria
Tel.: 012 420 4957
Fax: 012 420 2000
e-mail: albert.weideman@up.ac.za

Tobie van Dyk is a lecturer in the Unit for Language Skills Development. He is the project manager of the test of academic literacy levels (TALL) that has been developed within the unit, and is now used by his own university, as well as by the Universities of Potchefstroom and Stellenbosch. His particular interests lie in language assessment and testing.

Tobie van Dyk
Unit for Language Skills Development
University of Pretoria
0002 Pretoria
Tel.: 012 420 4834
Fax: 012 420 2000
e-mail: tobie.vandyk@up.ac.za