Journal for Language Teaching 58(2) 2024, 21 pages
https://doi.org/10.56285/jltVol58iss2a6269

Institutional responses to academic literacy challenges: An in-house test as an alternative for assessing academic literacy levels

Laura Drennan
University of the Witwatersrand, South Africa
E-mail: laura.drennan@wits.ac.za

Michelle Joubert
University of the Free State, South Africa
Email: joubertma@ufs.ac.za

Albert Weideman
University of of the Free State, South Africa
E-mail: albert.weideman@ufs.ac.za

ABSTRACT

The COVID-19 pandemic brought about disruptions to the administration of conventional academic literacy tests, which necessitated alternative ways of identifying students in need of literacy support. In response to this disruption, an existing in-house test was identified as a potential alternative for measuring incoming students' ability to handle the demands of academic discourse. Such an alternative was necessary for the effective identification of students needing additional academic literacy support and their placement in appropriate faculty-specific literacy courses. Although the first round of institutional piloting deemed the online version of the test appropriate for being employed for these purposes, a further administration of the test was conducted to confirm whether the traditional (pencil-and-paper) version of the test yielded similar results and whether the quality of the test is such that it is appropriate for implementation at an institutional level. This paper compares the results of these two administrations of the test and reflects on the potential value of such an in-house test to achieve and enhance the synchrony between assessment, curriculum and teaching, building on earlier work on this.

Keywords: academic literacy; design principles; language assessment; student placement; validity

CITATION

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

ISSN: 0259-9570 | eISSN: 2958-9320

1. Why *another* test of academic literacy?

The large-scale use and development of tests of academic literacy at South African universities has been well-documented (Van Dyk, 2021). For three decades, many South African institutions have used academic literacy tests to make decisions about students' academic language and literacy developmental needs (Sebolai, 2022a). Such tests are sometimes used for diagnostic purposes and are thus written by students about to enter university. Based on their performance in these tests, students who may need additional developmental academic literacy support are identified and the necessary support is put into place to ensure that these students are empowered to meet the demands of tertiary study. As Sebolai and Stanford (2020: 77–78) state,

> Such tests are, firstly, crucial for determining the degree of academic preparedness among incoming students and the degree to which a lack thereof will hinder their academic achievement. Secondly, they are important for ensuring that any support provided to empower students is sufficiently tailored to meet the demands and challenges they will encounter during their years at university.

Developmental support most often, though not always, takes the form of an additional credit-bearing module, which focuses on helping students develop foundational academic literacy practices or rather "the ability to read and write in socially legitimate ways in the academy" (Boughey & McKenna, 2016, p.5; see also Weideman, 2003; Boughey, 2013; Mihindou, 2019; Sebolai & Stanford, 2020; Joubert, 2023). At some institutions, these modules take the form of additional or augmented support modules, while at others they are embedded within a student's academic programme or via writing intensive programmes. Diagnostic tools play a key role in helping institutions make placement decisions. Engaging with the quality and validity of these tests is thus vital since they have the potential to impact students' academic literacy development as well as what content is taught (Sebolai, 2022a; Sebolai & Stanford, 2020). That they target the identification of students who are most at risk as a result of their levels of academic literacy is a first indication of the awareness of their institutions of their ethical responsibility to support incoming students. Their impact also raises questions of potential stigmatization. Hence, the results need not only be utterly fair, but also, when made public, need to ensure privacy (Weideman, 2024, Chapters 15 and 16).

Currently, the two most prevalent tests of academic literacy for first-time undergraduate students used nationwide are the National Benchmark Test (the NBT) and the Test of Academic Literacy Levels (the TALL), which are both sound tests of students' academic literacy abilities at a first-year level. However, lessons from the COVID-19 pandemic showed that having an alternative, in-house diagnostic test of academic literacy can enable academic literacy practitioners to make placement decisions when students have limited access to testing centres or when the cost of writing a national test is prohibitive, as may be the case for some South African students. Additionally, and as

Sebolai (2022a) points out, it is imperative that language needs analyses and diagnostic tests of academic literacy inform curriculum design. This is so that academic literacy practitioners are better able to design their interventions in a way that is "appropriate, accessible, theoretically defensible, useful, and potentially more effective as well" (Sebolai, 2022a p.2). In other words, academic literacy practitioners involved with curriculum design should have an in-depth knowledge and understanding of the constructs assessed in these diagnostic tools so that the curricula of the interventions do indeed address the developmental needs of the students as 'diagnosed' by the test. The adoption of an in-house test of academic literacy can thus be useful. The first step towards achieving such alignment between diagnostic tools and curricula (Weideman, 2019) is determining whether the test intended for these purposes is indeed an acceptable alternative to established measures of academic literacy.

As part of doctoral research which sought to determine the efficacy of a theoretically justifiable academic writing intervention for students at a tertiary level (Drennan, 2019), the Assessment of Preparedness to Present Multimodal Information (the APPMI) was designed. In the pilots for this study, the test discriminated well among undergraduate students and obtained satisfactory results in the Rasch and Classical Test Theory analyses done on them (Drennan et al., 2021). It was thus deemed appropriate for piloting at a large scale and for potential use as a diagnostic tool for first-time entering undergraduate students for placement into developmental academic literacy modules, which students are required to take in their first year of study. The purpose of this paper is, therefore, to investigate whether the results of the first 2021 large-scale institutional administration could be replicated in a different format, and as a further investigation of the quality of the test design employed in the APPMI.

