Tydskrif vir Taalonderrig - Journal for Language Teaching
- Ijenali yokuFundisa iLimi - IJenali yokuFundisa iiLwimi -
Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta Polelo
- Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo - Ijenali
Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo
- Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig -
Journal for Language Teaching - Ijenali yokuFundisa iLimi
- IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi
- Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo -
Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi
- Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa
Ririmi - Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa
iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta
Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo -
Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo
- Jenala yo Dyondzisa Ririmi
- Tydskrif vir Taalonderrig
- Journal for Language
Teaching - Ijenali
yokuFundisa iLimi -
IJenali yokuFundisa
iiLwimi - Ibhuku
Lokufundisa Ulimi
- Tšenale ya tša
Go ruta Polelo -
Buka ya Thuto
ya Puo - Jenale
ya Thuto ya Dipuo
- Ijenali Yekufundzisa
Lulwimi - Jena?a ya u
Gudisa Nyambo - Jenala yo
Dyondzisa Ririmi - Tydskrif vir Taalonderrig
- Journal for Language Teaching - Ijenali
yokuFundisa iLimi - IJenali yokuFundisa iiLwimi -
Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta
Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya
Dipuo - Ijenali Yekufundzisa Lulwimi - Jena?a ya
u Gudisa Nyambo - Jenala yo Dyondzisa Ririmi
- Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali
yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi -
Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo -
Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi
- Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa
Ririmi - Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa
iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta
Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo -
Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo
- Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig -
Journal for Language Teaching - Ijenali yokuFundisa iLimi
- IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi
- Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo -
Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi
- Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa
Ririmi - - Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa
iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta

Laura Drennan

Michelle Joubert

Albert Weideman

and

Rohan Posthumus
University of the Free State

Combined measures of identifying language risk for first-year students

ABSTRACT

The disruption that accompanied the pandemic in 2020 and 2021 also affected the administration of academic literacy tests. These are employed to place incoming students at institutions of higher education in appropriate courses for the development of their ability to handle the demands of academic discourse. For many students the conventional tests, deployed online, were inaccessible. We reflect here on how possible alternatives might be employed to identify students who are at risk as a result of low levels of academic literacy. The first is an algorithm that aims to predict whether the student might be a candidate for an intensive academic literacy course, and the other, constituting the primary focus of this paper, is a conventional post-admission academic literacy test available in-house, which had the potential of being refined for such a purpose. Since the test had initially been designed to assess prospective postgraduate students' preparedness to engage in academic writing, and had been piloted on a range of undergraduate students, this presented an opportunity to explore whether it might be possible to use it more widely. Analyses generated by programs yielding both descriptive, inferential and probability statistics are presented to show that this test was indeed capable of being employed thus, and could be refined further. At the same time, this exploration has had the further benefit of enhancing the assessment literacy of those presenting the actual academic literacy interventions. We envisage a further exploration of adapting tests of similar design for assessments that are more field-specific.

Keywords, academic literacy; language assessment; academic risk; student placement; assessment literacy; language test refinement

***Journal for Language Teaching*, Volume 55, Number 2, pp. 93 – 115. 2021.**
ISSN 0259-9570. https://dx.doi.org/10.4314/jlt.v55i1.4

1. Taking lessons from disruption

South African universities use tests of academic literacy on a large scale, and this is matched by the critical analyses of their worth: the 'Bibliography' tab of the website of the Network of Expertise in Language Assessment (NExLA 2021) lists more than a 100 scholarly articles, master's level dissertations and doctoral theses, detailing the effort that has gone into this over the last two decades (for a comprehensive analysis, see Van Dyk 2020). The tests are designed to combat the risk, associated with language ability, to which first-year students are exposed. This is usually accomplished in one of two ways: to identify students whose academic literacy levels are too low for successful study before access is granted to study; or, when measured after entry, to place students on the appropriate academic literacy course. The latter are interventions designed to develop the ability to handle the demands of academic language (Weideman 2003). At many higher education institutions in South Africa, these interventions are mostly modular and train students in language (generally, English language associated with cognitive academic language proficiency as outlined by Cummins, 2000, 2001) as well as academic literacy skills. The testing of language and literacy levels then, is an important issue to engage with since it has the potential to impact what is covered in terms of language and literacy in these interventions. The particular contribution that South Africa has made in this regard (discussed in Weideman 2021) has focused in particular on the definition and articulation of the construct of academic literacy, conceived of as the ability to handle the demands of academic discourse (Patterson & Weideman 2013a, 2013b; Weideman, Patterson & Pot 2016).

South Africa is not unique in trying to counter the effects of students arriving at university unprepared to handle academic language by first testing their academic literacy. Especially in the format of post-admission assessments, these kinds of tests are used much more widely (Read 2015, 2016), and often in addition to other measures. In South Africa these may be the results of the Grade 12 school exit-level examinations; elsewhere this may also include standardized international language tests. It is significant, however, that in addition to the commercially available large-scale international language tests, the post-entry English language tests (PELA's) in Australia and New Zealand, for example, indicate that universities still find a need to develop and use their own measures of academic language ability. There can be many reasons for this, but the need in South Africa to find an in-house language assessment became especially apparent during the pandemic that has, since March 2020, disrupted the usual administration of nationally available tests.

In the current case, we report on the implementation of one alternative and the pilot of another alternative that were devised when, for many students, the conventional tests, became inaccessible. This was because these tests were deployed only in online format, requiring computer and electronic resources that were not always available. These tests require that students have access to specialized equipment, such as webcams, but the version of the academic literacy test we used could be made available online, though without the need for sophisticated equipment. Before articulating the specific research questions below, we may note that the broader research question is: How can one

identify students who need support in developing their academic literacy, when the tests to determine their levels become inaccessible? Of course this is too comprehensive an issue to tackle in an article of this nature, so we turn next to the specific issues that we were able to address, to contribute at least a partial understanding to the broader research question.

In particular, we shall further report here on two strategies that could be employed as alternatives for identifying students who are at risk as a result of levels of academic literacy known or presumed to be too low. We examine both in the analyses below. When both measures are taken into consideration, we might move towards a more comprehensive understanding of students' academic literacy and language performance. The first measure is an algorithm that aims to predict whether the student might be a candidate for an intensive academic literacy course, and the other a conventional post-admission academic literacy test available in-house, which could be refined for such a purpose. That test, which is the primary focus of this paper, had initially been designed to test prospective postgraduate students' preparedness to engage in academic writing (Drennan 2019, 2021), but, since it had been piloted on a range of undergraduate students, it presented an opportunity to explore whether it might not be possible to use it more widely. It could also be made available on QuestionMark, an in-house electronic platform for assessment. This mode of delivery did, however, present various challenges that need to be taken into consideration regarding future administrations of online literacy and language tests.

