Albert Weideman

University of South Africa

Assessment literacy and the good language teacher: insights and applications

Abstract

There is currently a great deal of interest in language teachers' competence in assessing language ability. Their competence in this regard, or lack of it, has much to do with their initial training and professional biases. Taking as example the teaching and learning of one specific kind of language, academic discourse, this paper discusses a number of assessment techniques that language teachers could apply to language teaching at school, or in other contexts of language tuition. Its basis is four basic principles of language assessment: reliability, validity, interpretability of results, and efficiency. These four principles are important to all assessments designed by language teachers. Some assessment techniques that have not yet widely been used to their full potential by teachers are described when different formats of language assessment are discussed. In particular, examples of effective and efficient formats of assessment will be given by referring to an analysis of a test of academic literacy administered to senior secondary school students in their pre-university year. Those examples have clear applications in other language learning settings. The paper concludes with a challenge to teachers: to experiment with new assessment designs, and to learn how to interpret the results of assessment in order to plan language instruction more effectively.

Keywords: assessment literacy; language testing; academic literacy; design principles; language ability; reliability; validity; diagnostic information; efficiency

***Journal for Language Teaching*, Volume 53, Number 1, pp. 103 – 121. 2019.**
ISSN 0259-9570. http://dx.doi.org/10.4314/jlt.v53i1.5

Are good language teachers competent in assessment?

The international literature is literally abuzz with the idea of "assessment literacy" and how that ability affects language teaching (Wylie & Lyon, 2017), sometimes in highly localized contexts (e.g. Semiz & Odabaş, 2016; Sellan, 2017). Assessment literacy is usually defined as language teachers' awareness and knowledge of assessment. Another related contemporary theme is classroom-based assessment. In short: there is more current awareness about assessment among language teachers and those who train them than ever before. The current interest in the level of knowledge of language teachers about how language can be assessed most effectively is evident not only in the growing literature on it (Taylor, 2009; Fulcher, 2012), but also in the prominence it has in discussions at professional gatherings of language testing specialists. The annual conference of the International Language Testing Association (ILTA), the Language Testing Research Colloquium (LTRC) held in Bogota in 2017, for example, had "Language assessment literacy across stakeholder boundaries" as its main theme, and at LTRC 2018 in Auckland several papers and contributions dealing with assessment literacy took that discussion further. Though language teachers' knowledge of testing practice is not the only dimension of assessment literacy, since it applies equally to the knowledge of assessment held by users of test scores in a number of institutional environments (Taylor, 2009), it no doubt constitutes an important component of the ability to use tests responsibly. It is usually argued in the literature being referred to here that such competence will enable language teachers to become more accountable for the ways in which they design language assessments, and that they will gain professionally by becoming assessment literate.

Despite the international attention focussing on the assessment literacy of practising language teachers, very little thought has gone into this in South Africa. The question for us is therefore: Where does one begin in examining the levels of assessment literacy of teachers? And, once examined and determined, how would language teachers achieve higher levels of assessment literacy? At this early stage in our awareness of assessment literacy as a global issue, rather than coming up with a final methodology of how it might be probed (for, as can be expected, there are many different ways to go about this), this paper focuses instead on a number of essential assessment design principles that might in our context underlie any eventual evaluation of the levels of assessment literacy of language teachers. The paper will attempt to demonstrate that language teachers will gain professionally if they start by checking whether their current ways of assessing language ability conform to four prominent principles of language test design: test reliability, test validity (cf. Weideman, 2019a), the interpretability of results, and test efficiency, though there are many more (Weideman, 2019b). That means, first, that they will have to invest in considering potentially effective but underutilized language assessment techniques, which I shall return to when discussing the implications of the case study presented below. Second, such professional gain will contribute to their perspective on language assessment not being a mere classroom or curricular routine, but an accountable process, in which the responsible assessment of language ability – the kind of assessment that checks and considers whether it responds or conforms to certain principles of assessment constitutes a key element. In short, becoming more assessment literate can no doubt be associated with what is referred to in the title as the ‘good’ language teacher.

Language teachers’ current awareness of and ability to assess language competence both professionally and responsibly in the first instance have to do with what was emphasized during their training and development. One of the regrets of my professional life as a language teacher is that I realized too late how my students might have benefited if I was skilled in language assessment. My own training emphasized being ‘learner-centred’ as the prime consideration. There is a good reason for this emphasis: language teachers began to acknowledge at that time that teaching language does not automatically convert into learning another language. The task of the teacher, it was accepted, is to make learning possible. My regret is this: if I had known more about assessing language ability, I would have been much better equipped to create the conditions for language learning in the highly charged context of the classroom – highly charged because we have plenty of examples of language learning happening successfully outside the classroom, in environments with less stress, less anxiety, and less tension.

