International Journal of TESOL Studies (2024)
Vol. 6 (2) 24-42 https://doi.org/10.58304/ijts.20240203

Article

Validating Pecorari, Shaw and Malmström's (2019) Academic Vocabulary Test – Form 1: Evidence from Rasch-based Analyses and Retrospective Focus-group Interviews

Chi Duc Nguyen*
Hanh Thi Hoang
VNU University of Languages and International Studies
Vietnam National University – Hanoi, Vietnam

Received: 16 May, 2023/Accepted: 13 January, 2024/Published: 25 February, 2024

Abstract

This study aimed to assess the construct validity of Pecorari, Shaw and Malmström's (2019) Academic Vocabulary Test – Form 1 (AVT1). To this end, it first employed Rasch-based statistical evidence generated from the test responses of 989 high-school and university students in Vietnam to inspect five major aspects of Messickian construct validity: Content, Substantive, Structural, Generalizability and External (Messick,1995). It then moved on to thematically analyze the data collected from the follow-up focus-group interviews with 50 students randomly selected from the test-takers to detect any emerged patterns in their actual engagement with the test. Results from Rasch-based analyses showed that AVT1 could sufficiently measure the target ability – receptive academic vocabulary breadth – of an overwhelming 889 out of the 989 test-takers (90%) and there were seven statistically distinct groups of item difficulty in the empirical item hierarchy. In general, test items and test-takers performed as predicted by a priori hypotheses and displayed good fit to the Rasch model. Principal Component Analysis indicated that the test items formed a fundamentally unidimensional construct, suggesting that the test only measured one meaningful dimension, presumably the receptive academic vocabulary breadth. The evidence for the invariance in item calibration and person measure as well as the test's external reliability enabled the generalization of score properties and interpretations across populations, settings, and tasks. Results from the interview data analyses revealed that AVT1 did allow some room for guessing effects, which was predicted by the test developers. These findings altogether suggest that this test can be a useful tool to gauge receptive academic vocabulary breadth.

Keywords
Academic Vocabulary Test, breadth of academic vocabulary knowledge, test validation, Messickian construct validity, Rasch-based analysis

*Corresponding author. Email: ducnc@vnu.edu.vn

1. Introduction

English academic vocabulary plays a critical role in the language use (e.g., Cobb & Horst, 2004; Dang & Webb, 2014; Nagy & Townsend, 2012) and the general academic performance (e.g., Goldenberg, 2008; Jacobs, 2008) of both high-school and university students who attend education programs in English as a Medium of Instruction (EMI) environments. This is true not only for English as a Second Language (ESL) but also first language (L1) English-speaking students (Gardner & Davies, 2014; Masrai & Milton, 2018). Therefore, stakeholders in K-12 and tertiary education need to have a well-designed and well-validated academic vocabulary test to foster their teaching, learning, and researching practices (Nation, 2001; Read, 2000, 2007). This need was first highlighted one decade ago by Lin and Morrison (2010) as they found "there is no one commonly accepted standard test of academic vocabulary" (p.257). This, however, still holds true today.

In fact, the existing literature has witnessed the presence of some tests or, to be more precise, some test parts that assess the breadth of academic vocabulary knowledge such as the Vocabulary Levels Test (Nation, 1983; 1990), the two updated versions of the Vocabulary Levels Test (Schmitt, Schmitt & Clapham, 2001) and the Listening Vocabulary Levels Test (McLean, Kramer & Beglar, 2015). However, these tests are all limited in three ways. First of all, none of these tests treat academic vocabulary knowledge as an independent construct, but as an embedded trait in a larger construct of general vocabulary knowledge (Pecorari, Shaw & Malmströ̈m, 2019; Schmitt, Schmitt & Clapham, 2001), which makes the test score interpretation far from straightforward. Second, the numbers of sampled academic vocabulary items in these tests are so small (only ranging from 18 to 30 items) that ceiling effects (i.e., when the maximum test score is reached) have been reported several times in previous research (e.g., Edgarsson, 2018; Skjelde & Coxhead, 2020). Finally, all the tests mentioned above sample their test items from academic word lists that use word families, not lemmas, as the counting unit, such as Xue and Nation's (1984) University Word List or Coxhead's (2000) Academic Word List. Consequently, they run the risk of inflation in score interpretation, especially when test-takers do not have sufficient inflectional and derivational knowledge (see Brown et al. (2021) for a detailed discussion on this issue).

To address the aforementioned limitations, Pecorari, Shaw and Malmströ̈m (2019) developed a new academic vocabulary test (AVT) with two equivalent forms – Form 1 and Form 2 – which both aim to measure the breadth of academic vocabulary knowledge. Each form of the test sampled 57 target lemmas, together with another 57 lemmas working as distractors, from Gardner and Davies's (2014) Academic Vocabulary List and used the same written multiple-choice form-meaning matching test format as in Schmitt, Schmitt and Clapham's (2001) Vocabulary Levels Test. Regarding test validation, Pecorari, Shaw and Malmström (2019) provided some preliminary evidence for the test quality, which, however, only focused on item facility, item discrimination and the correlation between the test scores and the COCA-based frequency of the test items. Therefore, this test is in need of further validation. The present study aims to address this need by inspecting the construct validity of Form 1, using evidence from Rasch-based analysis and from retrospective focus-group interview data analysis.

2. Literature Review

2.1 Early endeavors to measure academic vocabulary breadth

2.1.1 Nation's (1983, 1990) Vocabulary Levels Test

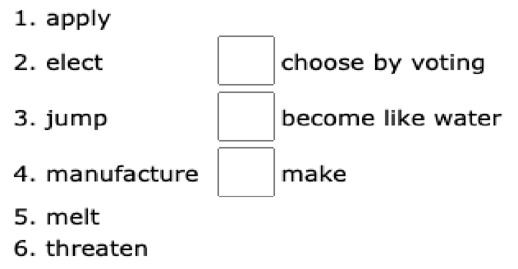
The first endeavor to measure academic vocabulary breadth was made by Nation (1983, 1990) with his Vocabulary Levels Test. This test was originally developed as a simple diagnostic test for English language teachers in New Zealand to estimate the vocabulary size of second-language (L2) English-speaking immigrants or international students who were seeking short-term study or long-term residence in that country. However, due to the lack of a standardized vocabulary test at that time and even till the end of the twentieth century, this test has been widely used both by vocabulary practitioners and researchers far beyond this context (Read, 2000).

In Nation's Vocabulary Levels Test, the construct of vocabulary knowledge was operationalized as the ability to recognize a word meaning from a given word form. Vocabulary breadth was operationalized by random sampling of 18 target words from a pool of a frequency-based 1000-word level in Thorndike and Lorge's (1944) English word list and in Campion and Elley's (1971) specialized university word list. Word families were used as the counting unit in these corpora and so were they in Nation's Vocabulary Levels Test.