2. Test design

The APPMI was designed with the notion that the processes of selecting and organising information precede the production of information in writing or any other format. When an experienced reader/writer is assigned the task of producing information with a clear objective, they engage in more advanced tasks such as gathering relevant information from primary sources and organising it to create connections between ideas that strengthen and support the stated goal (Spivey, 2001; Spivey & King, 1989). This necessitates adapting what they read to the task at hand and their understanding of the text organization principles of various source kinds. Their comprehension of the conventional organisation of discourse in various text types enables them to select information using significance criteria, understand the way concepts are connected in a text through textual cues, and draw conclusions between texts (Frederiksen 1975; Spivey & King, 1989; Van Dijk, 1979). As such, creating discourse and synthesizing details are believed to be intimately related to understanding discourse.

The APPMI was designed to measure the skills related to various cognitive phases associated with the processes of selecting and organising information (Drennan, 2019;

2021; Drennan et al., 2021). Table 1 shows the alignment between these cognitive phases, the subtests of the APPMI, and the construct of academic literacy proposed by Patterson & Weideman (2013a, 2013b). Accordingly, Table 2 shows the various subtests of the APPMI and their corresponding weightings.

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

Cognitive phases	Sub-processes	APPMI subtests	Alignment with construct
Conceptualization	• Task representation • Macro-planning	• Understanding text type and communicative function • Making academic arguments • Interpreting graphic and visual information • Text comprehension	• 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
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

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

Subtest	No. 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**

3. Population and administration

There have been five administrations of the APPMI in total – three pilots for refinement purposes and two large-scale administrations; the table below summarises the details of each of these. The last two iterations were administered to the target (first-year) cohort. As a result of COVID-19, students were unable to write the NBT, the results of which are typically used to place students in academic literacy courses. Consequently, the 2021 test cohort constituted students who had been identified as “at risk” by means of a machine-learning algorithm developed by the Centre for Teaching and Learning (CTL) and channelled into faculty-specific academic literacy courses accordingly (see Drennan et al., 2021 for further details). The 2023 cohort also involved students enrolled in the literacy courses; these students had obtained a score below 64% on the NBT. For those who had not written the NBT, the algorithm was used to identify the portion who were at risk and in need of additional academic literacy support. For both the 2021 and 2023 administrations, every effort was made to ensure equal representation across the various faculty-specific literacy courses. Ethical clearance was granted by the UFS ethics committee for the 2023 study (UFS-HSD2020/1475/2910/21/3), and students were asked to consent to take part in the study.

Table 3: Administration history (Drennan, 2019; 2021; Drennan et al., 2021)

Version	**Year**	**Test candidates**	
First version	2018	1175	Undergraduates
2nd Pilot (refined test)	2018	261	Undergraduates
Pre-test	2019	36	Honours
1st Institutional administration	2021	1088	First-years
2nd Institutional administration	2023	2292	First-years

Ideally, the APPMI should be written in the traditional pencil-and-paper format; however, this was not possible during the COVID-19 pandemic, when such an occasion would have served as a potential super-spreader event. As a result, the 2021 iteration was administered entirely online, in QuestionMark. The purpose of the 2023 administration was to determine whether the test would perform similarly when administered in the traditional format. However, to compare the results of the two administrations more accurately, a portion of the 2023 test cohort (309) completed the test online, while the bulk of the students (1983) completed the pencil-and-paper format. As with previous pilots, the test was divided into two parts which were completed in two separate sessions of 2 hours each, one week apart, during students’ scheduled academic literacy class time. The reason for splitting the test into two parts was that it was administered to students during their academic literacy class time. The test, however, takes roughly 2.5 hours to complete, and since the length of each academic literacy class is only two hours, the students needed two sittings to complete the entire test. Additionally, for the purposes of the pilot, the researchers needed to ensure that all students completed all parts of the APPMI.

4. Conditions of language test design

Given the potential impact of the test on students' literacy development, it is critical to assess its quality and validity (Sebolai, 2022a; Sebolai & Stanford, 2020). Essential to determining the quality of a test's design and its appropriateness for its intended purpose, is an assessment of the extent to which it fulfils various design principles. The researchers concentrated on the principles of design considered most crucial in the early stages of test validation. The evaluation of whether these design principles have been met is determined largely by the interpretation of quantitative data and the results of technical analyses of the empirical and factual properties of the test. The chosen principles are listed in this section along with follow-up questions that may be needed to illustrate how these principles have been met—or not (Van der Walt & Steyn, 2007).

The first of these principles concerns the technical integrity of the test and whether there is "unity within a multiplicity of components" (Weideman, 2019, p. 43). One measure of a test's technical quality is its homogeneity. This refers to the extent to which the test is an integral whole and "whether all the test items measure the same trait (one factor)" (CITO, 2005, p.19). Thus, the question posed to measure the fulfilment of this condition was the following:

1) *Which empirical measure(s) of homogeneity or heterogeneity may be provided to show an acceptable level of homogeneity for the test?*

To answer this question, the results were subjected to Rasch analyses, and a factor analysis was undertaken. The latter is a statistical criterion used widely, but also programs built on Classical Test Theory (CTT), to measure the degree of homogeneity of a test, and the first measure reveals, among other things, the degree of fit between test items in terms of individual ability and item difficulty.

The technical reliability of a test, in terms of its measurement of language ability, constitutes the second criterion. The question posed in this case was the following:

2) *Which measures reflect the test's level of reliability in terms of consistently measuring the ability being assessed?*

To answer this question, various CTT analyses were run to measure test level reliability in terms of Cronbach's alpha coefficient, Greatest Lower Bound (GLB) and "person reliability", item reliability across the test as a whole (the latter two deriving from Rasch), and average item-total correlations.