So teachers who were imbued, as I was, with the language teaching orthodoxy of the last two decades of the 20th century, the communicative approach to language teaching, often failed to attend adequately to language assessment. There is no space here to deal with the considerable literature on communicative language testing, but we should note that teachers’ attempts to implement communicative language teaching involved creating a language classroom without stress and anxiety. The creation of those non-threatening conditions seemed to preclude any overt evaluation of their learners’ ability. The dilemma in testing communicatively was: How does one evaluate without creating more tension?

Language assessment is therefore tied up with one’s approach to teaching. The approach adopted justifies not only the desired style of teaching (Weideman, 2002), but also influences how language ability must be assessed. If the approach calls for a stress-free language learning environment, one may be tempted to diminish the importance of evaluating performance. By steering away from, or reluctantly assessing their learners’ ability, however, teachers deprive themselves of a valuable source of information for their subsequent teaching. Add to this that pre-service language teacher training by all accounts pays inadequate attention to assessment techniques (Taylor 2009), and you have a recipe for neglect, and, in my case, regret.

This paper will consider one approach to teaching language for a specific purpose, and examine how that approach affects, and is aligned with, assessment. I shall take as an illustration the teaching and testing of academic literacy. Academic literacy is the ability that we desire students to possess when they intend to enrol at tertiary institutions, in order to handle the language demands of university or higher education. It must be remembered that not all language teachers teach languages at public schools; there is a world of private language tuition, with a wide variety of purposes and aims, outside of the school system. Of course, where teachers do teach language at primary or secondary school level, they teach it as a subject. That might tempt one to think that language teaching at school can be conceived of as 'achievement' testing, and contrast that with 'proficiency' testing, provided, of course, that one is able to uphold that conventional distinction. As soon as we examine the language syllabi (Department of Basic Education, 2011), however, we note that language ability ('proficiency') is the primary aim of the instruction (Du Plessis, Steyn & Weideman, 2016). What is more, though academic literacy is not a school subject, the curricula being referred to here require secondary school learners to become proficient in language used for education and learning (for a more complete treatment of the implications of this, see below, section 3, and Myburgh-Smit, 2015). So in its aims, language teaching at school looks beyond school, to the ability to handle language that will be used, for example, for study, in the workplace, and for being a responsible citizen. The teaching, the learning and the assessment of academic literacy therefore potentially hold a number of valuable insights for language teaching in general. Academic literacy is a communicative ability, because in educational institutions a student interacts with others through language in order to understand, develop, and produce analytically-characterized discourse, usually related to academic argument. Though academic discourse is but one type of language (Patterson & Weideman, 2013), and is acknowledged, also in the Grade 12 syllabus, to encompass an advanced level of language ability (Du Plessis, Steyn & Weideman, 2016), there will be obvious lessons for language teachers to learn from this example, and applications to be made to other (intermediate and beginner) levels of language teaching, inside and outside of the school system.

Below I shall therefore deal, first, with a new way of looking at language and assessing it. Though, as has already been noted, there are many more, four basic principles of language assessment will be highlighted: reliability, validity, the interpretability of results, and efficiency. These four principles should be important specifically to the assessments designed by the good language teacher, if we view such teachers as ones that can justify the ways in which they design their assessments of language ability by responding to the principles of language test design. Phrased differently: the good language teacher will treat assessment design as a process that needs to be accounted for. When the formats that language assessment might take are discussed below, some assessment techniques that have not yet widely been used to their full potential by teachers will be described. Those examples have applications in other language learning settings. The paper will conclude with a challenge to teachers: to experiment with new assessment designs, using information that these yield in order to plan instruction more productively.

Why good language teachers would want to become good assessors of language ability

Language teachers, as well as administrators who use test scores, are concerned with professional and responsible assessment practices, as defined in the previous section, not only for the sake of becoming accountable for the decisions that are taken on the basis of the scores they award to their students, but also for a number of further reasons.

The first of these is that good language teachers should wish to be competent in assessing language ability in order to have a measure of whether the language instruction that they provide has been successful. Without the measurement of outcomes, there can be no basis for claims that they have been achieved. Such a process speaks to the design principle of effectiveness. Second, if they had some reliable indication of the level of mastery of the language by their students, they would have a potentially trustworthy measure of those levels. Third, such a reliable measure would enable them to be better attuned to the further language development needs of their students. A reliable measure would allow them to identify the gaps in their students' language ability. Their measurement, and the interpretation of its results, would therefore yield diagnostic information: the scores may provide information essential for further instruction. If they carefully designed their assessment instruments to measure exactly the ability they were wanting to enable their students to develop – what is called 'validity' in testing jargon (Weideman, 2019a) – the results of the assessment will be more easily interpretable. Finally, if they could measure language ability in an efficient manner, it might eliminate some of the drudgery associated with 'marking'.