This test comprised five parts, representing five different frequency-based 1000-word levels in the aforementioned corpora: 2000-, 3000-, 5000-, 10,000-word Level and University Word Level. A form-meaning matching format was employed for this measure. For each test item, the written L2 word forms of three target words and three distractors were presented on the left-hand side, while the written L2 definitions of these three target words were put on the right-hand side. Test-takers were required to match the word form to the correct word meaning. Below is one example of these test items:

Figure 1

Test items in Nation's Vocabulary Levels Test



1. apply
2. elect □ choose by voting
3. jump □ become like water
4. manufacture □ make
5. melt
6. threaten

Read (1988) was the first to validate this test. He administered this test to a group of 81 ESL students before and after a three-month English for Academic Purposes course in New Zealand. He then moved on to test its underlying hypothesis as to whether there existed a positive correlation between the test scores and the corpus-based frequency of the test items. In both testing times, this hypothesis was supported, but not for the case of the University Word Level. Finally, he ran a Guttman scalogram analysis for the test scores to generate the coefficient of scalability. According to Hatch and Farhady (1982), this coefficient of scalability should be well above .60 for the test scores to reflect a truly implicational scale. For the first and second testing time, Read found a coefficient of .90 and .84, respectively.

Another test validation was made by Beglar and Hunt (1999). First, the two original test levels – the 2000-word Level and the University Word Level – were revised into four different new forms, two of which were subsequently given to 496 high-school and 464 university students in Japan. Results from Rasch-based analyses showed that both test forms were basically unidimensional. Only three test items displayed misfit to the Rasch model, which adversely affected as many as 108 test-takers. The internal reliability coefficients of these two test forms were .84 and .95. The test scores were also found to positively correlate with the TOEFL test scores, especially those in the reading and the grammar section.

2.1.2. Schmitt, Schmitt and Clapham's (2001) Vocabulary Levels Test

To improve the quality of Nation's (1983, 1990) Vocabulary Levels Test, Schmitt, Schmitt and Clapham (2001) substantially revised this test into two updated versions. They retained the purpose, underlying construct, construct operationalization, method, scoring protocol and score interpretation of the original test, but made four major changes. First, the number of test items at each level was increased from 18 to 30. Second, the lexical items for the 2000-word Test Level were sampled from both the first and the second 1000-word level in the aforementioned corpora, at a 12:18 instead of the original 6:12 ratio. Next, the source for the academic lexical items was changed from Campion and Elley's (1971) specialized university word list to Coxhead's (2000) Academic Word List. To enhance the representativeness of the sampled vocabulary items, Schmitt, Schmitt and Clapham (2001) also strictly followed a new ratio of 3 nouns – 2 verbs – 1 adjective for each test level. Finally, test-takers would get one point for every correct answer and they needed to obtain at least 26 out of the maximum score of 30 points at each level to be considered to have mastered that level.

These two test versions were then administered to a group of 801 test-takers, both L1 and L2 English speakers for test validation. The test scores for L1 English speakers showed that these test-takers often obtained the maximum or almost maximum test score (*M* = 309; maximum score = 312). Therefore, these tests posed no or little problem to (educated) L1 English speakers. As for the whole dataset, the results from classical item analysis indicated that the mean item facility score for each test level decreased when the corpus-based frequency of the test items decreased, supporting the hypothesized positive correlation between these two variables. Second, the mean item discrimination index for each level ranged from .51 to .67, suggesting that the test items displayed sufficient spread in terms of item difficulty. Further item analyses revealed that the test-takers strictly followed the test instruction and rarely made blind guesses during their test completion. Like Beglar and Hunt (1999), Schmitt, Schmitt and Clapham (2001) also examined whether different test levels altogether formed an implicational scale; however, they excluded the Academic Vocabulary Level from their final analysis based on the ground that there might be some overlapping in the corpus-based frequency between the general and academic vocabulary test items. After the exclusion above, they did obtain a far higher coefficient of .993 for the first test version and .995 for the second test version compared to that of .90 and .84 in Read (1988).

Principal Component Analysis (PCA) was also conducted for each test level to examine whether the two test forms were unidimensional or involved the measurement of a second meaningful dimension. In all test levels, including that of academic vocabulary, most of the variance in the test scores was explained by the primary dimension extracted by Rasch's PCA (which was presumed to be the test construct) with the explanatory power ranging from 95% to 98% of the variance in the test scores. When data of all test levels were combined into a unified PCA analysis, however, the explanatory power decreased to 78% and the remaining variance in the test scores might be explained by other factors which altogether could form a second meaningful dimension in item residuals. Test unidimensionality, however, might not be dependent on the amount of variance in the test scores accounted for by the test construct itself, but by the size and interpretation of residual contrasts (McLean, Kramer & Beglar, 2015). Unfortunately, Schmitt, Schmitt and Clapham (2001) did not look at the individual loadings of each test item, especially those with the absolute loadings of .40 and above to provide further evidence in relation to the test's unidimensionality.

Cronbach's alpha was also computed to examine the internal reliability of each test level in the two test forms above. These coefficients were all high, ranging from .92 to .96, especially for the two academic vocabulary sections with .95 and .96. Results from follow-up interview data analyses showed that test-takers rarely made blind guesses during their test completion. Most of the mismatches between their actual test scores and their interview data might be because of their partial vocabulary knowledge rather than their blind guessing. To provide evidence for the equivalence of the two test forms, Schmitt, Schmitt and Clapham (2001) compared the mean test scores and the score variances at each test level between the two test forms and for the whole population of the test-takers. No significant difference was found. However, when this sample was broken down by their first language, the test equivalence was not consistent for all test levels, except for the case of the 2000-word Level. As for the academic vocabulary section, the test scores did not differ between the two test forms among Romance language speakers, but they did so among non-Romance and Slovak language speakers.

2.1.3 McLean, Kramer and Beglar's (2015) Listening Vocabulary Levels Test

To address the need for a standardized aural vocabulary test, McLean, Kramer and Beglar (2015) constructed their Listening Vocabulary Levels Test. This test was conceptualized as "a diagnostic and achievement instrument that measures knowledge of English lexis from the first five 1000-word frequency levels and the Academic Word List (Coxhead, 2000) for either pedagogical or research purposes" (p.743). The receptive form-meaning connection of aural, but not orthographic, English vocabulary was used to operationalize the test construct. To measure the breadth of aural English vocabulary knowledge, a total of 120 lexical items were randomly selected from the first to the fifth frequency-based 1000-word lists in Nation's (2012) British National Corpus/Corpus of Contemporary American English (BNC-COCA) (24 lexical items per level) and another set of 30 lexical items was also randomly chosen from Coxhead's (2000) Academic Word List as the target words. The resulting test, therefore, consisted of 150 test items in total. As all word lists above used word families as the counting unit, this counting unit was also employed for the target words in the test.

McLean, Kramer and Beglar (2015) borrowed Nation and Beglar's (2007) Vocabulary Size Test's multiple-choice form-meaning matching format for their present test. For each test item, test-takers were asked to listen to the target word first in isolation and then in a non-defining sentence context. After that, they needed to select the correct L1 meaning for the target word out of four test options which were all presented in the written form on the test paper. Below is one example of these test items.

Figure 2

Test items in McLean et al.'s (2015) Listening Vocabulary Levels Test

[Examinees hear: 'School: This is a big school.']

(1) a.銀行 (bank)
b.海の動物 (sea animal)
c.学校 (school)
d.家 (house)

Note: Test-takers did not see or hear the English translation of the test options within brackets.