Another principle concerns the ability of the test to function as a "technically differentiated but whole assessment" (Weideman, 2020:64). This refers to the extent to which the various subtests, measuring different sub-abilities, work together with other subtests and the overall test as an organised whole. The question posed in this respect was the following:

3) *What measures demonstrate that the test is organised as a differentiated but technical whole, with each subtest functioning on its own and together with other subtests to contribute to the viability of the measurement?*

This was tested by analyses done in Iteman 4.4 and TiaPlus (CITO, 2005) to determine the correlations between the various subtests, as well as the correlations between the subtests and the test as a whole.

The question formulated to fulfil the fourth principle concerning the appropriateness and relevance of the test was the following:

4) *What empirical evidence demonstrates the technical appropriateness of the test in terms of exhibiting an acceptable degree of fit between candidate ability and difficulty?*

Evidence for this question was gleaned from CTT measurements of the mean *P*-value or facility of the test, the possibly normal distribution of scores, as well as Rasch analyses of item-to-person and person-to-item fit.

The final question concerning the principle of test fairness was the following:

5) *Does the test measure candidates fairly?*

To determine the answer to this question, the CTT and Rasch analyses were revisited to determine if any candidates had potentially been misclassified by the test and whether such misclassified candidates could be provided a fair chance of taking a similar test.

The following section discusses the results of various sets of analyses to assess the extent to which the specified design principles have been met.

5. Analysis and discussion of results

An important first step in the comparison of the two administrations was to determine whether the difference in the results of the online and traditional iterations was statistically significant. A t-test was conducted to determine the difference between the 2023 online and 2023 traditional; 2021 (online) and 2023 traditional; and 2021 (online) and 2023 combined (online and traditional) versions. The results in Table 4 show that there was no significant difference (p=0.6519) between the results of the 2023 online and traditional iterations of the test. While not statistically significant (p<0.05), the comparisons between the 2021 and 2023 iterations suggest a potential difference, which indicates that the difference in the mean score (p=0.0526 and p=0.0589) is more likely to be related to test circumstances (notably COVID-19 and lockdown) than to mode of delivery. For this reason, the 2023 online and traditional results were combined for further comparison with the 2021 test results.

Table 4: t-Test: Two-Sample assuming unequal variances

Test version	Mean	Obs	df	T Stat	P(T<=t) two-tail
2023 APPMI online	52.57	309	414	-0.451	0.6519
2023 APPMI paper	52.97	1983			
2021 APPMI (online)	51.91	1088	2308	-1.939	0.0526
2023 APPMI paper	52.97	1983			
APPMI 2021	51.91	1088	2202	-1.889	0.0589
APPMI 2023	52.92	2292			

To answer the first question concerning the homogeneity of the test, one may refer to the results of the factor analyses depicted in Figure 1 (Drennan et al., 2019) for the 2021 online pilot and Figure 2 for the 2023 pencil-and-paper administration of the APPMI. For both iterations, the results show an acceptable degree of homogeneity, except for a few items associated with one subtest (*Organisation of Text*). Upon closer inspection, this section's discrimination value (*Rit)* for the 2023 administration was within the desired range (>0.3). Still, the facility value (i.e. difficulty level) was slightly lower (*P*-value of 45%) than the desired 50%. However, the 2021 pilot's facility and discrimination values for this subtest were within range (*P*-value of 49%; average *Rit* value of 0.75), which could suggest that the issue may lie with the 2023 test cohort and not necessarily with the subtest in question. Furthermore, the three outlying items (36, 60 and 62) that were flagged for undesirable (*Rit*) values in the 2021 pilot were not flagged in the 2023 administration.

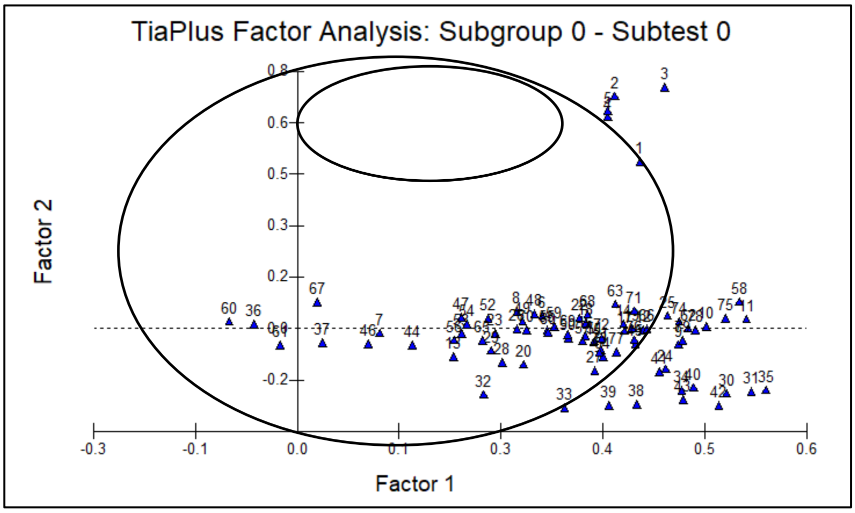


Figure 1: Factor analysis of the APPMI (2021, online)