The example referred to below will provide evidence of how these four basic principles of language testing (reliability, validity, interpretability and efficiency) can be employed to make assessment of language more consistent, more effective, more useful, and less of a chore.

Method, instructional context of the assessment, description of population, and sampling

The instructional setting for language instruction in the case study described below is one in which the language teacher was involved by invitation, to make up for a shortcoming in the language development of senior secondary school learners. The learners in this case are South African high school students, of about 16-17 years of age, who are in their final year of school, about to write their final school exit examinations. The results of these examinations determine to a large extent whether these senior pupils will be allowed entry into tertiary study at a university. The trouble that they face is that the universities they apply to now have an additional requirement: in many cases applicants have to demonstrate their ability to use language for academic purposes at university level by writing tests of academic literacy in the year before entering university. The universities may use the results for placement of students on academic literacy development courses, but increasingly, since more and more school leavers are competing for a limited number of places at university, they are also using them for access, to determine whether students will be allowed to enter certain courses, especially ones that are in high demand (Myburgh-Smit, 2015; Sebolai, 2016).

Why (with English now the dominant language of higher education) does these students' instruction in English-as-subject at secondary school not adequately prepare them for using language at university level? The reasons are complex (Du Plessis, Steyn & Weideman, 2016), but the short answer is: university authorities no longer trust the deteriorating results of the school exit examinations. What is more, language instruction at school has demonstrably been drifting away from the stipulations of the national syllabus (Weideman, Du Plessis & Steyn, 2017), so that it in fact does not emphasise, as required, mastering language for the purposes of higher education. Language instruction neglects language for academic purposes, though the syllabus requires substantial attention to its development. One reason for this neglect is that the examination papers of previous years set the tone (a phenomenon called 'washback'), not the syllabus. Teachers merely want to see their pupils gaining good marks in the exit examinations, and the exit examinations make little provision for the assessment of the high level of language ability required by the curriculum, which includes the assessment of academic literacy.

It is in this context, then, that supplementary language teaching is called for. Additional language instruction is offered to minimize the risk for prospective entrants into higher education failing to obtain a mark on an academic literacy test that will allow them entry either into university, or into certain highly sought-after courses. The universities who require it have a point: they have seen a trend, following the massification of higher education globally since the mid-1990s, that for them establishes a link between success at university and academic language ability (Van Rensburg & Weideman, 2002; Van Rooy & Coetzee-Van Rooy, 2015). By their reckoning, lower levels of academic literacy carry a risk for university students not achieving the pass rates required for receiving government subsidies, still by far the main source of income for universities. They do not wish to place their income at risk, especially when the risk is not of their making, but lies elsewhere in the education system.

The population of this study is made up of 105 senior secondary school pupils that were conveniently sampled in a series of language development workshops at several schools in 2017. The pupils whose assessment results are being used were fully informed of the aim of the test, and an undertaking was given that their results, when analysed, would be comprehensively anonymized. In addition, both the location and the number of the schools involved have been withheld.

Satisfying the principle of validity by first defining the ability to use language for education and study.

Though, as we have noted in the previous section, the school syllabus (Department of Education, 2011: 4, 9) provides amply for instruction to gain "access to higher education", and has specific stipulations, for example, for the mastery of advanced vocabulary, making inferences, doing critical analysis, identifying main and peripheral issues, categorizing, sequencing, recognizing connections between texts, and many similar functions of language that are associated with the mastery of language for academic purposes, there is almost no evidence in the final assessment of students that this ability is either taught or assessed (Weideman, Du Plessis & Steyn, 2017). Yet these are important components of language of which to have mastery, if we look at the substantial literature on academic literacy (see NExLA, 2019), as this ability is referred to. A widely-used definition of academic literacy (Patterson & Weideman, 2013) emphasizes the analytically stamped nature of academic discourse, that has already been referred to above. Its various elements can be summarized in a table (Table 1), that in the first column identifies the component of academic literacy that should be taught and assessed, and in the second column gives examples of the possible task types (for teaching) or the subtests that will allow components of academic literacy to be assessed (Weideman, 2017)