Finally, we shall consider the benefits of possibly employing a combination of these measures in future, by comparing the outcomes they yield, especially in making decisions about the placement of at-risk students on appropriate academic literacy courses.

2. Selection of instruments

A machine-learning algorithm

Rationale for selection

The first instrument selected was an algorithm that was developed to predict whether the student might be a candidate for an intensive academic literacy course. The concept of "justice as fairness" (Tjabane & Pillay 2011: 11; Deygers 2017, 2019; Deygers & Malone 2019; Deygers, Van den Branden & Van Gorp 2017), when applied to literacy and language testing, was brought to light when the COVID-19 pandemic struck in 2020. With much conventional testing going online that utilized proctoring software, one of the dilemmas faced by tertiary institutions involved making judgments about students' academic literacy and language levels when many students did not have the means (stable network, laptops, sufficient data) to be reliably tested on an online platform. This

was especially problematic where the post-admission tests of academic literacy and language which are used to place students into academic literacy development modules included test takers who are first-generation students, who are likely to come from lower socio-economic backgrounds (Cloete 2016: 6).

In an attempt to mitigate these social justice risks and provide students with an equal chance to be exempted from these developmental modules based on their putative ability and not their socio-economic background, a machine-learning algorithm was developed at the University of the Free State (UFS). More detailed information about the development of the algorithm will be published in a forthcoming paper, but in brief, the algorithm was developed using the National Senior Certificate (NSC) and National Benchmark Test (NBT) data of 27 528 UFS students from the years 2014 to 2020. The aim of the algorithm is to determine which variables, or combination of variables, predict performance on the academic literacy portion of the NBT.

Data-collection and analysis procedures

In order to create the machine learning algorithm, the historical NSC data set (from 2014 to 2020) of 27 528 UFS students was used as training data for the algorithm, as the NBT scores were available for these students. The key variables used included the school code (as a proxy for socio-economic status), and school mark-subject combination, as well as home language (see also Sebolai 2016). These variables were then used to predict literacy and language performance on the NBT. The dataset was processed, normalized and split into a training and testing dataset. The algorithm trained on one dataset and predicted the score of the literacy portion of the NBT test without 'knowing' the results of the testing dataset. The historical results were compared with the in-sample predicted results, and an accuracy score was calculated in a confusion matrix. The process is set out in Figure 1.

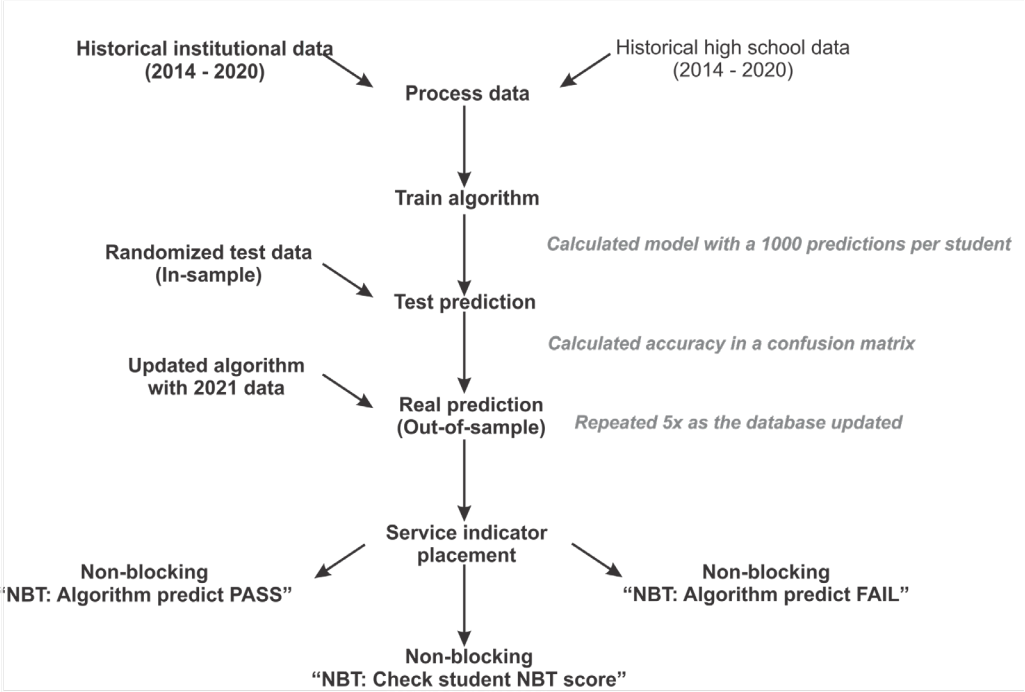

Figure 1*: Process of predicting placement on academic literacy courses*

Results

The accuracy score served to satisfy the probability of making a correct out-of-sample prediction. Figure 2 shows that the in-sample predictions were 85% accurate. Seventy percent (70%) of the in-sample were true negatives (correctly predicted) and 15% of the sample were true positives (correctly predicted). The remaining 15% were incorrect predictions. Students who managed to write the NBT in 2020 were excluded from the algorithm, as placement based on the NBT academic literacy results took precedence over the algorithm's predictions. The reason for this is that the NBT actually tests for academic literacy performance, while the algorithm tries to predict it. In 2021, the algorithm was applied to 7 370 mainstream programme students and, based on the resulting predictions, students were placed into (or exempted from) the developmental literacy modules.