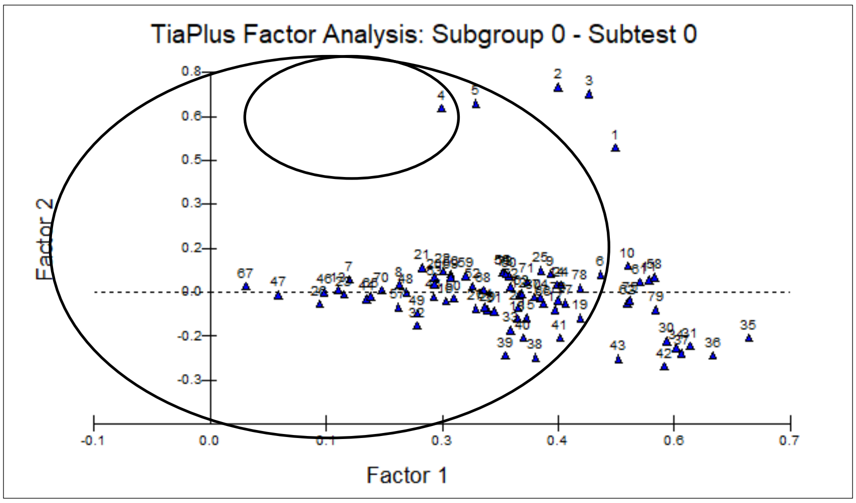
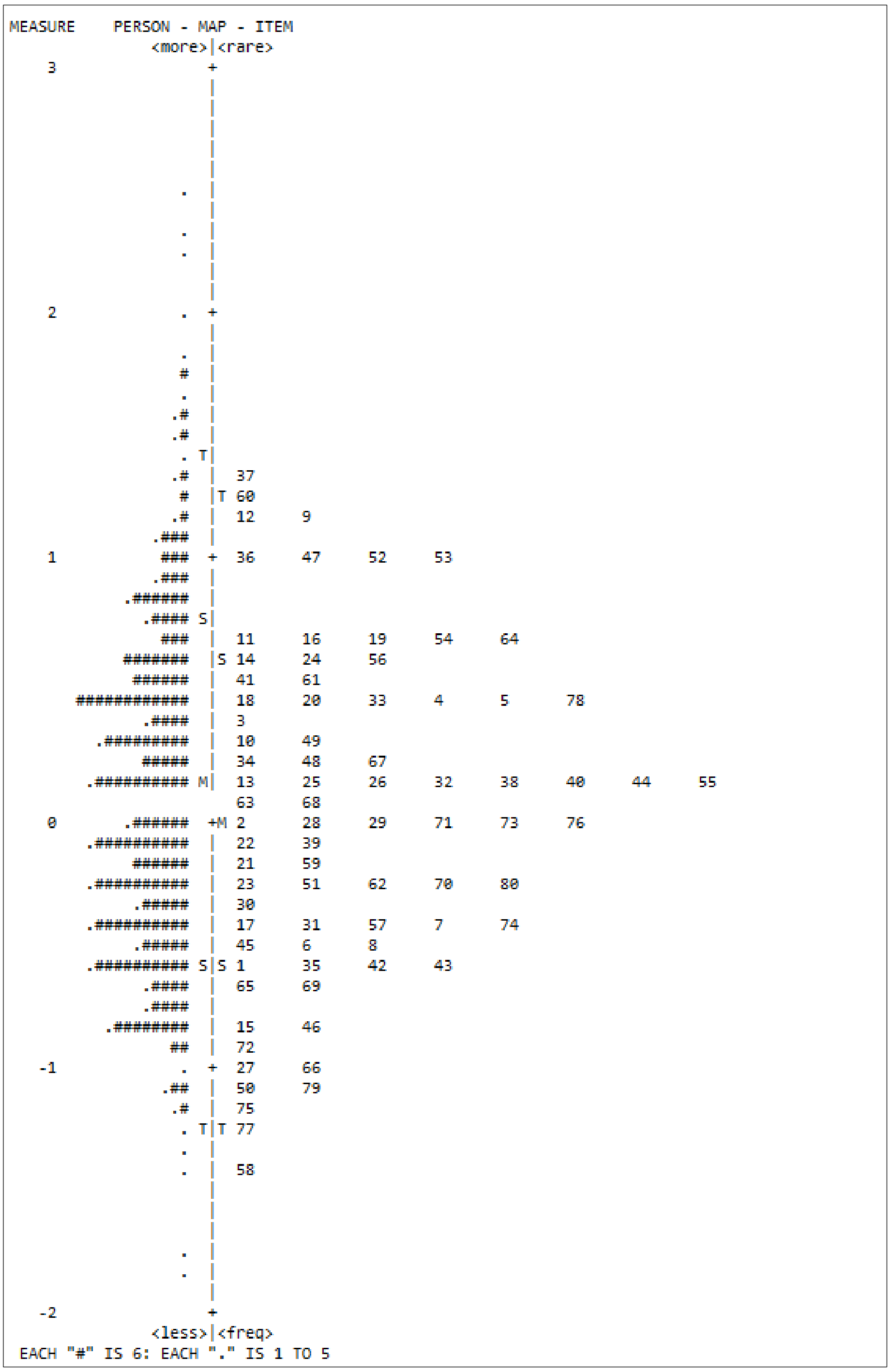


Figure 2: Factor analysis of the APPMI (2023, P&P)