Table 1: The construct of academic literacy and its operationalizing

Understand / interpret / have knowledge of	Task type / Subtest
vocabulary and metaphor	Academic vocabulary (one word) Academic vocabulary (two word) Text comprehension (in larger context) Text editing Grammar & text relations (modified cloze)
complex grammar, and text relations	Grammar & text relations (cloze) Scrambled text / organisation of text Text editing Making academic arguments
communicative function	Understanding text type and 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 Organising information visually
essential/non-essential information, sequence and numerical distinctions, identifying relevant information and evidence	Text comprehension Interpreting graphic & visual information Making academic arguments
inference, extrapolation, synthesis of information, and constructing an argument	Making academic arguments Text comprehension Scrambled text / organisation of text Writing task

There are more subtests that can be associated with the components listed here, so those listed in the second column constitute only a provisional set. It is important to note that while the components of academic literacy in the first column can be assessed by means of a range of subtests, the subtests, in turn, potentially can assess more than one component. It is important, furthermore, to note that the components of academic literacy listed in the first column define that specific language ability functionally. That is a new perspective on language: it asks what one needs to be able to do with and through language. It is different from the traditional way of defining it as being made up of sounds, vocabulary, and grammar, or as the skills of listening, speaking, reading and writing (Weideman, 2017).

Having now defined what is being measured (the ‘construct’ of academic language ability), and broken it down into components, we have taken the first step towards fulfilling the requirement of test validity: we have a theoretically defensible idea of what we are measuring (Read, 2010: 288); defensible, too, because it is a current rather than an outdated one. The problematic truth about language teaching at senior secondary school level in South Africa, as we have shown above, is that it does not adequately measure the ability to use language for education and study. The observation here is that a definition of that ability, already articulated to a substantial extent in the curriculum prescriptions for language teaching at school, would be a good starting point. Once we have deliberately defined the ability we wish to assess, we have also taken the next step towards making the results of the assessment interpretable and meaningful; if the test measures effectively, we may be able to see whether our students lack mastery of one or more functional components of academic interaction through language, for example of seeing relations between different parts of a text, or of making inferences, and so on.

The application of this knowledge should already be apparent: if we can find a test that assesses the ability to use appropriate vocabulary, or the competence to extrapolate, or to making meaningful connections, or to measure sensitivity to genre, or perhaps to do all of these, we would have a measure of language ability that is strongly related to what we have to be assessing in many other language classrooms as well. Before returning to these possible applications, let me first present, as an example of such a test, the measuring instrument that was used in the case being described and analysed.

The measuring instrument: a multi-component, and potentially comprehensive assessment

The learners in this example whose academic literacy needs further development are in their final or pre-final year of high school. They need to be able to demonstrate to the universities they will be entering after finishing school that they are able to cope with the language demands of tertiary education. The first step towards the development of this ability to use language for a specific purpose is the assessment of their existing levels of academic literacy. In the relatively short period they have to prepare for an assessment by the universities the identification of their weaknesses and strengths would be ideal for designing the language instruction they need in order to develop their language ability.

To measure this ability, a theme-based test of academic literacy was used (taken from Weideman, 2018), that assesses as comprehensively as possible the components of academic literacy referred to in the previous section, and is made up of several of the subtests (sections) that measure them, which have been selected from those mentioned in the second column of Table 1:

Section 1: Scrambled text
This subtest scrambles the sentences of a paragraph, and requires the learner to unscramble them by asking which sentence should be placed first, second, third, and so on.

(5 marks)

Section 2: Vocabulary knowledge
Based on words taken from Coxhead's (2000) Academic Word List, this subtest assesses the learner's familiarity with words used frequently in academic language.

(10 marks)

Section 3: Verbal reasoning
This subtest gives the test taker the opportunity to demonstrate an ability to make inferences, extrapolate, and know what counts as evidence.

(5 marks)

Section 4: Interpreting graphic and visual information
The ability to handle data presented graphically or visually is tested, for example recognizing trends, or making proportional comparisons. It also tests the ability to handle different genres.

(10 marks)

Section 5: Register and text type
Five sentences/phrases from five different genres (a newspaper report; an advertisement; a set of instructions; a novel; a scholarly text) must be matched with five further sentences from the same sources.

(5 marks)

Section 6: Text comprehension
Unlike a conventional comprehension test, the questions are carefully designed to assess whether one is able to distinguish between the essential and the peripheral; to see connections among words, clauses, and paragraphs; to recognize sequence and order; to know how different communicative functions are used; to use metaphor and idiom in context, etc.

(45 marks)

Section 7: Grammar & text relations
This subtest is a modification of cloze procedure, where every fifth, seventh, or ninth word may be deleted. Test takers are required not only to fill in

the right word, but also to indicate where the gap is. The subtest tests grammatical awareness, vocabulary knowledge, use of prepositions, relations between different elements of text, and even communicative function.