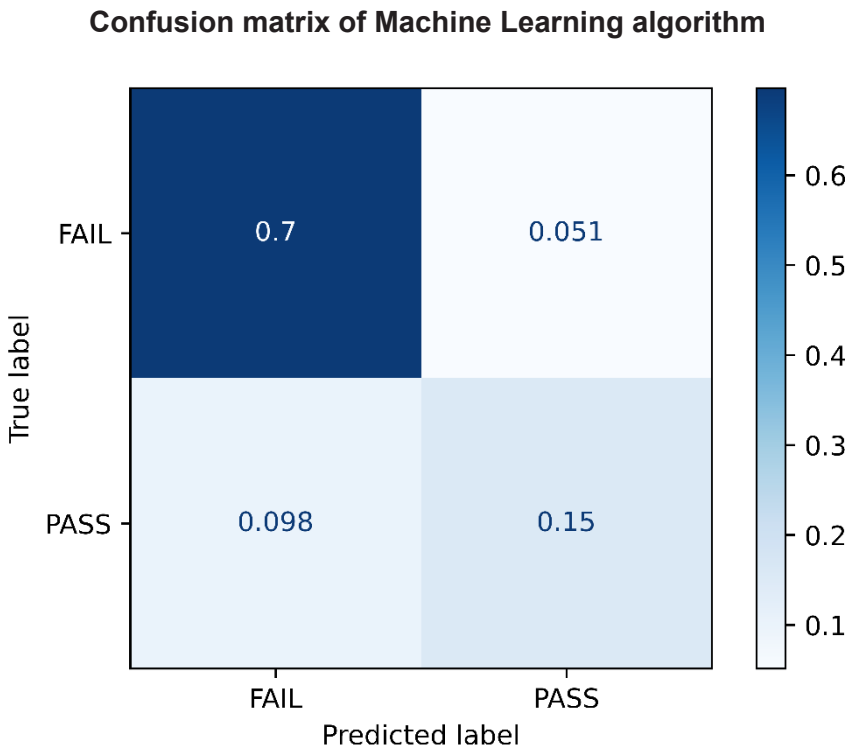

Figure 2: Confusion matrix (accuracy testing) for the machine learning algorithm

An Assessment of Preparedness to Present Multimodal Information (APPMI)

Rationale for selection

The second instrument which was piloted for potential use on incoming students was the Assessment of Preparedness to Present Multimodal Information APPMI test. There are several reasons for selecting this test which will be briefly mentioned here before a more in-depth discussion on the design and performance of the test is provided in the section that follows. The first of these reasons was that it employs a view of language ability that is functional instead of skills-based (Weideman 2020b). It furthermore utilizes the perspective that, in order to be able to 'write' or present academic information to a relevant audience, students need the functional abilities first to find academic information (by listening, reading, or other means) before processing such information (initially by taking notes, but also through categorizing and classification, comparing, sifting, tabulating, and discussion). All of that has to happen before presenting the new, processed academic information to others, and that presentation can take the form not

only of writing, but embrace several media or modalities: *viva voce* presentation with or without PowerPoints further discussions, written assignments and papers, and the like. It is obvious that when one takes a skills-neutral, functional view of the process of using academic language – as a process of first finding, then processing, and finally producing new information – the various 'skills' used are interdependent. One cannot, for example, conceive of the phases of this process without observing that it involves an intertwinement of what might be called 'listening', 'reading', 'speaking' and 'writing'. Also in the mix are cognitive 'skills' such as comparing, inferencing, and extrapolating. Yet the skills are so closely interwoven that an isolation of one or the other, in an attempt to assess such a skill, is just about impossible. So a skills-neutral test allows the test designer to probe the mastery by the student of functional and cognitive sub-abilities, rather than to focus on what are now generally admitted to be restrictive 'skills' (listening, speaking, reading, and writing) (Bachman & Palmer 1996: 75f.). We return to this point in the next section.

A second reason for selecting the APPMI as a possible stand-in assessment of academic literacy lay in the history of its development: it was piloted also on undergraduate students, and was found to discriminate well among them (Drennan 2019). Evidence of the degree of fit between the test takers' ability and the difficulty of the items can be gleaned from the results of Rasch and Classical Test Theory analyses, to be discussed in more detail hereafter.

There was also a third reason: the test was not monotone in its design, but a differentiated test, with various subtests that seemed to function well. Traditionally, the relations among various subtests in a test of language ability are viewed as indications of validity (Weideman 2019a, 2019b), but there is another way of looking at test-subtest correlations, which need to be above 0.6, and even 0.7. That would indicate that each subtest is functioning well, and doing its part to contribute to the overall measurement, more or less in line with the others. Similarly, the subtest inter-correlations have to fit into specified parameters: the inter-correlations of subtests preferably need to be lower, between 0.2 and 0.5, in order for the test to be technically viable, since that would indicate that each subtest is measuring something different. Weideman (2021: 11) therefore views these relations as organic rather than physical analogical concepts – those usually associated with validity – since they echo the functional contribution of each to a viable technical whole. This view will take alignment with these parameters as a demonstration of how the test fulfils the requirement of technical *differentiation across functionally different tasks*.

3. A skills-neutral approach to test design

The design of the APPMI is based on the premise that the production of information, in the form of writing or any other presentation of information to a particular audience, is preceded by the processes of gathering (selecting) and processing (organizing) information. Proponents of discourse synthesis (Spivey 2001; Spivey & King 1989) support

the notion that when a proficient reader/writer is tasked with producing information with a particular objective in mind, they engage in the higher-order processes of gathering relevant information from sources and organizing it to create links between ideas that develop and support this objective. This involves adapting their reading depending on their objective and their prior knowledge of the text structure conventions of various source text types. This knowledge of how discourse is usually organized in specific text types allows them to use criteria of importance to select information, understand from textual signals how ideas are linked in a text, and make inferences across texts (Frederiksen 1975; Spivey & King 1989; Van Dijk 1979). It is for this reason that the production of discourse and synthesis of information are thought to be closely linked to discourse comprehension.

Drennan (2019) discusses the cognitive phases associated with the processes of gathering and processing information, and the various subtests of the APPMI that were designed to measure the skills related to these processes. Table 1 illustrates the relationship between the cognitive phases, the subtests of the APPMI, and abilities required to handle the demands of academic discourse (i.e. the construct of academic literacy: Patterson & Weideman 2013a, 2013b).