In terms of answering the second question pertaining to the technical unity of the test, the following Rasch analysis results serve to measure the degree of ‘fit’ between individual ability and item difficulty. The 2021 results reflected on the Wright map in Figure 3 show that no items (on the right) fell outside the desired parameters of between -3 and 3 (Van der Walt, 2012; Van der Walt & Steyn, 2007) or outside the parameters (- 2 and +2) of more conservative, high-stakes tests (Keyser, 2017).



```
MEASURE    PERSON - MAP - ITEM
               <more>|<rare>
    3                +
                     |
                     |
                     |
                     |
                     |
                  .  |
                     |
                  .  |
                  .  |
                     |
                     |
    2             .  +
                     |
                  .  |
                  #  |
                  .  |
                 .#  |
                 .#  |
                  . T|
                 .#  |  37
                  #  |T 60
                 .#  |  12     9
               .###  |
    1           ###  +  36     47     52     53
               .###  |
            .######  |
              .#### S|
                ###  |  11     16     19     54     64
            #######  |S 14     24     56
             ######  |  41     61
       ############  |  18     20     33     4      5      78
              .####  |  3
         .#########  |  10     49
              #####  |  34     48     67
        .########## M|  13     25     26     32     38     40     44     55
                        63     68
    0       .######  +M 2      28     29     71     73     76
        .##########  |  22     39
             ######  |  21     59
        .##########  |  23     51     62     70     80
             .#####  |  30
        .##########  |  17     31     57     7      74
             .#####  |  45     6      8
        .########## S|S 1      35     42     43
              .####  |  65     69
              .####  |
          .########  |  15     46
                 ##  |  72
   -1             .  +  27     66
                .##  |  50     79
                 .#  |  75
                  . T|T 77
                  .  |
                  .  |  58
                     |
                     |
                     |
                  .  |
                  .  |
                     |
   -2                +
               <less>|<freq>
 EACH "#" IS 6: EACH "." IS 1 TO 5
```

Figure 3: Wright map: person-item distribution map for the 2021 APPMI

Table 5 shows a truncated version of the Rasch analysis results for the degree of fit for the 2023 administration of the test. While the benchmark for average fit (Infit MNSQ) is 1.0, with problematic values exceeding 1.5, the more conservative limits, ranging from 0.75 to 1.3 (Weideman, 2020, p. 66), were applied to measure average fit. The infit mean square values for all 80 items were within range, including the two (shaded) terminal values (0.86 and 1.17). These results, together with the results of the factor analyses discussed above, confirm the technical soundness of both versions of the APPMI in terms of homogeneity and overall fit.

Table 5: Misfit order: items in the 2023 APPMI

Item	**Total count (n)**	**Infit MNSQ**	**Ptmeasure-AL Corr**	**Expected**
47	2291	1.14	0.10	0.30
67	2291	1.17	0.08	0.31
46	2291	1.08	0.14	0.27
23	2291	1.09	0.16	0.30
26	2291	1.12	0.14	0.31
53	2291	1.06	0.25	0.31
22	2291	1.05	0.26	0.31
56	2291	1.04	0.27	0.31
59	2291	1.04	0.27	0.31
64	2291	1.04	0.28	0.31
Further better fitting items not shown				
2	2291	0.97	0.35	0.31
17	2291	0.97	0.34	0.30
24	2291	0.97	0.35	0.31
41	2291	0.97	0.35	0.31
66	2291	0.97	0.31	0.26
79	2291	0.91	0.42	0.29
31	2291	0.90	0.44	0.29
37	2291	0.90	0.44	0.31
36	2291	0.88	0.47	0.31
35	2291	0.86	0.47	0.28

The third question to be answered involves the technical reliability of a test in terms of its measurement of language ability. A series of tests was therefore conducted to determine the reliability of the APPMI. The first of these measured the APPMI's test-level reliability in terms of Cronbach's alpha coefficient and Greatest Lower Bound (GLB); the preferred coefficients for these are typically above 0.85 and 0.9, respectively. The results in Table 6 demonstrate the adequacy of the APPMI's technical consistency, as the alpha coefficient was consistently higher than the desired 0.85 in the 2021 and 2023 administrations of the test, which were 0.86 and 0.88 respectively. The results for the 2023 administration were confirmed by the Rasch “person reliability” measure, which was above the acceptable 0.8 (McNamara, Knoch and Fan, 2019:52). Similarly, the GLB values for the 2021 and 2023 iterations were above the desired 0.90, as was the case for all previous pilots of the test; the GLB value for the combined 2023 results were not calculated.

Table 6: Reliability (Cronbach alpha and GLB) and related indicators: APPMI

APPMI results	2021 pilot (n=1088)	2023 traditional (n=1983)	2023 Online (n=309)	2023 Combined (n=2292)
Cronbach alpha	0.86	0.88	0.87	0.88
GLB	0.93	0.93	0.97	-
Rasch	-	-	-	0.88
Avg *P*-value	51.29	52.14	53.03	52.91
Avg *Rit* value	0.29	0.30	0.31	0.31

Further evidence of the productivity of test items comes from the overall facility (*P*-value) and discrimination (*Rit*) measures in Table 6. An observation that should be made is the improvement in the average *P*-value for the 2021 and 2023 iterations. Problematic items that were flagged in previous pilots were adjusted or discarded (Drennan, 2019, 2021) resulting in acceptable *P*-values within the vicinity of 50% for the last two pilots. The average *Rit* values were also consistently above 0.30, which is indicative of a high level of item reliability.