(20 marks)

Here is an example of a part of this last kind of subtest:

In the following, you have to indicate the possible place where a
word may have been deleted, and which word belongs there.

> Goodyear claimed that he **81&82** $\boxed{\text{i}}$ discovered $\boxed{\text{ii}}$ vulcanization $\boxed{\text{iii}}$ 1839 $\boxed{\text{iv}}$ but did **83&84** $\boxed{\text{i}}$ not $\boxed{\text{ii}}$ patent $\boxed{\text{iii}}$ the $\boxed{\text{iv}}$ until June 15, 1844.

81. Where has the word been deleted?	**82. Which word has been left out here?**
A. At position (i).	A. first
B. At position (ii).	B. rubber
C. At position (iii).	C. year
D. At position (iv).	D. in

83. Where has the word been deleted?	**84. Which word has been left out here?**
A. At position (i).	A. then
B. At position (ii).	B. apparently
C. At position (iii).	C. invention
D. At position (iv).	D. fully

The full test, of 100 marks, reflects in the weighting of its subtests the judgement of the test designer regarding the relative importance of the different components of academic literacy that were listed in Table 1.

The test theme ("Rubber: an ordinary, everyday thing") is evident in all subtests. Having a theme-based test contributes to the sense that the test takers have of interacting in an academic fashion with a single, coherent issue, stimulating by its topicality their engagement with it. That in itself enhances another facet of the test, which is often called its "face validity".

Efficient assessment may spring from new test formats

Many teachers would, perhaps even regularly, employ the kinds of tests mentioned above, or variations of them. Yet some are certainly less familiar. For example, the last subtest, Grammar & text relations, that assesses high-level grammatical skill in addition to vocabulary, communicative function and cohesion, is a potentially productive kind of subtest that has been neglected by teachers. In the next section, that deals with the analysis of the results of the application of the instrument described in section 5, above, I shall present some statistics on just how well such a test performs, and why it should not be avoided.

The format in which the test is administered is also one that teachers may often neglect. All of the subtests in the test outlined above are in multiple-choice format, with four or five choices provided. Classroom teachers sometimes have predictable, yet largely unwarranted, biases against closed-ended instead of open-ended formats of assessment, and these prejudices are to some extent explicable, since answer-constrained formats, as their name implies, indeed limit the variety of possible responses. If no other formats for assessment are utilized, it would perhaps be a mistake to rely solely on this limited-answer format. But if one already has sufficiently emphasized other formats, such as assignment writing, putting together a portfolio of best work, awarding marks for classroom participation and for homework, and if one regularly utilizes a range of open-ended student responses in several modes, there is no reason to steer clear entirely of a multiple-choice format. This format has the advantage of allowing one to re-use items that have worked well, thus saving future assessment design time. It has demonstrable efficiency gains. Incidentally, it is also demonstrably more reliable than even strictly moderated forms of hand-marked open-response assessments. Moreover, its marking is much easier, and, if the answers are completed on an optical reader form, its results can be scanned and captured on a spreadsheet to facilitate further analysis. It not only saves on the drudgery of marking, but, if the test ranges over a multiplicity of components and subtests, and is long enough (usually more than 40 items in length), its overall result will closely correlate highly with reliably scored, open-ended assessments that require much more effort and time to administer and score. Finally, the format is easily adaptable to computer-based testing, that is becoming ever more available also in schools, and it is a pre-requisite for computer-adaptive testing (Read, 2010: 293), where, by using a limited number of pre-tested items, the test taker's ability level can efficiently be obtained. The application here of the principle of efficiency is evident.

Adopting such new or underutilized formats therefore brings the language teacher to consider to yet another test design principle, that of efficient measurement. That kind of consideration is aimed at removing drudgery from language testing, and freeing up instructional energy that can usefully be employed towards realising the real goal of language teaching: the development of the learner's ability.

The principles of reliability and interpretability: What a rudimentary analysis of results yields

In this section, we show how the results of this wide-ranging test of academic literacy, the instrument described in section 5, were analysed (see Berg, Schaugency, Van der Meer & Smith, 2018). For such analysis there are freely downloadable programs for classroom teachers, that I would encourage teachers to explore. I have chosen to employ the freeware called TiaPlus (for Test and item analysis +) from http://

tiaplus.cito.nl/ (see Cito 2013), but there are numbers of others, e.g. jMetrik (https://itemanalysis.com/) (for a more comprehensive list, see Clauser & Hambleton 2018).

The TiaPlus analysis yields performance data both at test level, for the test as a whole, and at subtest and item level, as in Table 2.