Table 1*: Alignment of cognitive phases, APPMI subtests and construct (Drennan 2019, 2021)*

Cognitive phases	Sub-processes	APPMI subtests	Alignment with construct
Conceptual-ization	Task represen-tation Macro-plan-ning	Understanding text type and communicative function Making academic arguments Interpreting graphic and vi-sual information Text comprehen-sion	Communicative function Text type (including visual rep-resentations) Essential/non-essential infor-mation, sequence and numerical distinctions, identifying rel-evant info for evidence Employment and awareness of method Inference, extrapolation, syn-thesis of information, and con-struction of argument

Cognitive phases	Sub-processes	APPMI subtests	Alignment with construct
Meaning construction	Global careful reading Selecting relevant ideas Connecting ideas from multiple sources	Organizing information visually Understanding academic vocabulary Text comprehension Making academic arguments Organization of text/scrambled text	Vocabulary and metaphor Complex grammar and text relations Communicative function Text type (including visual representations) Essential/non-essential information, sequence and numerical distinctions, identifying relevant info for evidence Employment and awareness of method Inference, extrapolation, synthesis of information, and construction of argument
Organizing ideas (based on mental task representation)	Organizing intertextual relationships between ideas Organizing ideas in a textual structure	Interpreting graphic and visual information Organization of text/scrambled text Understanding text type and communicative function Making academic arguments Grammar and text relations Text editing	Vocabulary and metaphor Complex grammar and text relations Text type (including visual representations) Communicative function Employment and awareness of method Inference, extrapolation, synthesis of information, and construction of argument

As part of a larger study (Drennan 2019), the APPMI was initially designed to measure the preparedness of prospective postgraduate social science students to present information in their field of study, the results of which were used to inform the development of discipline-specific academic writing initiatives. Thus, a discipline-specific approach was taken to the design of the various subtests, incorporating relevant reading texts that formed part of students' undergraduate prescribed reading. In this way, the test was

designed to measure these students' 'readiness' to negotiate the discourse relevant to the particular discourse community. Table 2 reflects the nine subtests of the test and their corresponding weightings.

Table 2*: Test specifications: APPMI (Drennan 2019, 2021)*

Subtest	Number of items	Weighting
Organizing information visually	8	8
Organization of text	5	5
Understanding academic vocabulary [two-word format]	6	12
Interpreting graphic and visual information	8	8
Understanding text type and communicative function	5	5
Text comprehension	18	18
Making academic arguments	8	16
Grammar and text relations	16	16
Text editing	6	12
Totals	**80**	**100**

The **APPMI** had thus been through various rounds of piloting and implementation before it was considered for further piloting in the current study. The initial plan was to conduct a traditional pencil-and-paper pilot test. However, with COVID-19 measures in place on the university campus, the logistics of having students write the test at scale became problematic. In order to accommodate students in a socially-distanced way, the pilot would have had to be spread over several weeks. This raised an obvious concern: the longer the test extended into the academic year, the more academic literacy content would be covered in the literacy courses, which could potentially influence students' performance on the test. In addition, a major concern was that the pilot would become a super-spreader event, which would have put the health and well-being of students and staff at risk. Thus, the decision was made to administer the pilot using the online platform, Question Mark, which had its limitations. One of these is the extent to which students' digital and computer literacy skills influenced their navigation and comprehension of the various texts and items. In addition, adapting a test that was initially intended to be written in pencil-and-paper format for online administration was

technically and logistically complex. The test had to be piloted by a core team (of test designers and administrators) more than 19 times to ensure that the final version was user-friendly and that texts and items displayed properly. This involved creating two versions of the test – a desktop and mobile version; students selected the appropriate version depending on the device they used to complete the test. Moreover, the number of student queries received during the pilot was immense and if the APPMI were to remain online for the foreseeable future, an entire team would need to be dedicated to its administration.

4. The administration of the APPMI

Pilot history

The pilot history of the APPMI involves the administration of the first version of the APPMI with 1175 social sciences students, after which problematic items were either revised or removed in the refined test (referred to hereafter as the 2nd pilot) that was administered to a further 261 undergraduate students. During the 2nd pilot, the full test (Test 1) was split into two parts (Test 2 and 3) for various logistical and administrative reasons (see Drennan 2020). Following further refinement, the final version of the APPMI was administered to 36 honours students as a measure of their preparedness to present information, particularly in written format (see Drennan 2019). The Classical Test Theory analysis results gleaned from the pilot and pre-test versions of the APPMI pointed to a technically sound test. It was on the basis of these initial results that the decision was taken to investigate whether the test would yield similar results when piloted on first-year students who had already been identified as at-risk, using an algorithm, and accordingly channelled, through that algorithm, into academic literacy courses. Thus, all 1088 students who participated in the current (2021) pilot had been identified by means of the algorithm as in need of literacy support and enrolled accordingly in one of the developmental academic literacy courses at the UFS.

Analysis and discussion of results

The TiaPlus (CITO 2005) analyses of results reflected in Table 3 show that the APPMI obtained consistently acceptable reliability scores, as measured by both Cronbach alpha and Greatest Lower Bound (GLB), although the latter was not available for all the versions of the test. The Cronbach alpha values range between 0.82 and 0.91, which are well above the benchmark measure of 0.73, and the GLB values, where available, were also high (0.93 and 0.95). In terms of the overall facility of the test, a first indication of its appropriateness for the ability of the test candidates, the average P-values are all in the vicinity of the desired 50%, particularly in the case of the current pilot which obtained an average score of 51%. A further measure of technical quality can be derived from the average *Rit* values which measure how well the test differentiates between candidates.

All these values are either very close to or above the satisfactory 0.3 mark, indicating a high level of conformity with the expectations.

Table 3*: Reliability and related indicators for APPMI*

APPMI results	2nd pilot			Pre-test (n=36)	Current pilot (n=1088)
	Test 1 (n=56)	Test 1+2 (n=159)	Test 1+3 (n=158)		
Cronbach alpha (reliability)	0.90	0.87	0.82	0.91	0.86
GLB			0.95		0.93
Ave P value		57.73	58.52		51.29
Ave Rit Value	0.33	0.39	0.35	0.36	0.29

The homogeneity of a test is a further measure of technical quality, indicating whether the multiple components of a test make up an instrumental unity. Although a certain degree of heterogeneity can also be indicative of a rich construct (Van der Slik & Weideman 2005), a more homogenous test is typically associated, furthermore, with a more reliable test. The TiaPlus (CITO 2005) factor analysis results depicted in Figure 3 (2nd pilot) and Figure 4 (pre-test) depict a homogenous construct. The outlying items in the 2nd pilot were omitted from or revised for inclusion in the pre-test version of the test. The pre-test factor analysis (Figure 4) helped to flag items that had performed well in the 2nd pilot, so the recommendation at the time was to see if these items would again be flagged as problematic in future pilots of the test. As seen in Figure 5, the factor analysis illustrates a predominantly homogenous test, except for a few items associated with one subtest. Three outlying items (36, 60 and 61) had undesirable discrimination (*Rit*) values, although they were not flagged as problematic in previous versions of the test. Closer investigation would be required to determine whether these items need further refinement should the test be considered for future use. However, as is the case with test pilots, new items are likely to be flagged with each administration.