This test was then carried out with a group of 214 Japanese university students to validate its construct validity. These test-takers were granted one point for a correct answer. No cut-off point for the level mastery was reported in their test manual. Results from Rasch-based analyses showed that the test-items had such sufficient spread of item difficulty that they could measure the target ability of 207 test-takers (96.7%). Almost all items (145 items, about 97%) and test-takers (207 persons) performed as predicted by *a priori* hypotheses and displayed good fit to the Rasch model. Evidence from Principal Component Analysis suggested that the test was basically unidimensional. Split-half reliability check indicated that the test had good internal consistency. A moderate coefficient of .54 was found for the correlation between the test scores above and those of TOEIC listening tests. Results from analyzing follow-up interview data also provided supportive evidence for the face validity as well as the interactiveness of the Listening Vocabulary Levels Test.

The review above clearly shows that tests of receptive vocabulary breadth to date have treated academic vocabulary knowledge as an embedded trait in a larger construct of general vocabulary knowledge. In addition, the overlap in terms of corpus-based frequency between academic and general lexical items in those tests led to difficulties with score interpretation. The number of the academic lexical items in these tests were also relatively small, ranging from 18 (Nation's Vocabulary Levels Test) to 30 items (Schmitt, Schmitt and Clapham's updated Vocabulary Levels Test or McLean, Kramer and Beglar's Listening Vocabulary Levels Test). It is thus unsurprising that ceiling effects have been reported several times in the existing literature (e.g., Edgarsson, 2018; Skjelde & Coxhead, 2020). The most serious limitation of these tests, however, is that they all used word families, not lemmas, as the counting unit, which could lead to an overestimation of test-takers' academic vocabulary knowledge if the test-takers did not have sufficient knowledge of relevant word derivation and inflection. Such overestimation might misguide academic vocabulary instructors' and researchers' educational decisions and recommendations, respectively (Brown et al., 2021).

2.2 Pecorari, Shaw and Malmström's (2019) new Academic Vocabulary Test

To address the limitations of all vocabulary tests reviewed above, Pecorari, Shaw and Malmström (2019) developed two versions of a new Academic Vocabulary Test – Form 1 and Form 2. This test aimed to measure "receptive knowledge (form-meaning) rather than productive, and breadth rather than depth" of academic vocabulary knowledge (Pecorari, Shaw & Malmström, 2019, p. 5). To operationalize the construct of vocabulary knowledge, they resorted to the word form-meaning connection. Although there is consensus in our field that vocabulary knowledge, including academic vocabulary knowledge, involves more than merely the word form-meaning relationship (e.g., Boers, 2021; Nation, 2001; Read, 2000), this form-meaning link is still regarded as the most important among all aspects of vocabulary knowledge (e.g., Laufer & Goldstein, 2004; Nation, 2001), especially when it comes to the case of measuring the breadth, but not the depth of vocabulary knowledge (Read, 2000). As a result, it is not surprising that all the tests reviewed above have used this aspect of vocabulary knowledge as the primary latent trait in their measures.

The breadth of academic vocabulary in this test was operationalized by stratified random sampling of lexical items from Gardner and Davies's (2014) Academic Vocabulary List. Specifically, these test developers first divided Gardner and Davies's (2014) Academic Vocabulary List (AVL) into 30 frequency-based 100-lemma sets and then randomly selected two target words and two respective distractors from each set. Another screening process was then carried out to remove any word whose (a) morphology was identical to that of another, but not in the AVL (e.g., *study* (*n*) and *study* (*v*)), (b) meaning could be easily deduced by knowledge of word parts alone (e.g., *dissatisfied* and *satisfied*), (c) meaning was too close to that of another selected word in the same set (e.g., *paraphrase* and *reformulate*) and (d) meaning was difficult to define. The resulting list for each AVT test form consisted of 57 target lemmas, together with another 57 lemmas working as distractor test items. This word selection was well-justified for at least three main reasons. First, compared to other academic word lists, Gardner and Davis's (2014) AVL is the only word list that uses lemmas, not word families, as the counting unit. From a vocabulary assessment perspective, this helps minimize inflation in score interpretation, especially when test-takers did not have sufficient derivational and inflectional knowledge. In addition, the AVL was constructed based on a 120-million-word academic sub-corpus of the 425-million-word Corpus of Contemporary American English – COCA, which was by far larger in size, more updated and covered more academic disciplines than earlier corpora. Using much stricter ratio, frequency, and dispersion statistics in its construction process, the AVL was also found to be abler to separate its "*core*" academic words from "*general high-frequency words*" (*cf.* Dang, Webb and Coxhead's (2017) Academic Spoken Word List) as well as from "*academic technical words*" (Gardner & Davis, 2014, p.312) (*cf.* Lu and Coxhead's (2020) Traditional Chinese Medicine Word List or Coxhead and Demecheleer's (2018) Technical Vocabulary of Plumbing).

Pecorari, Shaw and Malmström (2019) also used the same written multiple-choice form-meaning matching test format from the previous tests as in Nation's original Vocabulary Levels Test (1990) or Schmitt, Schmitt, and Clapham's updated Vocabulary Levels Test (2001). In the scoring procedure, test-takers were also awarded one point for each correct answer. To validate the AVT, Pecorari, Shaw, and

Malmström (2019) administered the two test forms to a group of 455 EFL university students. They examined item facility, item discrimination, test reliability as well as the equivalence between the two test forms. Results from item analyses respectively indicated a mean item facility and item discrimination score of .73 and .43 for Form 1 and .72 and .46 for Form 2. These together suggested that the test items displayed sufficient spread of difficulty and were able to discriminate the target ability of test-takers into statistically distinct groups. To test the hypothesized positive correlation between the test scores and the COCA-based frequency of the test items, a Pearson correlation analysis was then run for the two variables. A coefficient of .54 was found. The same Cronbach's alpha of .91 was also reported for the internal reliability of the two test forms. No significant difference was found between the mean test scores of the two test forms, suggesting that they were equivalent in terms of difficulty.

Pecorari, Shaw, and Malmström (2019) also carried out one-to-one interviews with 14 test-takers. The purpose of these interviews was twofold. On the one hand, the interview data would help triangulate the quantitative data collected from the test. On the other hand, these test developers could also further explore the extent to which blind guesses (if any) influenced the test scores. They found out that, in 43 out of the 294 cases (14.6%), the interviewees' oral responses differed from their written ones. Further qualitative data analyses suggested that the differences might be due to blind guesses where test-takers did not know the word meanings. Fortunately, most of their blind guesses were incorrect.

As shown in the review above, the AVT is potentially a useful measure of academic vocabulary breadth. However, since the test validation above only focused on item facility, item discrimination, and test validity, it cannot guarantee that the test served its intended purpose. Therefore, a more thorough validation process was required and the present study contributed towards filling this gap.

3. Present Study

3.1 Research aim

This study aimed to further validate Pecorari, Shaw, and Malmström's AVT. It assessed the construct validity of AVT's Form 1. To this end, it first used Messickian construct validity (1995) as the conceptual framework for test validation. Messickian construct validity has been widely accepted as a useful framework for validating language tests (e.g., Bachman & Palmer, 1996; McNamara, 2006; Read & Chapelle, 2001) and a vocabulary test (e.g., McLean, Kramer & Beglar, 2015; Schmitt, Schmitt & Clapham, 2001). The statistical evidence from Rasch-based analyses was then employed to inspect five major aspects of Messickian construct validity of the AVT – Content, Substantive, Structural, Generalizability, and External (1995)[1]. The final aspect of Messickian validity – Consequential – was not examined herein as this went far beyond the scope of the present study. Finally, follow-up focus-group interviews were carried out with some test-takers to examine their actual engagement with the test, especially their guessing behaviors – an emerging pattern that Pecorari, Shaw, and Malmström (2019) found in their test validation. The combination of quantitative and qualitative evidence was expected to better substantiate the construct validity of the AVT.