As measures of organised differentiation and technical functionality, Tables 7 and 8 show the test-subtest correlations and inter-correlations. For half of the subtests in both administrations (2021 and 2023), the test-subtest correlations were below the desired parameters, which should be higher than 0.6 (Weideman, 2020). However, all but one of these subtests consisted of five or fewer items, which could have affected the correlation values. Regarding subtest inter-correlations for the 2023 iteration, the five that fell outside the required parameters (0.2 – 0.5) were those within the *Organising information visually* and *Text type and communicative function* subtests indicated below in bold. The 2021 pilot results produced similar results for these two subtests, with test-subtest correlations of 0.53 and 0.37 respectively. The three inter-correlation values that were outside the parameters (in the 2021 pilot) were also within these two subtests. In terms of the other three subtests with test-subtest scores below 0.60, the estimated Coefficient alpha (Spearman-Brown) scores ranged between 0.85 and 0.97 had they been a standard norm length of 40 items. Given this and the specific language abilities measured in these subtests, it may be premature to remove these from the test at this point. Future administrations of the APPMI may be needed to determine whether this is necessary.

Table 7: Test-subtest correlations and subtest inter-correlations (n=1088) for 2021

		Subtest									Alpha (40+)
Sub -test	Total test	1	2	3	4	5	6	7	8	9	10	
1	0.45										0.97
2	0.61	0.23									0.81
3	0.62	0.21	0.34								0.85
4	0.47	0.15	0.28	0.24							0.85
5	0.54	0.18	0.32	0.32	0.28						0.86
6	0.71	0.17	0.32	0.35	0.30	0.34					0.87
7	0.53	0.15	0.24	0.30	0.13	0.21	0.25				0.78
8	0.37	0.09	0.18	0.17	0.15	0.16	0.15	0.13			0.91
9	0.75	0.24	0.44	0.37	0.26	0.32	0.36	0.38	0.20		0.77
10	0.61	0.21	0.32	0.32	0.27	0.26	0.33	0.30	0.17	0.41		0.90

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	1.85	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	52.73	50.00	53.92	37.10	55.45	62.91
Coefficient apha	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

Table 8: Test-subtest correlations and subtest inter-correlations (n=2292) for 2023

Sub-test	Total test	1	2	3	4	5	6	7	8	9	10	Alpha (40+)
1	0.45											0.97
2	0.60	0.26										
3	0.61	0.16	0.33									0.85
4	0.39	0.16	0.20	0.17								0.85
5	0.47	0.16	0.26	0.24	0.23							0.85
6	0.75	0.21	0.34	0.36	0.19	0.29						0.92
7	0.48	0.11	0.21	0.25	0.13	0.16	0.26					0.74
8	0.43	0.14	0.21	0.24	0.10	0.13	0.19	0.15				0.94
9	0.77	0.25	0.44	0.41	0.25	0.30	0.42	0.32	0.27			0.80
10	0.63	0.21	0.35	0.34	0.17	0.24	0.40	0.27	0.20	0.43		0.91

		1	2	3	4	5	6	7	8	9	10
No.of items	**80**	5	6	8	4	4	16	8	5	18	6
Avg. test score	**42.33**	2.23	2.62	3.45	2.09	2.24	9.14	4.66	2.07	10.18	3.64
Std. deviation	**11.76**	1.83	1.47	1.81	1.16	1.13	3.92	1.61	1.62	3.29	1.66
SEM:	**4.13**	0.67	1.06	1.17	0.88	0.86	1.10	1.23	0.82	1.85	1.02
Avg. P-value:	**52.91**	44.69	43.72	43.16	52.16	55.98	57.11	58.26	41.37	56.58	60.71
Coefficient alpha	**0.88**	0.80	0.45	0.52	0.37	0.37	0.80	0.36	0.68	0.64	0.61
GLB	**0.00**	0.86	0.49	0.60	0.42	0.46	0.92	0.43	0.75	0.70	0.63
Asymptotic GLB	**0.00**	0.87	0.48	0.58	0.42	0.42	0.92	0.42	0.74	0.68	0.62

In terms of the overall facility of the test, the average *P*-values for the 2021 and 2023 administrations are 51% and 53% respectively, which are in the vicinity of the desired 50%. In terms of the various subtests and their progression from easy to more difficult, this might be improved if the two problematic subtests *Organising information visually* (subtest 7) and *Understanding text type and communicative function* (subtest 8) are removed and the remaining subtests reordered as follows: *Text editing* (subtest 10), *Text comprehension* (subtest 9), *Grammar and text relations* (subtest 6), *Academic arguments* (subtests 4 and 5), *Organising text* (subtest 1), *Understanding academic vocabulary* (subtest 2) and *Interpreting graphic and visual information* (subtest 3). Therefore, some amendments may be necessary to improve the degree to which the differentiated parts function together to satisfy the condition of technical viability.

For the sake of thoroughness and to better inform such a decision to remove the problematic subtests, a second analysis was run to see if the differentiation and functionality, as well as overall reliability of the APPMI improved if they were removed. The reduction yielded a reliability (Cronbach's alpha) score of 0.87. Furthermore, the overall facility (53%) and discrimination (0.33) values were also within the desirable range. The Rasch analysis results for degree of fit were also within the desired range, with terminal values of Infit Mean Square ranging from 0.86 to 1.19. These results were comparable to the full test; thus, for reasons of technical economy, it may therefore be worthwhile removing these two subtests for future administrations of the APPMI. The abilities measured in these subtests (*Organising information visually* and *Understanding test type and communicative function*) are still tested in various other subtests (see Table 1). It is important to record such design decisions, particularly when they relate to test development and refinement, in order to satisfy yet another design principle, that of clear technical signification. It is a principle that is fulfilled, among other things, when the blueprint and specifications of a test, that express the detail of how the test must be shaped, are determined and articulated. The same applies to the recording of technical design decisions relating to the refinement of the test.