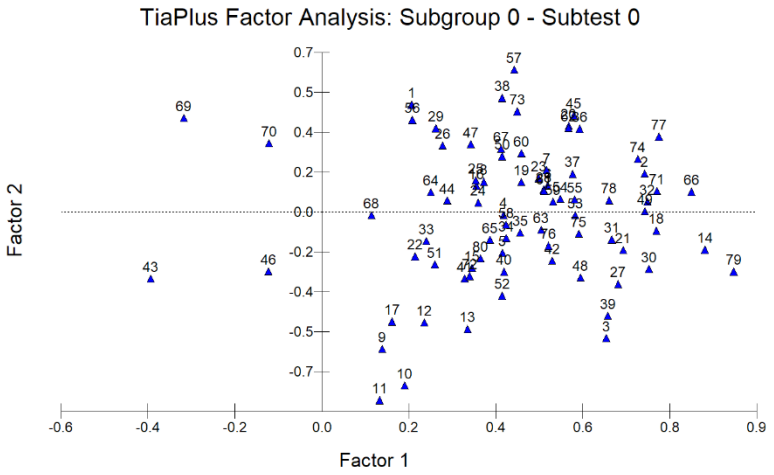

Figure 3: Factor analysis for APPMI 1 of 2nd pilot

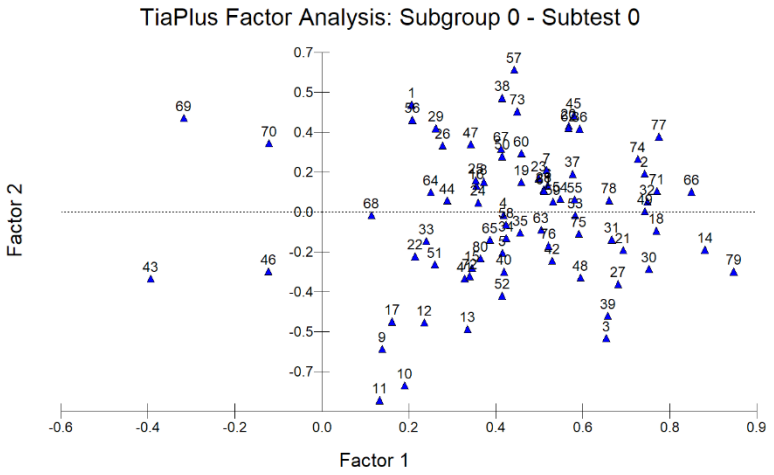

Figure 4: Factor analysis for pre-test

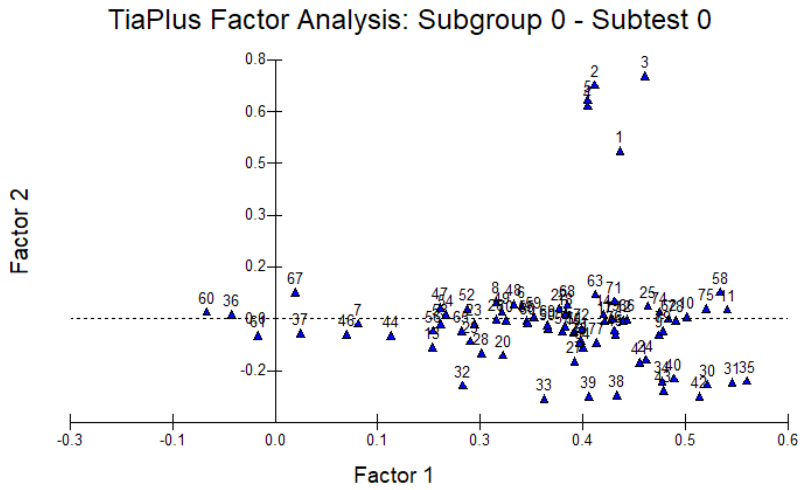

Figure 5: Factor analysis for current pilot

To illustrate how the administration of the current pilot of the APPMI fared in terms of fulfilling the requirements of technical differentiation across functionally different tasks, we may consider the analysis in Table 4.

Table 4: Test-subtest correlations and subtest inter-correlations (n=1088)

		Subtest									
Sub-test	**Total test**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	**0.45**										
2	0.61	0.23									
3	0.62	0.21	0.34								
4	**0.47**	0.15	0.28	0.24							
5	**0.54**	0.18	0.32	0.32	0.28						
6	0.71	0.17	0.32	0.35	0.30	0.34					
7	**0.53**	0.15	0.24	0.30	**0.13**	0.21	0.25				
8	**0.37**	**0.09**	0.18	0.17	0.15	0.16	0.15	**0.13**			
9	0.75	0.24	0.44	0.37	0.26	0.32	0.36	0.38	0.20		
10	0.61	0.21	0.32	0.32	0.27	0.26	0.33	0.30	0.17	0.41	

Number of items:	80	5	6	8	4	4	16	8	5	18	6
Average test score:	41,03	2,45	2,93	3,55	2,06	2,11	8,00	4,31		9,98	3,77
Standard deviation:	11,18	1,83	1,43	1,84	1,16	1,15	3,50	1,69	1,44	3,10	1,57
SEM:	4,15	0,65	1,12	1,22	0,88	0,87	1,22	1,24	0,89	1,95	1,01
Average P-value:	51,29	48,92	48,90	44,43	51,61	52,37	50,00	53,92	37,10	55,45	62,91
Coefficient Alpha:	0,86	0,80	0,39	0,52	0,35	0,38	0,74	0,41	0,57	0,60	0,57
GLB:	0,93	0,87	0,45	0,61	0,44	0,45	0,88	0,48	0,64	0,00	0,61
Asymptotic GLB:	Na	0,87	0,39	0,56	0,42	0,42	0,88	0,47	0,62	0,00	0,59

It is evident that the degree of differentiation and functionality is such that in only half of the subtests, the test-subtest correlation was slightly lower than ideal (<0.6) (for a discussion and rationale for these parameters, see Weideman 2020a). Almost all of these (four out of five) contained either four or five items, which could also have played a part in the lower correlations in these instances. In addition, only three out of 36 subtest inter-correlations are outside of the desired parameters (0.2 – 0.5), and they occur only in subtests with fewer items. Moreover, in previous administrations of the test (Drennan 2019, 2020), the average test-subtest correlations increased from 0.58 in the 2nd pilot to 0.63 in the pre-test.