3.2 Research participants

Research participants were 989 high-school[2] (*n* = 531; 209 males and 322 females) and university (*n* = 458; 21 males and 437 females) students recruited from three cities/provinces – Hanoi (the capital city), Bac Giang (a rural province) and Lang Son (a remote mountainous province) – with different socio-economic conditions in Vietnam. The high-school students were 16 to 18 years old (*M* = 16.75; *SD* = 0.75) and the university students were between 20 and 23 of age (*M* = 21.02; *SD* = 0.82). The former was coded in the present study as HS001 to HS531, while the latter as US001 to US458.

According to the benchmark for foreign language proficiency level in Vietnam, high-school and university students needed to achieve Level 2 and Level 3 on the Six-level Foreign Language Proficiency Framework for Vietnam or Level A2 and B1 on the Common European Framework of Reference for Languages (CEFR) as a requisite for their high-school and tertiary-education enrolment, respectively. The scores from their CEFR-based in-house placement tests, however, showed that while English language proficiency of the high-school students widely ranged from Level A2 to C1, that among the university students tended to be more consistent, mostly between Level B2 and C1 and for some cases Level C2. These research participants included both English-majored and English-non-majored students. They all confirmed in their consent form that they agreed to participate in this study as well as provide their personal information, test, and interview data for the research purpose and that they had not seen the AVT before.

3.3 Data collection and data analysis

Data for this study came from two sources. One was from the test scores of the 989 test-takers above. They all took AVT Form 1 in their classroom settings and for 15 to 20 minutes at most. They were allowed to make guesses during their test completion, but no reference to other available resources (e.g., a dictionary, a peer, or a language teacher) was allowed. The test responses of the 989 test-takers above were then run through Winsteps Version 5.4.3 to solicit relevant Rasch-based statistical evidence.

Right after the test, 50 test-takers (25 high-school students and 25 university students) were randomly recruited and then divided into 10 different groups of five test-takers for retrospective focus-group interviews. The present study used focus-group, but not one-to-one interviews for two reasons. First, they allowed ideas from several interviewees to be collected at the same time. Second, the interaction between the interviewees within the same group might indicate whether a specific behavior in their test engagement was shared among themselves, which in turn fostered the theme detection in the following data analysis.

In these interviews, the focus was placed on the test-takers' guessing practices (if any). Probing questions in these interviews centered around five common vocabulary test-taking strategies proposed by Paul, Stallman and O'Rourke (1990), Schmitt, Ng and Garras (2011) and Gyllstad, Vilkaite and Schmitt (2015). These strategies included: (a) guessing from word parts; (b) guessing from knowing another member of a word family; (c) guessing from elimination and association; (d) guessing from co-textual clues (if any) and (e) blind guessing. The final strategy – blind guessing – was further broken down into choosing the option (e1) that is the most uncommon; (e2) that is the shortest or longest; (e3) that is understood best, or (e4) that just sounds the best.

Interviewees could choose either to use English or Vietnamese or both to best convey their ideas. Before each interview, their test papers were given back to them so that they could refer to any specific test items to illustrate their points in their responses to the probing questions. Since the purpose of these interviews was not to triangulate the validity of their test scores, but to examine their guessing behaviors (if any), no test items were pre-selected. Any reference to a specific test item was initiated by the interviewees themselves. All interviews were audio-recorded and the interviewer (also the primary author of this paper) also took note of details that audio-recording failed to capture.

The total length of the resulting recordings was 5 hours and 11 minutes, with an average of about 31 minutes per group interview. These audio-recordings were then transcribed and inductively analyzed by two independent experienced qualitative researchers who used the same coding scheme of the eight vocabulary test-taking strategies above first to identify the instances of these strategies in the written transcript and then to tally the frequency of each strategy. Any strategy that was mentioned by at least three test-takers in a group interview and at least three groups in each population (i.e., high-school students and university students) qualified as a theme which would then be included in the research findings. In this way, the representativeness of the theme could be enhanced. After independently coding the transcripts, these coders worked together and compared their coding results. There were four differences between these two coders, which were resolved through discussions.

4. Findings

4.1 Rasch-based evidence

4.1.1 Content aspect

The content aspect of Messickian construct validity concerns whether a language test measures what it intends to measure and nothing else. To this end, test developers need to ensure that all test contents should tap into the underlying knowledge, skills, and other attributes used to define and operationalize the test construct (i.e., content relevance). In case the test, for practical reasons for example, fails to cover all knowledge, skills, and attributes above, the selected test contents should be representative of such knowledge, skills, and attributes (i.e., content representativeness). In addition, the process of item writing and scoring should avoid introducing construct-irrelevant factors into the measure (i.e., technical quality). Therefore, to examine the content aspect of Messickian construct validity, one needs to gather supporting evidence for the content relevance, the content representativeness as well as the technical quality of the test being validated (Messick, 1995).

Regarding content relevance, the AVT appears to satisfy this criterion as already discussed in Section 2.2. Turning to content representativeness, it is obviously impractical to measure the vocabulary knowledge for all 3,015 lexical items in Gardner and Davies's (2014) Academic Vocabulary List in a single test. Therefore, the AVT resorts to gauging the receptive form-meaning connections of only 57 target lemmas, together with another 57 lemmas used as respective test distractors, from the list above as a representative sample of the test construct. To examine the representativeness of this selected test content, evidence needs to be gathered to answer four critical questions: (a) Is a sufficient number of test items included in the AVT? (b) Does the empirical item hierarchy show sufficient spread? (c) Are there any noticeable gaps in the above empirical item hierarchy? and (d) Is there a sufficient number (at least 2) of statistically distinct item difficulty levels in the dataset? This evidence can first be drawn from the Wright Map displayed in Figure 3 below.