The final consideration for this paper concerns the fairness of the test. As mentioned earlier, there was no significant difference between the pencil-and-paper and online versions of the APPMI, indicating that neither group was advantaged or disadvantaged in terms of mode of administration. Another early and quantifiable measure of test fairness is to calculate the number of candidates who may potentially have been misclassified by the test. This can be determined by means of the alpha-based, same test or parallel test [Rxx or Rxt case] results produced by the Tiaplus analysis (CITO 2005). Considering there is an equal chance of misclassification, Table 9 shows that, at worst, 6.3% of candidates may have been misclassified.

Table 9: Potential misclassifications in the administration of the APPMI (2023)

Alpha-based	
- Rxx' case: Percentage	6.3%
Number	145
- Rxt case: Percentage	4.6%
Number	105

To determine a potential cut-off for performance on the APPMI, a correlation analysis was run to measure the potential relationship between students' 2022 NBT and 2023 APPMI scores. The results of the 2021 administration indicated a strong correlation of 0.6, which was statistically significant (<0.0001). These results were replicated in the 2023 iteration, with a correlation of 0.6, which is highly significant (p-value <0.0001). That is, the regression analysis results indicate that the APPMI scores have a statistically significant (p-value<0.05) relationship with the NBT scores. The adjusted R-Square value indicates that APPMI scores account for ~33% of the variation in NBT scores. It should be noted, however, that the NBT scores are highly variable, and taking all results into consideration, the APPMI test can be considered to be a good predictor of language ability. The NBTP (National Benchmark Test Project) sets a cut-off score of 64% to determine which students are required to take literacy courses; an APPMI score of 73% predicts this cut-off score of 64% on the NBT. To accommodate potential misclassifications (6.3% per Table 9), and since there is an even chance that misclassification can either be advantageous or detrimental, a second test opportunity should be afforded only to 3.15% (or 73 students), with that calculation starting at the 73% cut-off point. i.e., the first 73 students with scores below 73% would qualify. However, it should be noted that second opportunities do not necessarily involve the same test, since that could only be administered six months afterwards, and might therefore be impractical and a potential limitation. A construct equivalent test, with the same specifications as the first, may provide a possible solution.

6. Conclusion

Challenges brought about by the COVID-19 pandemic regarding the administration of widely used tests of academic literacy necessitated alternative means of identifying students in need of literacy support. The results of an institutional pilot of an in-house test of academic literacy (APPMI) deemed it appropriate as a diagnostic tool for incoming undergraduate students. A further administration of the test was necessary to determine whether the results of the previous online pilot could be replicated in a traditional (pencil-and-paper) format and to further justify the quality of the test's design. The methodological tools discussed in this paper are useful in the initial phases of what is commonly referred to as test validation, which has been described here as a process of demonstrating responsible design. Based on the results of CTT and Rasch analyses, the APPMI has generally met the various principles of test validity and could thus be considered as a possible in-house alternative to conventional academic literary tests, such as the NBT. However, the principles investigated above constitute but an early start of test validation; not every principle of the framework for responsible design (Weideman, 2017) has been examined here. Further justification is required, for example, in terms of construct, face validity, alignment with language policies and instruction, and reputability (Weideman, 2020).

As regards comparisons between the two administrations of the test, this study has concluded that the mode of administration does not affect the performance of the test, as there was no statistically significant difference between the results of the online and traditional (pencil and paper) iterations.

An important first step in the comparison of the two administrations was to determine whether the difference in results of the online and traditional iterations was statistically significant. A t-test was done to determine the difference between the 2023 online and 2023 traditional; 2021 (online) and 2023 traditional; and 2021 (online) and 2023 combined (online and traditional) versions. The results in Table 4 show that there was no significant difference (p=0.6519) between the results of the 2023 online and traditional iterations of the test.

Important to note with any research on large-scale language and literacy testing are the limitations of such generic tests in credibly determining all students' academic literacy needs across all disciplines. Additional sources of information are necessary to ensure that literacy interventions intended to address these needs are appropriate, accessible, theoretically defensible, useful, and potentially more effective (Sebolai, 2022a). These additional sources include lecturer perceptions of students' needs, students' perceptions of their own needs, and in-depth analyses of genre- and discipline-specific text types to better understand disciplinary conventions. Furthermore, collaborative work with lecturers on the design of academic literacy curricula is of vital importance (Jacobs, 2005, 2007a, 2007b, 2010; Clarence, 2011; Wingate, 2015; Dison & Moore, 2019; Bond, 2020). It is, therefore, necessary to supplement the results of the APPMI, should it be administered institutionally, with these additional sources of information to work towards aligning the test as a diagnostic tool and the institutional literacy support on offer (Weideman, 2019; Sebolai, 2022b).

References

Bond, B. (2020). *Making language visible in the university: English for academic purposes and internationalisation.* Multilingual Matters.

Boughey, C. (2013). What are we thinking of? A critical overview of approaches to developing academic literacy in South African higher education. *Journal for Language Teaching 47*(2), 24–42. https://hdl.handle.net/10520/EJC148781

CITO. (2005). *TiaPlus user's manual*. M & R Department.

Clarence, S. (2011). Writing in the academy. In A. Archer, & R. Richards (Eds.), *Changing spaces: Writing centres and access to higher education,* Sun Press.

Dison, L., & Moore, J. (2019). Creating conditions for working collaboratively in discipline-based writing centres at a South African university. *Per Linguam 35*(1), 1–14. https://doi.org/10.5785/35-1-851

Drennan, L. (2019). Defensibility and accountability: Developing a theoretically justifiable academic writing intervention for students at tertiary level [Doctoral thesis, University of the Free State]. KovsieScholar Repository. 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 A. Weideman, J. Read, & T. du Plessis, (Eds.), *Assessing academic literacy in a multilingual society: Transition and transformation* (pp. 196–216). Multilingual Matters.