A Rasch analysis (Linacre 2018) can provide further evidence of the probability of whether there is a desirably high degree of fit between the ability of test takers and the items used. Weideman (2020a) uses this kind of analysis to show, on a Wright map (as in Figure 6), that in the refined, post-pilot version of a test of proficiency in language for economics and finance, persons and items fit into the desired parameters of between -3 and 3 logits (Van der Walt 2012; Van der Walt & Steyn 2007). In the present case, Figure 6 shows no items (on the right) that fall outside of these parameters, or outside the more conservative parameters (-2 and +2) of high-stakes tests that Keyser (2017) proposed for another postgraduate test of academic literacy (in Afrikaans). Furthermore, the distribution of candidates (the 'persons', on the left) indicates a fairly normal curve, meaning there is an adequate fit between candidates' ability and the test, and a fair likelihood that test-takers will be able to cope with both easy and difficult test items.

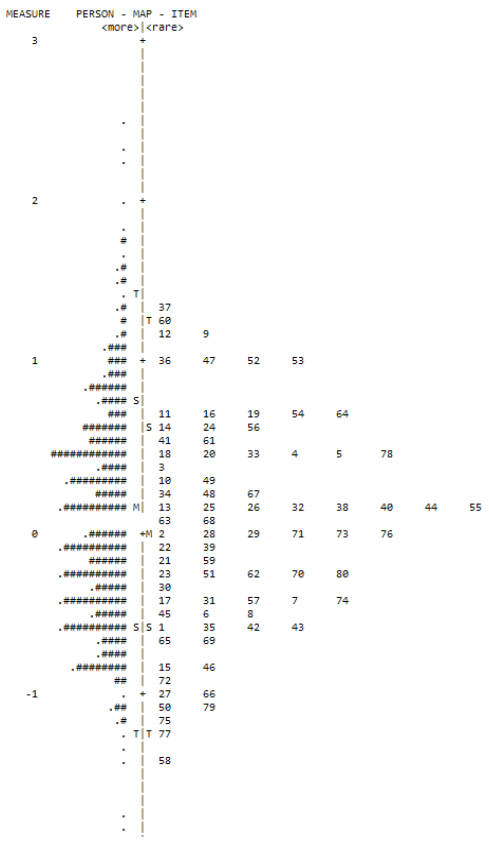

Figure 6*: Wright map: person-item distribution map for APPMI*

The next step in the analysis process was to determine whether there was a correlation between students' APPMI and NBT scores. Since the algorithm was designed to predict performance on the NBT, the results of which are typically used to place students in academic literacy and language courses at the UFS, it followed that the potential relationship between the two sets of scores should be investigated. A similar correlation analysis was not appropriate for the APPMI and algorithm scores, as the algorithm is a machine-learning algorithm that predicts outcomes based on historical data and thus does not measure academic literacy proficiency, as do the NBT and the APPMI. The results are however included out of academic curiosity.

In order to run the correlation analysis, only the scores of those students who had written both the APPMI and NBT could be used, which presented a very small sample size (n=46). For this reason, the scores of the refined pilot version of the APPMI (administered during the final stages of the test refinement process in 2018) were also included in the analysis to see if similar results could be produced for earlier versions of the test. Although the latter sample size (n=41) was also very small, both analyses yielded similar results. For the APPMI and algorithm correlation, the scores of students

who received a prediction and had written the APPMI in 2021 were used. The sample size (n=610) was more than ten times larger than the other two correlations. Table 5 shows the results of the Spearman's rank-order correlation the non-parametric version of correlation that measures the strength and direction associated between two ranked variables. These results show a strong correlation (0.60 and 0.70) between the NBT and APPMI scores. Furthermore, the p-values in both cases are significant (at $\alpha = 0.001$) indicating that there is sufficient evidence to conclude that there is a significant monotonic association between the APPMI and NBT scores, since the correlation coefficient is significantly different from zero. The APPMI and the algorithm scores had a significant moderate correlation (0.26) albeit much lower than those for the first two correlations.

Table 5*: Spearman R correlation between measures*

Analysis	Correlation	p-value	Sample size
APPMI (refined) vs NBT	0.70	< 0.0001	41
APPMI (pilot) vs NBT	0.60	< 0.0001	46
APPMI (pilot) vs Algorithm	0.26	< 0.0001	610

At face value, one would expect the correlation between the APPMI and the algorithm (0.26) to be much closer to the 0.85 accuracy that was reported by the confusion matrix during testing, or at least match the other two correlations. However, in contrast to the APPMI, the algorithm predicted a binary outcome namely 1 for "Algorithm predict PASS" or 0 for "Algorithm predict FAIL". The binary cut-off for the training data was determined by the UFS literacy policy, stating that an NBT score of 64% and higher would classify a student as proficient. The benefit of this cut-off for binary predictions is that it simplifies placing students, but it also results in very strict classification. The South African national matric pass rate decreased by 5.1% from 2019 to 2020 due to the effects of COVID-19. One may speculate that since the algorithm uses all matric marks (and not only literacy related marks) as variables, the effect of COVID-19 may have had a direct influence on the algorithm's prediction outcomes. For example, a 5% decline in overall marks does not necessary imply a significant decrease in a matriculant's literacy, but it could potentially be enough evidence for the algorithm's prediction to cross below the 64% cut-off and cause the algorithm to classify a student as "FAIL" instead of "PASS" (under normal conditions). Predicting a whole cohort's literacy marks in a pandemic shows that machine learning algorithms are prone to statistical bias when the surrounding environment changes substantially. One may assume that high-performing matriculants in 2020 would not be affected by this statistical bias as their marks provide enough of a buffer to remain above the 64% cut-off. Although the algorithm served its purpose well by exempting high-performing

students that could not write the NBT, the algorithm's predictions cannot be interpreted as a direct measure of students' literacy.

Although the sample sizes linked to the correlation results are too small to make inferences on a large scale, they do begin to support the case being made for the soundness of the APPMI as a measure of academic literacy. The results suggest similarities in the APPMI and NBT's assessment of students' academic literacy proficiency, and the NBT is a reputable assessment tool used on a national level in South Africa. Furthermore, given the accuracy of the algorithm in identifying "at risk" students in need of literacy support, the results above also support the case for using a combination of the APPMI and algorithm to arrive at a more comprehensive understanding of students' literacy and language performance, and corresponding needs.