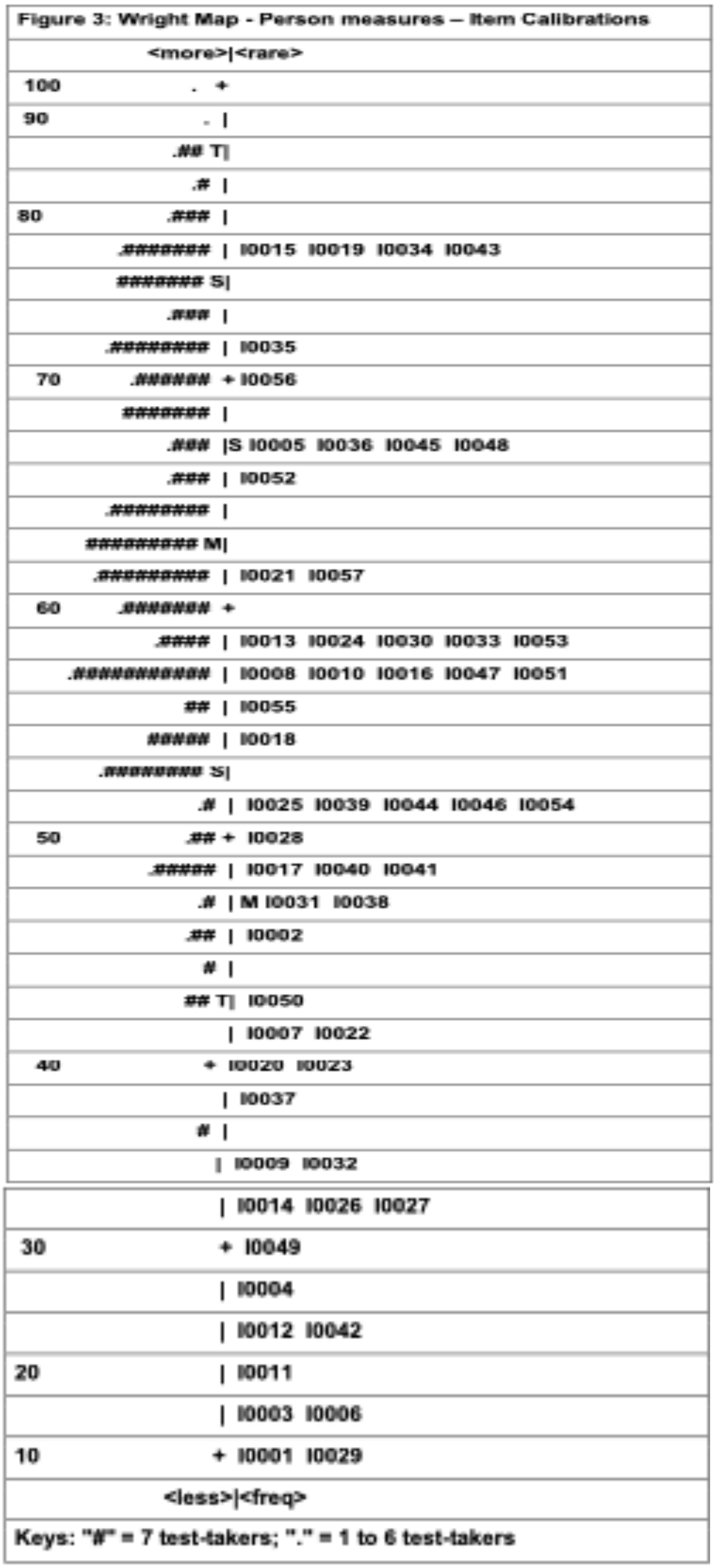
This Wright Map (or also known as the Person-Item Map) showed the linear relationship between the Rasch calibrations for the 989 test-takers (on the left-hand side) and the 57 test items (on the right-hand side). In this map, measurement unit was the Rasch logit, which, however, was then transformed into CHIPS for easier interpretation (Smith, 2000). More able test-takers and more difficult test items were located toward the top of the figure, while less able test-takers and less difficult test items were placed toward the bottom of the same figure. A "#" in the above map represented 7 test-takers and a "." did so for 1 to 6 test-takers.

On the above CHIPS scale, the difficulty levels of the 57 test items in AVT1 ranged between -22.53 (Item 29) and 76.29 CHIPS (Item 19). Meanwhile, the target abilities (i.e., the receptive academic vocabulary breadth) of the 989 test-takers ranged between 37.57 (Test-takers 2, 152, 304, 478, 630, 789 and 940) and 111.43 (Test-takers 126 and 763) CHIPS. AVT1 was found to sufficiently measure the receptive academic vocabulary breadth of 90% (or to be more precise 889) of the above test-takers. While no flooring effect (i.e., when test-takers obtained absolutely no score for the whole test) was found, the ceiling effect occurred twice (Test-takers 126 and 763). A few gaps did exist in the empirical item hierarchy, especially in the region with the item difficulty of below 35 and above 80 CHIPS.

These together suggest that although AVT1 could measure the receptive academic vocabulary breadth for the vast majority of test-takers, more items might be needed, especially those with difficulty of 80 CHIPS onwards. To examine the number of statistically distinct item difficulty levels that the measure could create in the dataset, Item Strata was used instead of Item Separation since the outliers in the test items were not normally distributed, but heavily tailed. In this case, Item Strata was 7.76, suggesting that there existed almost eight groups of test items in AVT1 which had distinct difficulty levels.

Figure 3

The Wright Map



```
Figure 3: Wright Map - Person measures – Item Calibrations
              <more>|<rare>
 100                . +
 90                 . |
                  .## T|
                   .# |
 80               .### |
             .####### | I0015  I0019  I0034  I0043
             ####### S|
                  .### |
            .######## | I0035
  70         .###### + I0056
             ####### |
                 .### |S I0005  I0036  I0045  I0048
                 .### | I0052
            .######## |
          ######### M|
           .######### | I0021  I0057
  60       .####### +
                .#### | I0013  I0024  I0030  I0033  I0053
         .########### | I0008  I0010  I0016  I0047  I0051
                   ## | I0055
                ##### | I0018
           .######## S|
                   .# | I0025  I0039  I0044  I0046  I0054
  50              .## + I0028
               .##### | I0017  I0040  I0041
                   .# |M I0031  I0038
                  .## | I0002
                    # |
                  ## T| I0050
                      | I0007  I0022
  40                  + I0020  I0023
                      | I0037
                    # |
                      | I0009  I0032
                      | I0014  I0026  I0027
 30                   + I0049
                      | I0004
                      | I0012  I0042
 20                   | I0011
                      | I0003  I0006
 10                   + I0001  I0029
               <less>|<freq>
Keys: "#" = 7 test-takers; "." = 1 to 6 test-takers
```

As for the technical quality, the degree to which the test items fitted the test construct was first examined by inspecting Rasch Standardized Item Weighted Mean-square Fit Statistics estimated with the 989 test-takers (Infit MNSQ). Infit MNSQ is expected to range between -2.0 and + 2.0 to ensure the construct fit (Smith, 2000). Since the sample size of this study was relatively large (N = 989), a stricter range of Infit MNSQ of [0.5; 1.5] was applied (Linacre, 2007). The Infit MNSQ of the 57 items in AVT1 ranged from 0.83 to 1.41, falling well within the aforementioned range. For a valid score interpretation, test items should be independent of each other (Bond & Fox, 2015). Test items with Infit and Outfit MNSQ smaller than -2.0 might indicate local independence violation. As Infit MNSQ of all 57 items ran from 0.83 to 1.41 and their Outfit MNSQ from 0.78 to 1.63, no thread to the local independence was found. The final issue related to the technical quality is whether all test items behave in the same direction and as that of the latent variable. The point-measure correlation or the correlation between the actual observation on a specific item and the Rasch estimate for the measure as the whole, excluding the observation on that item, was examined. The point-measure correlation coefficients of the 57 items in AVT1 were all positive, ranging from .11 to .55. Therefore, it was very likely that these items behaved in the same manner and as that of the latent construct.

4.1.2 Substantive aspect

According to Messick (1995), the substantive aspect of construct validity "refers to theoretical rationales for the observed consistencies in the test responses, including process models of task performance…, along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks" (p. 6). In other words, test tasks should activate mental processes that are typical for the target language use in real life and test-takers are both affectively and cognitively engaged in the test tasks as expected.