Table 2: Subtest intercorrelations, test-subtest correlations, and basic properties of Rubber test

Subtest		Total test	1	2	3	4	5	6	7
Scrambled text	1	0.43							
Vocabulary	2	0.47	0.19						
Verbal reasoning	3	0.46	0.22	0.20					
Interpreting graphic & visual info	4	0.35	0.10	0.07	0.07				
Register and text type	5	0.49	0.19	0.23	0.16	0.20			
Text comprehension	6	0.90	0.37	0.39	0.39	0.24	0.43		
Grammar & text relations	7	0.71	0.09	0.18	0.22	0.14	0.18	0.43	
Number of testees		105							
Number of items		100	5	10	5	10	5	45	20
Average score		63.43	1.51	7.40	2.89	7.41	2.91	27.67	13.64
Average score (%)		63	30	74	58	74	58	61	68
Average Rit		0.28	0.83	0.34	0.44	0.40	0.65	0.81	0.87
Coefficient Alpha		0.88	0.79	0.26	-.04	0.41	0.49	0.81	0.87

The program has calculated familiar statistics, for example average scores and percentages for the test overall, as well as for the subtests. What we can learn from the average of 63% overall illustrates the working of the third important principle: interpretability. Was the test too easy since it has an average of above 50%? To know that, we need to interpret the scores with reference to other administrations of the same test. For example, when this test was administered several years earlier, first year students at a reputable South African university scored 68%. That means that the population whose results are being analysed here is actually not yet on par with what that university would have expected, despite what looks like a high average. We can only interpret a score properly by bringing into our calculation all the information we have at our disposal. Lesson? An average mark of 50% is meaningless as a kind of magical ‘pass’. Any mark needs interpretation.

We can also see from this analysis that the average Rit (a measure of how well an item discriminates between top-performers and those in the lowest overall quartile of marks; see Cito, 2013: 29) of the test, at 0.28, is well above 0.15, which is the lowest acceptable value. But when we look at the average Rit of the subtests, we notice that, apart from the longer (45 mark) subtest that predictably performed well, the much shorter subtest 7 (Grammar & text relations) fared remarkably better, outperforming the other subtests. What further counts in its favour is that, though it makes up only one fifth of the test (20 marks), its correlation with the overall test is excellent, at 0.71.

Our suspicion, then, that the Grammar & text relations subtest is one that performs well, can be tested further by examining the reliability of the test as a whole, and those of the various subtests. This is measured by looking at the Coefficient Alpha, sometimes called Cronbach’s Alpha (Cito, 2013: 29). Generally, for class tests, one would be looking for a reliability of above 0.6. For school examinations, we may perhaps find 0.7 acceptable. For higher stakes tests, for example those that grant access to further opportunity or work, one would certainly think that 0.8 is a minimum, and above 0.9 desirable. The fact that this particular test has an overall Alpha of 0.88 means that it is already highly reliable. But equally heartening is that subtest 7, with only 20 items, is once again the top performer, managing to score 0.87 on a very strict index. Compare that to the negative -0.04 of Subtest 3 (Verbal reasoning).

A more sophisticated analysis would have checked to see whether the subtest-intercorrelations were within the conventional parameters of 0.15 (a low correlation) and 0.5 (a moderate correlation), since we do not want the subcomponents of a test to correlate too highly: that would indicate that they are measuring the same component, and thus not measuring as effectively as they could. Several subtests here would therefore have drawn attention by scoring too low. As regards the subtest-test correlations, we would of course seek higher correlations. Predictably, Text comprehension, being the longest subtest, scores highest (0.90), but once again the Grammar & text relations subtest catches the eye: it has a correlation with the test overall of 0.71, the second highest of all the subtests. Its effectiveness is almost beyond doubt.

Some of the statistical measures above, like the reliability index Cronbach Alpha, directly gauge the reliability or consistency of the test, and so indicate wholly whether at test level, the language assessment instrument that was used conforms to the important principle of reliability. Others, that were not referred to in detail here, are measures that traditionally indicated whether the test satisfies the condition of effectiveness, or validity. But the further important principle of test design illustrated here is that these numbers, when interpreted, give us additional indications of how the test conforms to the principle of interpretability. That is not the only way in which a test like this gives meaningful results. There are more, and to that we turn in the next section.

Assessment that informs the design of instruction

What further lessons are there from these analyses? We can see that Subtest 3 (Verbal reasoning) is in need of repair. In fact, the overall results would have been more reliable if it had been omitted altogether. We can further observe that the longer the test is, the more reliable it is likely to be. And finally, we can see from the average marks that the students fared worst in the Scrambled text subtest (at 30% average). That means that their ability to see connections between different parts of a text is perhaps not on par.