5. Limitations

There are various limitations that need to be considered. The first is in terms of the online mode of delivery in the current pilot. Several students reported problems navigating the online platform; this was made evident by the number of student queries received by the administration team during the pilot. An informal report from the administrator mentions queries that included 1) students generally not understanding what to do or how to navigate the online interface; 2) scheduling errors; 3) '50201 errors' (which appear when a student loses internet connectivity); and 4) '50038 errors' (which appear when a student tries to access the test once it has closed). Over 140 student emails were received and over 441 discrete WhatsApp messages were sent between the team and facilitators during the pilot. In addition, there were several students who ran out of data, or whose internet connection was lost or dipped during the test. This is again evidenced by the number of student and facilitator queries that were received notifying the administration team of test loading errors. Further evidence of the technological struggles that students experience in general comes from the Students' Access to and Use of Learning Materials Survey (SAULM) that was commissioned by the Department of Higher Education and Training in 2020. UFS students commented that they struggled with power outages, slow internet speeds, and a shortage of data (Department of Higher Education and Training 2020). In addition, the UFS SAULM survey shows that only 60% of students own laptops, while 90% use smartphones to engage with their studies (Department of Higher Education and Training, 2020).

The second limitation pertains to the small sample sizes in the correlation analyses. Thirdly, it is very difficult to interpret the APPMI and the algorithm correlation without considering the algorithm's limitations. The APPMI and algorithm were created with different outcomes in mind the APPMI for testing literacy and the algorithm for placement based on statistical modelling using past literacy data. An inherent requirement for creating an algorithm is to have sufficient, recent and high-quality training data. Even though the algorithm used data from 27 528 students, the majority of UFS students did not write the NBT from 2014 to 2020, effectively reducing the training data available. Furthermore, this period includes disruptions such as the #feesmustfall campaign,

making statistical generalizations more difficult, since algorithms assume that the past will repeat itself. It is therefore difficult to quantify the effects of these limitations.

Although no definitive extrapolations can be made using these results, they may serve as an important start to a more in-depth analysis in future research.

6. Findings

This paper has examined different strategies that could be used to identify students who arrive at universities and show risk of not performing successfully as a result of language ability. Despite the various limitations of the online mode of delivery in the current pilot, the results show that the APPMI performed consistently well and is technically sound. It conforms to expectations in respect of its technical reliability, its being a technical whole or unity within a multiplicity of components, with subtests having different functions organically working towards the same measurement goal. It also possesses a high degree of technical appropriateness ('fit') as regards the level of ability being tested and the measurements employed. In short, it fulfils many of the early indications of what is often presented as warrants for the technical validity of a test, its ability to perform and measure as it should (Weideman 2019a, 2019b). It could therefore be considered for further refinement and wider use than initially envisaged as an alternative way of assessing language risk for incoming students at the UFS.

We have concluded that the crisis of the pandemic yielded an opportunity for those involved in offering language development interventions to become more fully acquainted with language assessment. As a consequence, the numerous academic literacy facilitators and other colleagues involved themselves have gained in assessment literacy.

7. Conclusion

Several challenges remain. Within the sub-organisational unit where this work is done, our strategic goal is to move towards greater relevance both in coursework and assessment. That means that we shall have to consider field-specific tests of academic literacy rather than the generic ones (such as the APPMI as it was used in this particular study) that are currently in use. That in itself will give rise to at least two problems. First, we shall need to answer the question: How specific must such tests be? If they are highly specific, what is the difference between them and the language encountered in assessments within the discipline already? If they are not very specific, will they still be relevant, or will they more likely reflect the content of a low-grade textbook (*Criminology for dummies*)? Second, if tests are to be used fairly across fields, how does the test designer ensure their equivalence? What is desirable is not always immediately possible, so before we proceed further we should have at least preliminary answers to these questions, and potential strategies to deal with them adequately.

The exploration we undertook and reported on in this paper shows that there is still work to be done, and much left to explain. Further investigation is needed, in particular on the refinement of the algorithm we employed (and the assumptions underlying it), and its possible closer future alignment with the results of academic literacy tests, with which it was combined in this analysis. Our productive employment of these tools in a time of disruption may yet lead to further insights to ensure just and fair outcomes in language assessment.

References

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

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

Cloete, N. 2016. *Free higher education. Another self-destructive South African policy*. Centre for Higher Education Trust: Cape Town. Available: http://hdl.voced.edu.au/10707/409532

Cummins, J. 2000. *Language, power and pedagogy: Bilingual children in the crossfire*. Clevedon: Multilingual Matters Ltd, 2000.

Cummins, J. 2001. Bilingual children's mother tongue: Why is it important for education? *Sprogforum* 7(19): 15-20.

Department of Higher Education and Training. 2020. Students' access to and use of learning materials survey report. Pretoria: DHET.

Deygers, B. 2017. Just testing: Applying theories of justice to high-stakes language tests. *International Journal of Applied Linguistics* 168(2): 143–163. DOI: 10.1075/itl.00001.dey

Deygers, B. 2019. Fairness and justice in English language assessment. In Gao, X. (Ed.) *Second handbook of English language teaching*. Cham, Switzerland: Springer. (Handbooks of Education). DOI: 10.1007/978-3-319-58542-0_30-1

Deygers, B., & Malone, M.E. 2019. Language assessment literacy in university admission policies, or the dialogue that isn't. *Language Testing*. DOI: 10.1177/0265532219826390

Deygers, B., Van den Branden, K. & Van Gorp, K. 2017. University entrance language tests: A matter of justice. *Language Testing* 1–28. DOI: 10.1177/0265532217706196

Drennan, L. 2019. Defensibility and accountability: Developing a theoretically justifiable academic writing intervention for students at tertiary level. PhD thesis, University of the Free State. URI: https://hdl.handle.net/11660/10888

Drennan, L. 2021. Assessing readiness to write: The design of an Assessment of Preparedness to Present Multimodal Information (APPMI). In Weideman, A., Read, J. & Du Plessis, T. (Eds.) Assessing academic literacy in a multilingual society: Transition and transformation, pp. 196–216.

Frederiksen, C.H. 1975. Representing logical and semantic structure of knowledge acquired from discourse. *Cognitive Psychology* 7: 317–458.

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

Linacre, J.M. 2018. *A user's guide to WINSTEPS Ministep: Rasch-model computer programs*. Chicago: Winsteps.

Network of Expertise in Language Assessment [NExLA]. 2021. Bibliography of language assessment. Available: https://nexla.org.za/research-on-language-assessment/. [Accessed 25 May 2021.]

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. DOI: 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. DOI: 10.4314/jlt.v47i1.6

Read, J. 2015. *Assessing English proficiency for university study*. Basingstoke: Palgrave Macmillan.

Read, J. (Ed.). 2016. *Post-admission language assessment of university students*. Cham, Switzerland: Springer.