This aspect of construct validity was first investigated by looking at the degree to which the empirical item hierarchy was in the hypothesized order. In AVT1, it was hypothesized that the test items would create an item difficulty continuum which was in line with their frequency in the COCA corpus. In the existing literature, this hypothesis is empirically supported as research in both L1 (e.g., Miller & Gildea, 1987; Stenner et al., 1983) and L2 (e.g., Greidaneus & Nienhuis, 2001; Laufer & Nation, 1999) contexts consistently shows that high-frequency words are more likely to be known than their low-frequency counterparts. To test this hypothesis in the case of the present study, the 57 lexical items in AVT1 were first divided into two sub-groups: one including all items with the frequency per million words in the COCA corpus of under 10 times (n = 33) and the other including the remaining items with the frequency per million words in the COCA corpus of 10 times and above (n = 24). Since test scores in both datasets were normally distributed, a t-test for independent samples was run to gauge the between-group difference. The test scores on the higher frequency words indeed far surpassed those on the lower frequency words: t = 4.44 (df = 55, p < .001). To further examine the linear relationship between the empirical item difficulty hierarchy and the item COCA-based frequency, a Pearson correlation coefficient was also calculated for the total Rasch-based test scores for each test item on the one hand and their COCA-based frequency per million words on the other hand. A positive correlation coefficient of .65 (p < .01) was found. In addition, as consistently shown in previous corpus-based research, the frequency ratio of Noun: Verb: Adjective in actual language use is roughly 3: 2: 1, suggesting that nouns stand a better chance to be learnt than verbs, which in turn are also more likely to be learnt than adjectives (Webb, Sasao & Ballance, 2017). A linear regression model in which the above test scores were set as the dependent variable while the COCA-based word frequency and the word-class (i.e., whether the target words were nouns, verbs, or adjectives) as the two predictors was also generated. This model was indeed able to explain the variance in the test scores: F (2,54) = 4.90 (p < .001), R^2 = .54. While the COCA-based word frequency could predict well for the test scores (t = 3.97, p < .01, b = .46), the word class failed to do so (p = .83). These altogether supported the above hypothesis.

As for the target abilities of the test-takers (i.e., their receptive academic vocabulary breadth), it was also hypothesized that those with better L2 proficiency might also be the ones with a larger receptive academic vocabulary. As mentioned above, out of the 989 test-takers, 458 were university students while the remaining 531 were high-school students. In the context of Vietnam, students were all required to achieve at least Level 3 on the Six-level Foreign Language Proficiency Framework or Level B1 on the Common European Framework of Reference for Languages to graduate from their K-12 education and be eligible for university entrance. Therefore, the university students were also hypothesized to have a larger receptive academic vocabulary than the high-school students. On AVT1, the former obtained a mean score of 41.83 out of 57 (*SD* = 7.37), whereas the latter had only a mean score of 38.85 (*SD* = 8.25). The result from a *t*-test for independent samples affirmed that the university students did score significantly higher on this measure than their high-school counterparts: $t = 5.92$ ($p < .001$). The Rasch model also provides evidence for the degree to which test-takers' responses align with what the model predicts for their target abilities (Smith, 2000). This alignment is manifested via Rasch person fit statistics. Again, Infit and Outfit MNSQ for person measures should be between -2.0 and +2.0. Infit MNSQ for the 989 test-takers ranged from .66 to 1.87 while their Outfit MNSQ was between .42 and 4.17. In total, 47 test-takers had Outfit MNSQ larger than +2.0, which accounted for 4.8% of the above sample size. This misfit rate was less than 5% and was expected to occur by chance (given the above huge sample size).

4.1.3 Structural aspect

As AVT1 aimed to measure receptive academic vocabulary breadth, the test construct was by nature unidimensional and this unidimensionality should be well-reflected in the test scores. This is often referred to as the structural aspect of Messickian construct validity (Messick, 1995). According to Linacre (2007), the dimensionality of a language test should be investigated by detecting items that appear to measure a secondary dimension based on Principle Component Analysis (PCA) of item residuals. The Rasch model does this by extracting the first major dimension in a dataset, which is the systematic variation explained by the test construct. If the data are unidimensional and they fit the Rasch model, no other systematic relationship should be present in the residuals. PCA was run for the AVT1 scores. The Rasch model explained 33.3% of the variance in the test scores (eigenvalue = 82.44). According to McLean, Kramer and Beglar (2015), however, test unidimensionality is not dependent on the amount of variance in the test scores accounted for by the test construct itself, but by the size and interpretation of residual contrasts. Therefore, the first residual contrast in the above PCA was given a closer look. In fact, the first residual contrast was able to explain 4% of the variance in the test score with an eigenvalue of 3.28, which was noticeably higher than the chance level of 2.0 as suggested by Linacre (2007). Thus, individual test items with loadings above + .40 or loadings below - .40 should be inspected. Table 1 presents four test items as such.

Table 1

Test Items Potentially Measuring a Second Meaningful Dimension

Test items	Target words	Rasch scores	Loadings	Raw scores
I55	Creation	56.25	+ .60	814
I50	Formulate	43.54	+ .53	693
I09	Empathetic	35.95	+ .40	935
I19	Prerogative	76.29	- .41	267

As shown in Table 1, these four test items were three relatively easy items (Creation, Formulate and Empathetic) and one very difficult item (Prerogative). As these items all belonged to different clusters of test items in AVT1, no second meaningful dimension appeared in this case (item difficulty is not a meaningful dimension). In addition, according to Stevens's (2002) guidelines, the chance for a second meaningful dimension to exist in a dataset is slim if there are fewer than three, four and ten items with the absolute loadings over .80, .60 and .40, respectively. In conclusion, the PCA analysis of item residuals indicates that the AVT1 measured only one construct, presumably receptive academic vocabulary breadth.

4.1.4 Generalizability aspect

The generalizability aspect of construct validity examines to what extent score properties and interpretation could be generalized across population groups, settings, and tasks (Messick, 1995). To provide evidence for this aspect of construct validity, previous researchers (e.g., McLean, Kramer & Beglar, 2015; Schmitt, Schmitt & Clapham, 2001; Webb, Sassao & Balance, 2017) often looked at the invariance in both item calibrations and person measures.

To investigate the item calibration invariance, Differential Item Functioning (DIF) was examined to detect any unexpected behaviour per item compared to the modelled error. The DIF analysis was first performed between male (*n* = 401) and female (*n* = 488) test-takers using a Mantel-Haenszel test. No statistically significant DIF was found for any item (a = .05). The same test was then carried out for the test-takers coming from three different cities/provinces of Vietnam: Hanoi, Bac Giang and Lang Son. This time, significant DIF was detected for one test item – Item 19 "prerogative", which was also the most difficult item in AVT1 with Rasch difficulty estimate of 76.29 CHIPS. In this case, test-takers in Hanoi were more likely to know this word than their counterparts in the other contexts. This was not surprising as the research participants in Hanoi included all university students (*n* = 458), who were shown to have a larger receptive academic vocabulary than the high-school students.