Drennan, L., Joubert, M., Weideman, A., & Posthumus, R. (2021). Combined measures of identifying language risk for first-year students. *Journal of Language Teaching 55*(2), 93-116. https://doi.org/10.4314/jlt.v55i2.4

Frederiksen, C. H. (1975). Representing logical and semantic structure of knowledge acquired from discourse. *Cognitive Psychology 7*, 317–458. https://doi.org/10.1016/0010-0285(75)90016-X

Jacobs, C. (2005). On being an insider on the outside: New spaces for integrating academic literacies. *Teaching in Higher Education 10*(4), 475–487. https://doi.org/10.1080/13562510500239091

Jacobs, C. (2007a). Towards a critical understanding of the teaching of discipline-specific academic literacies: Making the tacit explicit. *Journal of Education 41*, 59–81. https://hdl.handle.net/10520/AJA0259479X_23

Jacobs, C. (2007b). Mainstreaming academic literacy teaching: implications for how academic development understands its work in higher education. *South African Journal of Higher Education 21*(7), 870–881. https://hdl.handle.net/10520/EJC37399

Jacobs, C. (2010). Collaboration as pedagogy: consequences and implications for partnerships between communication and disciplinary specialists. *Southern African Linguistics and Applied Language Studies 28*(3), 227–237. https://doi.org/10.2989/16073614.2010.545025

Joubert, M., Larsen, A., Magnuson, B., Waldron, D., Sabo, E., & Fletcher, A. (2023). Global challenges: South African and Australian students' experiences of emergency remote teaching. *Journal for University Teaching & Learning Practice 20*(4). https://doi.org/10.53761/1.20.4.09

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

Mihindou, G. R. (2019). Language and academic literacies development at the University of Johannesburg. In C. L. Scott, & E. N. Ivala (Eds.), *Transformation of higher education institutions in post-apartheid South Africa*. Routledge.

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

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

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

Sebolai, K. (2022a). Determining the faculty specific academic literacies needs of first year university students: A mixed methods approach. *Journal for Language Teaching 56*(2), 1-20. https://doi.org/10.56285/jltVol56iss2a5702

Sebolai, K. (2022b). The Academic Development approach to academic literacy in higher education South Africa: A disconnect between teaching and assessment. *Journal for Language Teaching*, *56*(2), 1–16. https://doi.org/10.56285/jltVol56iss2a5387

Sebolai, K., & Stanford, F. (2020). Validating the highest performance standard of a test of academic literacy at a South African University. *Per Linguam, 36*(2), 76–89. https://hdl.handle.net/10520/ejc-perling-v36-n2-a7

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

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

Van Dijk, T. A. (1979). Relevance assignment in discourse comprehension. *Discourse Processes 2*, 113–126. https://doi.org/10.1080/01638537909544458

Van Dyk, T. (2021). Postscript: What the data tell us: An overview of language assessment research in South Africa's multilingual context. In A.Weideman, J. Read,. & T. du Plessis (Eds.), *Assessing academic literacy in a multilingual society: Transition and transformation* (pp. 217–235). Multilingual Matters.

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

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

Weideman, A. (2019). Definition and design: Aligning language interventions in education. *Stellenbosch Papers in Linguistics (SPIL) Plus*), *56*(1), 31–46. https://doi.org/10.5842/56-0-782

Weideman, A. (2020). Complementary evidence in the early-stage validation of language tests: Classical Test Theory and Rasch analyses. *Per Linguam*. *36*(2), 57–75. https://hdl.handle.net/10520/ejc-perling-v36-n2-a6

Weideman, A. (2024). *A theory of applied linguistics: Imagining and disclosing the meaning of design*. Springer. https://doi.org/10.1007/978-3-031-67559-1

Wingate, U. (2015). *Academic literacy and student diversity:Tthe case for inclusive practice,* Multilingual Matters, Bristol.

ABOUT THE AUTHORS

Laura Drennan

University of the WItwatersrand

Email: laura.drennan@wits.ac.za **ORCID**: https://orcid.org/0000-0002-9416-2590

Laura Drennan is currently a lecturer in the Division of Languages, Literacies and Languages at the Wits School of Education. She previously served as the Head of the Writing Centre at the University of the Free State, where she specialised in academic writing development. She obtained her doctorate in English Language Studies and academic literacy development. Her research interests include language teaching and learning, academic literacy, academic writing development, and language testing.

Michelle Joubert

University of the Free State, South Africa

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

Michelle Joubert is the head of the Academic Language and Literacy Development at the Centre for Teaching and Learning, University of the Free State. Previously, she was an academic literacy curriculum coordinator at Durham University, UK and a Fulbright scholar. Her areas of expertise include academic literacies, academic writing and reading development, curriculum design, professional identity and positionality, multilingualism, and language teaching and learning. She has a doctoral degree in English, specialising in academic literacy development.

Albert Weideman

University of the Free State, South Africa

Email: albert.weideman@ufs.ac.za **ORCID**: https://orcid.org/0000-0002-9444-634X

Albert Weideman is Professor of Applied Language Studies at the University of the Free State. He is the chairperson of the Inter-institutional Centre for Language Assessment and Development (ICELDA), and the deputy chairperson of the Network of Expertise in Language Assessment (NExLA). His research focuses on how language assessment, course design, and policy relate to a theory of applied linguistics.