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

Spivey, N. N. 2001. Discourse synthesis: Process and product. In McInnis R.G. (Ed.) *Discourse synthesis: Studies in historical and contemporary social epistemology, pp.* 379–396. Westport, CT: Praeger.

Spivey, N.N. and King, J. R. 1989. Readers as writers composing from sources. *Reading Research Quarterly* 24: 7-26. Available: https://www.jstor.org/stable/748008

Tjabane, M., & Pillay, V. 2011. Doing justice to social justice in South African higher education. *Perspectives in Education,* 29(2): 10–18. URI: http://hdl.handle.net/10520/EJC87624

Van der Slik, F. & Weideman, A. 2005. The refinement of a test of academic literacy. *Per Linguam,* 21(1): 23-35. DOI: 10.5785/21-1-70

Van der Walt, J. 2012. The meaning and uses of test scores: An argument-based approach to validation. *Journal for Language Teaching* 46(2): 141–155. DOI: 10.4314/jlt.v46I2.9

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

Van Dijk, T.A. 1979. Relevance assignment in discourse comprehension. *Discourse Processes* 2: 113–126. DOI: 10.1080/01638537909544458

Van Dyk, T. 2020. Postscript: What the data tell us: An overview of language assessment research in South Africa's multilingual context. In Weideman, A., Read, J. & Du Plessis, T. (Eds.). *Assessing academic literacy in a multilingual society: Transition and transformation,* pp. 217–23. DOI: 10.21832/9781788926218-014

Weideman, A. 2003. Assessing and developing academic literacy. *Per Linguam* 19(1 & 2): 55–65. DOI: 10.5785/19-1-89

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

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

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

Weideman, A. 2020b. A skills-neutral approach to academic literacy assessment. In Weideman, A., Read, J. & Du Plessis, T. (Eds.). *Assessing academic literacy in a multilingual society: Transition and transformation*, pp. 22–51.

Weideman, A. 2021. Context, construct, and validation: A perspective from South Africa. *Language Assessment Quarterly*. DOI: 10.1080/15434303.2020.1860991

Weideman, A., Patterson, R. & Pot, A. 2016. Construct refinement in tests of academic literacy. In Read, J. (Ed.) *Post-admission language assessment of university students*, pp. 179–196. Cham: Springer. DOI: 10.1007/978-3-319-39192-2_9

Weideman, A., Read, J. & Du Plessis, T. (Eds.). 2020. *Assessing academic literacy in a multilingual society: Transition and transformation. (*New Perspectives on Language and Education: 84.) Bristol: Multilingual Matters. DOI 10.21832/WEIDEM6201

ABOUT THE AUTHORS

Laura Drennan

University of the Free State

https://orcid.org/0000-0002-9416-2590 : Email: drennanl@ufs.ac.za

Laura Drennan specialises in academic literacy development, and academic writing in particular. She holds a PhD in English language studies and academic literacy development. She is a lecturer/researcher at the Unit for Academic Language and Literacy Development, and head of the Write Site.

Albert Weideman

University of the Free State

https://orcid.org/0000-0002-9444-634X : Email: weidemanaj@ufs.ac.za

Albert Weideman is Professor of Applied Language Studies and Research Fellow at the University of the Free State. He recently published *Assessing academic literacy in a multilingual context: Transition and transformation* (2021, Multilingual Matters). He focuses on language assessment and a theory of applied linguistics.

Michelle Joubert

University of the Free State

https://orcid.org/0000-0002-5814-4053 : Email: joubertma@ufs.ac.za

Michelle Joubert is head of the academic literacy unit at the University of the Free State, South Africa. She has worked in academic literacy in the USA, UK and South Africa. Her research interests include AL practitioner identity, curriculum design, decolonisation and multilingualism in HE.

Rohan Posthumus

University of the Free State\

Email: posthumusjj@ufs.ac.za : https://orcid.org/0000-0002-8899-0111

Rohan Posthumus is a data analyst for the Centre for Teaching and Learning (CTL) at UFS. He is responsible for developing data analytics by assisting in all phases of research projects, including conceptualization, data collection, processing, data management, statistical analysis and report writing.

Tydskrif vir Taalonderrig - Journal for Language Teaching
- Ijenali yokuFundisa iLimi - IJenali yokuFundisa iiLwimi -
Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta Polelo
- Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo - Ijenali
Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo
- Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig -
Journal for Language Teaching - Ijenali yokuFundisa iLimi
- IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi
- Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo -
Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi
- Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa
Ririmi - Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa
iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta
Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo -
Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo
- Jenala yo Dyondzisa Ririmi
- Tydskrif vir Taalonderrig
- Journal for Language
Teaching - Ijenali
yokuFundisa iLimi -
IJenali yokuFundisa
iiLwimi - Ibhuku
Lokufundisa Ulimi
- Tšenale ya tša
Go ruta Polelo -
Buka ya Thuto
ya Puo - Jenale
ya Thuto ya Dipuo
- Ijenali Yekufundzisa
Lulwimi - Jena?a ya u
Gudisa Nyambo - Jenala yo
Dyondzisa Ririmi - Tydskrif vir Taalonderrig
- Journal for Language Teaching - Ijenali
yokuFundisa iLimi - IJenali yokuFundisa iiLwimi -
Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta
Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya
Dipuo - Ijenali Yekufundzisa Lulwimi - Jena?a ya
u Gudisa Nyambo - Jenala yo Dyondzisa Ririmi
- Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali
yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi -
Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo -
Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi
- Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa
Ririmi - Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa
iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta
Polelo - Buka ya Thuto ya Puo - Jenale ya Thuto ya Dipuo -
Ijenali Yekufundzisa Lulwimi - Jena?a ya u Gudisa Nyambo
- Jenala yo Dyondzisa Ririmi - Tydskrif vir Taalonderrig -
Journal for Language Teaching - Ijenali yokuFundisa iLimi
- IJenali yokuFundisa iiLwimi - Ibhuku Lokufundisa Ulimi
- Tšenale ya tša Go ruta Polelo - Buka ya Thuto ya Puo -
Jenale ya Thuto ya Dipuo - Ijenali Yekufundzisa Lulwimi
- Jena?a ya u Gudisa Nyambo - Jenala yo Dyondzisa
Ririmi - - Tydskrif vir Taalonderrig - Journal for Language
Teaching - Ijenali yokuFundisa iLimi - IJenali yokuFundisa
iiLwimi - Ibhuku Lokufundisa Ulimi - Tšenale ya tša Go ruta