The conventional split-half reliability check was also done. The 57 test items of the original AVT1 were first randomly divided into two halves. The Rasch-based analyses were subsequently run for each half to gauge Rasch person ability estimates. Pearson correlation coefficients were then computed for the Rasch person ability estimates between these two halves and between each half and that of the whole original test version (with the 57 items). The correlation coefficient of the Rasch person ability estimates between the two halves was high with $r = .98$ (disattenuated correlation = 1.00, $p < .001$). A similar high correlation coefficient was also found between each half of the test and the whole test with an identical $r = .96$ (disattenuated correlation = .98 and .97, $p < .001$). A stricter method to test item calibration invariance was proposed by Linacre (2007). Following this method, the original 57 test items in AVT1 were also divided into two groups, but this time based on whether they had a positive or negative residual loading in the Rasch analysis outcomes. The respective Rasch person ability estimates on these two item groups were then pulled out and correlated with each other. Again, their correlation coefficient was high with $r = .97$ (disattenuated correlation = .99, $p < .001$).

When it comes to person measure invariance, Differential Person Functioning (DPF) was also inspected to detect any unexpected behavior per person compared to the modelled error. Rasch person ability estimates were first generated for three groups of the test items: Nouns (*n* = 27), verbs (*n* = 18) and adjectives (*n* = 12). The DPF analysis was then implemented using a *t*-test approach to compare each set of the above Rasch person ability estimates against that of the whole dataset (all 57 test items). Again, no significant DPF was found for any test-takers in any pairwise comparisons.

The Rasch model also provides another set of evidence for the generalizability aspect of Messickian construct validity via its person/item reliability and separation. Recall that in the present study, AVT1 with its 57 test items was administered to a total of 989 test-takers. The item separation was 5.57 and the item reliability was .97. The person separation was 2.44 and the person reliability was .86. As we can see, all these indices were relatively high, suggesting that the score properties and interpretations could be generalized across tasks, populations and contexts.

4.1.5 External aspect

This aspect of construct validity examines the relationship between the observations on the test being validated and those on another already-validated test which aims to measure the same or similar construct (Messick, 1995). As the 457 university students all took Schmitt, Schmitt, and Clapham's (2001) Vocabulary Levels Test (VLT) as a requisite for their university entrance, their test scores, especially those of 30 academic vocabulary items, were retrieved and correlated with their present AVT1 scores. The results showed that their AVT1 scores were moderately correlated with their scores on the academic vocabulary section and the whole VLT with $r = .64$ and .78 ($p < .001$), respectively. The lower correlation coefficient between their AVT scores and their scores on the academic vocabulary section of the VLT might be attributed to the ceiling effects frequently occurring in the latter. Nonetheless, it is still clear from the above report that the AVT did have fair external reliability.

4.2 Qualitative evidence

Coding of the qualitative data revealed some major patterns in the guessing behaviors of the 50 selected test-takers. First, most of the interviewees (42/50, 84%) admitted that they did make guesses at some point during their test completion. It was because the test did prompt them to make guesses in the test instruction "*There is no penalty for guessing*" (p. 1). Second, some of their guesses were totally blind as they had no clues for guessing. This especially held true for items 17, 18 and 19 which, as most of them believed, were the hardest items in the test and they tended to resort to the option that sounded the best in these cases. Some also made blind guesses for these three test items just because they were "*running out of time*" (e.g., HS251, 287 or US25, 39 and 311) – a finding that Pecorari, Shaw, and Malmström (2019) did not find in their study. Table 2 provides the specific results from the aforementioned coding.

Table 2

Test-takers' Guessing Behaviors

	Guessing behaviors	YES	NO
01	Guessing from word parts	√	
02	Guessing from knowing another member of a word family	√	
03	Guessing from elimination and association	√	
04	Guessing from co-textual clues		X
05	Blind guessing	√	
	• Choosing the option that is the most uncommon		X
	• Choosing the option that is the shortest or longest		X
	• Choosing the option that is understood best		X
	• Choosing the option that just sounds the best	√	

Key: YES/√ = guessing behaviors mentioned by at least three members in each focus group and at least three focus groups; NO/X = the remaining cases

However, most of their guesses were based on some plausible clues, which could be further categorized into three types of guesses. One was that test-takers did not know a target word before taking this test. However, they could deduce the word meaning from the word parts (e.g., word stem, prefix, or suffix). One often-mentioned example for this was Item 13 with the target word "*disintegration*". Some interviewees reported that they never met this word in their past learning, "*…but I do know the meaning of the word "integration" and the prefix "dis" means something opposite so I select "the process of falling apart" as the correct meaning for "disintegration"*", shared US268.

Another type of guess was associated with their "*process of elimination*" ["*chiến lược loại trừ dần*" (HS137)]. For example, test-takers knew the meanings of two target words in one test item and they also already knew the meanings of the three distractors. Therefore, even though they did not know the final target word before the test, they were still able to find the correct meaning for this lexical item. Their final guessing behavior, albeit not helping get the correct answer, did enable them to increase the chance of doing so. Just like the above type of guesses, test-takers this time knew for sure the meaning of only one out of the three target words in each test item as well as those of the three distractors. Apparently, the chance for their correct guesses now increased from about 17% (1:6) to 50% (1:2).

5. Discussion and Implications for AVT1 Developers and Users

This report showed that AVT1 had sufficient spread of item difficulty (with seven statistically distinct item difficulty groups in the empirical item hierarchy) which could measure the target ability – the receptive academic vocabulary breadth – of an overwhelming 90% of the 989 test-takers. However, there still existed some noticeable gaps in the empirical item hierarchy, especially in the region with item difficulty of below 35 and above 80 CHIPS. Therefore, the test developers might consider adding new items into the existing test or modify some easy test items which were placed at the bottom of the above Wright Map. Test users should also be aware of this limitation when they put AVT1 into practice because ceiling effects can occur and the target abilities of strong L2 students might not be properly measured. All test items and almost all test-takers (95.2%) displayed good fit to the Rasch model. Obviously, Pecorari, Shaw, and Malmström's (2019) item facility and item discrimination evidence in their test validation could not enable a direct comparison of item calibrations and person measures on the same psychometric (Rasch) scale as above.

Rasch-based evidence suggested that test items and test-takers performed according to *a priori* hypotheses or, in other words, both test items and test-takers behaved in the same way as the Rasch model predicted. However, the results of the follow-up focus-group interview data analyses showed that AVT1 did allow a relatively large room for guessing effects, both blind and cued. While all other similar vocabulary tests such Schmitt, Schmitt, and Clapham's updated Vocabulary Levels Test (2001), or McLean, Kramer and Beglar's Listening Vocabulary Levels Test (2015) did their utmost to mitigate the guessing effects, Pecorari, Shaw and Malmström's Academic Vocabulary Test (2019) encouraged this practice (for unknown reasons). Two common techniques that Read (2000) highly recommended for vocabulary test developers to minimize the guessing effects are (a) to increase the number of options and (b) to incorporate some strong distractors into these options.

As for classroom use, if L2 instructors would like to employ AVT1 as a summative assessment instrument to gauge the outcome of past learning (i.e., assessment of learning), caution should be taken into score interpretation because the examples in Section 4.2 showed that some students did not know a target word before the test, but the test itself could be useful input for them to guess the meaning of this new word. Consequently, their test scores did not reflect only the result of their past learning but also that of their vocabulary learning while sitting the AVT. This finding calls for a more specific test guideline or manual from the test developers which will specify the test purposes, the way to handle guessing effects in test implementation as well as in score interpretation against the above test purposes.