That lack of ability may also be evident when one examines the statistics of some items that also measure this ability, in context, in Section 6 (Text comprehension) of the test. In the item statistics that TiaPlus generates, we see further evidence in the low percentage correct answers to questions that test this same subskill.

These numbers provide empirical grounds for the language teacher (in this case, where the instruction is aimed at the development of academic literacy) to emphasize those tasks that enable learners to practice sequencing of information, cohesive ties, and seeing relations between different parts of a text, either by designing appropriate tasks, or using ones from textbooks (cf. Weideman, 2018; 2007). In short: diagnostic information supports instructional design, and helps the teacher to identify what should be emphasized in subsequent language teaching. We find such information when we apply the assessment principle or interpretability.

Some further applications

As an application of the principle of validity – at its basis the idea that we should be measuring what we set out to measure, and do it in a theoretically justifiable way (Weideman 2019a) we have now considered how a new idea of language allows us to assess more responsibly. That new perspective on language as used in higher education settings – conceptualized as a means of communication in academic work – affords us the opportunity to view it functionally instead of conventionally. Perhaps teachers in secondary schools may have noticed that the ideas mentioned in this paper are not so novel: they are embodied in the very syllabi that they use in their everyday language teaching. Indeed, given the syllabus demands in the South African case (Department of Basic Education, 2011), one could claim that if language teachers gave more attention to developing academic language ability at school, they would align their language teaching much more effectively with those policy requirements.

Similarly, the kinds of assessment are not entirely unfamiliar, though some may have suffered neglect. An example of a neglected format is the multiple-choice, closed-ended one employed in the example we looked at. There may have been similar neglect in respect of task type: the modified cloze procedure in Subtest 7, that was described in section 5 above, provides an example. This is a highly productive test, in the sense of satisfying the principles of test design that have been the focus of the argument here: reliability, effectiveness, interpretability and efficiency. The empirical analyses of its answers have shown that even in a test this short, its results correlate well with the overall mark. I would encourage experimentation with it, since it may serve well as a quick and efficient – though not comprehensive – assessment of language ability.

As to the bias among teachers against assessing in multiple-choice format, I would urge them to have an open mind, to examine and test out its advantages. Given enough imagination and creativity, there is no reason why it cannot be used, say, to assess knowledge of metaphor and idiom, or inferencing. Compare the following example from the sample test:

67. We can infer from the phrase “gums… herbivorous insects” in paragraph six that

A. the insects eating plants producing rubber have gums but no teeth.

B. insects usually take longer to adapt to their circumstances than plants.

C. a good number of plants that have rubber use it to protect themselves.

D. to appreciate the congealment of rubber depends on one’s point of view.

The key to application of these ideas is imaginative design, and the application of the four principles of assessment design that have been discussed here.

Conclusion

This paper has examined some of the implications for South African language testers of the growing international attention to the assessment literacy of language teachers and, though I have focussed less on them, of the administrators who use the results of language assessment, for example to decide whether applicants are granted access to university. Rather than prescribing or fixing ways of examining levels of assessment literacy among these professionals, the paper has suggested that we first consider four prominent principles of language test design: reliability, effectiveness (validity), interpretability and efficiency. Those principles should form the basis, nonetheless, for the methodological means, such as questionnaires, interviews, focus group discussions, ethnographic investigations and the like, through which levels of assessment literacy can eventually be measured.

This paper has been therefore been able to deal with only a small slice of the issue. Its argument and illustrations have been offered to encourage language teachers to assess language ability more professionally and more responsibly, so that the scores that they give the language learners in their charge become more useful, and also more publicly defensible. Teachers are neither immune to the global requirements of increased accountability, nor should they be the unhappy victims of neglect or ignorance of new professional challenges.

To become a good assessor of language ability depends on staying informed of what is happening. There are many good introductions, and even excellent shorter briefings (such as Read, 2010). For analysis, there is sufficient software available, of which one example has been used here to demonstrate the useful interpretations that language teachers may derive from the scores of a well-designed test. In a time when electronic means are no longer foreign to professional language teaching, I would encourage language teachers to learn to use at least one of these statistical programs, and not to shy away from experimenting with new formats of language test design. In short: there is every reason to attempt experimentation and imaginative design, and, as I said at the beginning, much to gain.

References

Berg, D.A.G., Schaugency, E., Van der Meer, J. & Smith, J.K. 2018. Using Classical Test Theory in higher education. In Secolsky, C. & Denison, D.B. (Eds.) Handbook on measurement, assessment, and evaluation in higher education. pp. 178-190.

Cito. 2013. TiaPlus users manual. Arnhem: M & R Department, Cito. Available: http://tiaplus.cito.nl/. Accessed: 10 November 2017.

Clauser, J.C. & Hambleton, R.K. 2018. Item analysis for classroom assessments in higher education. In Secolsky, C. & Denison, D.B. (Eds.) Handbook on measurement, assessment, and evaluation in higher education. pp. 355-369.

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

Department of Basic Education. 2011. Curriculum and assessment policy statement: Grades 10-12 English Home Language. Pretoria: Department of Basic Education.

Du Plessis, C. Steyn, S. & Weideman, A. 2016. Die assessering van huistale in die Suid-Afrikaanse Nasionale Seniorsertifikaateksamen: die strewe na regverdigheid en groter geloofwaardigheid. LitNet Akademies 13(1): 425-443. Available: https://www.litnet.co.za/die-assessering-van-huistale-in-die-suid-afrikaanse-nasionale-seniorsertifikaateksamen-die-strewe-na-regverdigheid-en-groter-geloofwaardigheid/. Accessed: 8 Jan. 2019.

Fulcher, G. 2012. Assessment literacy for the language classroom. Language Assessment Quarterly 9(2), 113-132. DOI: 10.1080/15434303.2011.642041.

Myburgh-Smit, J. 2015. The assessment of academic literacy at pre-university level: a comparison of the utility of academic literacy tests and Grade 10 Home Language results. MA dissertation. University of the Free State.

NExLA (Network of Expertise in Language Assessment). 2019. Bibliography of language assessment. Available https://nexla.org.za/research-on-language-assessment/. Accessed 8 Jan. 2019.

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

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

Sebolai, K. 2016. The incremental validity of three tests of academic literacy in the context of a South African university of technology. PhD thesis, University of the Free State. Available: http://hdl.handle.net/11660/5408.

Secolsky, C. & Denison, D.B. (Eds.) 2018. Handbook on measurement, assessment, and evaluation in higher education. New York: Routledge.

Sellan, R. 2017. Developing assessment literacy in Singapore: How teachers broaden English language learning by expanding assessment constructs. Papers in Language Testing and Assessment 6(1): 64-87.

Semiz, Ö, and Odabaş, K. 2016. Turkish EFL teachers' familiarity of and perceived needs for language testing and assessment literacy. Proceedings of LILA '16: III. International Linguistics and Language Studies Conference, organized by DAKAM (Eastern Mediterranean Academic Research Center). pp. 66-72. Istanbul: Dakam Publishing.

Taylor, L. 2009. Developing assessment literacy. Annual Review of Applied Linguistics 29, 21-36.

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

Van Rooy, B., & Coetzee-Van Rooy, S. 2015. The language issue and academic performance at a South African University. Southern African Linguistics and Applied Language Studies: 33(1): 31-46, DOI: 10.2989/16073614.2015.1012691

Weideman, A. 2002. Designing language teaching: on becoming a reflective professional. Pretoria: BE at UP. Available: https://albertweideman.files.wordpress.com/2016/04/ebook_designing_language_teaching_by_albert_weideman.pdf. Accessed: 8 Jan. 2019.

Weideman, A. 2007. Academic literacy: Prepare to learn. Pretoria: Van Schaik.

Weideman, A. 2017. A skills-neutral approach to academic literacy assessment. Contribution to symposium on Assessing the academic literacy of university students through post-admission assessments, Language Testing Research Colloquium (LTRC) 2017, Bogota, Colombia.

Weideman, A. 2018. Academic literacy: five new tests. Bloemfontein: Geronimo Distribution.

Weideman, A. 2019a. Degrees of adequacy: the disclosure of levels of validity in language assessments. Koers 84(1). doi.org/10.19108/KOERS.84.1.2451.

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

Weideman, A., Du Plessis, C. & Steyn, S. 2017. Diversity, variation and fairness: Equivalence in national level language assessments. Literator 38(1): 9p. DOI: 10.4102/lit.v38i1.1319

Wylie, C. & Lyon, C. 2017. Supporting teacher assessment literacy: a proposed sequence of learning. Teachers College Record. Available: http://www.tcrecord.org. ID Number: 22193. Accessed: 14 November 2017.

ABOUT THE AUTHOR

Albert Weideman

University of South Africa
Email: albert@lcat.design

Albert Weideman is Professor of Applied Language Studies at the University of the Free State. He is the coordinator of the Network of Expertise in Language Assessment (NExLA). His research focus is on how language assessment, course design, and policy relate to a theory of applied linguistics.