The present dataset does not indicate that AVT1 measured a second meaningful dimension, apart from the latent variable – the receptive academic vocabulary breadth. Both the number and strength of the test items with the absolute loadings over .40 did not suffice to constitute a second meaning dimension in item residuals. Therefore, AVT1 was basically unidimensional, which in turn allowed a straightforward interpretation of the test scores. Pecorari, Shaw, and Malmström's (2019) test validation failed to provide evidence for this critical aspect of construct validity. Therefore, they might need to do so, especially with Form 2.

Results from both DIF and DPF analyses suggested that test items and test-takers consistently behaved in the same manner and within the modelled error tolerance. The university students' AVT1 scores also correlated well with their scores on Schmitt, Schmitt and Clapham's Vocabulary Levels Test, including the academic vocabulary section in this test.

6. Conclusion

Pecorari, Shaw and Malmström's (2019) Academic Vocabulary Test proved to be a useful test to measure the receptive academic vocabulary breadth. The current number of test items was relatively sufficient to measure the above target ability; however, it could be better if more difficult test items were added to the existing version. All test items displayed good fit to the Rasch model and performed as predicted by *a priori* hypotheses. The test itself was unidimensional and possessed acceptable internal and external reliability, which together enabled test score properties and interpretations to be generalized across populations, settings, and tasks. However, as the test still allowed room for guessing, caution should be taken in administering the test and interpreting the test score. Finally, as the present study did not validate Form 2 of the AVT, research is needed to fill this gap.

Notes

1. These aspects will be further elaborated in the Findings Section.
2. The initial number of research participants was 1,008; however, data provided by 19 test-takers were removed from the final analyses because these test-takers did not finish all the test items or they would like to withdraw their participation.

Funding

This research has been completed under the sponsorship of the University of Languages and International Studies (ULIS, VNU) under the Project No. N.23.04.

References

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

Beglar, D., & Hunt, A. (1999). Revising and validating the 2000 word level and university word level tests. *Language Testing, 16*, 131-162. https://doi.org/10.1177/026553229901600202

Boers, F. (2021). *Evaluating second language vocabulary and grammar instruction: A synthesis of the research on teaching words, phrases, and patterns*. Routledge.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Routledge.

Brown, D., Stewart, J., Stoeckel, T., & McLean, S. (2021). The coming paradigm shift in the use of lexical units. *Studies in Second Language Acquisition, 43*(2), 462-471. https://doi.org/10.1017/S0272263121000668

Campion, M.E., & Elley, W.B. (1971). *An academic vocabulary list*. New Zealand Council for Educational Research.

Cobb, T., & Horst, M. (2004). Is there room for an AWL in French? In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 15-38). John Benjamins.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly, 34*, 213-38. https://doi.org/10.2307/3587951

Coxhead, A., & Demecheleer, M. (2018). Investigating the technical vocabulary of Plumbing. *English for Specific Purposes, 51*, 84-97. https://doi.org/10.1016/j.esp.2018.03.006

Dang, T.N.Y., & Webb, S. (2014). The lexical profile of academic English. *English for Specific Purposes, 33*, 66–76. https://doi.org/10.1016/j.esp.2013.08.001

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The Academic Spoken Word List. *Language Learning, 67*(4), 959-997. https://doi.org/10.1111/lang.12253

Edgarsson, G. (2018). Academic vocabulary proficiency and reading comprehension among Icelandic secondary school students. In B. Arnbjörnsdóttir & H. Ingvarsdóttir (Eds.), *Language development across the life span* (pp. 95-112). Springer.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*, 305-327. https://doi.org/10.1093/applin/amt015

Goldenberg, C. (2008). Teaching English language learners: What the research does and does not say. *American Educator, Summer*, 8-44.

Greidanus, T., & Nienhuis, L. (2001). Testing the quality of word knowledge in a second language by means of word associations: Types of distractors and types of associations. *The Modern Language Journal, 85*, 567-577. https://doi.org/10.1111/0026-7902.00126

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL-International Journal of Applied Linguistics, 166*(2), 278-306. https://doi.org/10.1075/itl.166.2.04gyl

Hatch, E. & Farhady, H. (1982). *Research design and statistics for applied linguistics*. Newbury House Publishers, Inc.

Jacobs, V. A. (2008). Adolescent literacy: Putting the crisis in context. *Harvard Educational Review, 78*, 7-39.

Laufer, B., & Nation, I.S.P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16*, 33-51. https://doi.org/10.1177/026553229901600103

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning, 54*(3), 399-436.

Lin, L. H. F., & Morrison, B. (2010). The impact of the medium of instruction in Hong Kong secondary schools on tertiary students' vocabulary. *Journal of English for Academic Purposes, 9*, 255-66. https://doi.org/10.1016/j.jeap.2010.09.002

Linacre, J.M. (2007). *A user's guide to WINSTEPS*. Chicago: winsteps.com.

Lu, C., & Coxhead, A. (2020). Vocabulary in Traditional Chinese Medicine: Insights from corpora. *ILT-International Journal of Applied Linguistics, 171*(1), 34-61. https://doi.org/10.1075/itl.18020.lu

Masrai, A., & Milton, J. (2018). Measuring the contribution of academic and general vocabulary knowledge to learners' academic achievement. *Journal of English for Academic Purposes, 31*, 44-57. https://doi.org/10.1016/j.jeap.2017.12.006

McLean, S., Kramer, B., & Beglar, D. (2015). The creation and validation of a listening vocabulary levels test. *Language Teaching Research, 19*, 741-760. https://doi.org/10.1177/1362168814567889

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly, 3*(1), 31-51. https://doi.org/10.1207/s15434311laq0301_3

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Miller, G.A., & Gildea, P.M. (1987). How children learn words? *Scientific American, 257*, 94-99.

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly, 47*, 91-108. https://doi.org/10.1002/RRQ.011

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nation, I.S.P. (1990). *Teaching and learning vocabulary*. Newbury House.

Nation, I.S.P. (1983). Testing and teaching vocabulary. *Guidelines, 5*, 12-25.

Nation, I.S.P., & Beglar, D. (2007). A Vocabulary Size Test. *The Language Teacher, 31*, 9-13.

Paul, P., Stalhnan, A., & O'Rourke, J. (1990). *Using three test formats to assess good and poor readers' word knowledge (Tech. Rep. No. 509)*. Urbana-Champaign: University of Illinois, Center for the Study of Reading.

Pecorari, D., Shaw, P., & Malmström, H. (2019). Developing a new academic vocabulary test. *Journal of English for Academic Purposes, 39*, 59-71. https://doi.org/10.1016/j.jeap.2019.02.004

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal, 19*, 12-25. https://doi.org/10.1177/003368828801900202

Read, J., & Chapelle, C. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 3-32. https://doi.org/10.1177/026553220101800101

Read, J., & Dang, T. N. Y. (2022). Measuring depth of academic vocabulary knowledge. *Language Teaching Research*. https://doi.org/10.1177/13621688221105913

Read, J. (2007). Teaching and learning vocabulary: Bringing research into practice. *Studies in Second Language Acquisition, 29*(1), 128-129. https://doi.org/10.1017/S0272263107220066

Schmitt, N., Ng, J. W. C., & Garras, J. (2011). The word associates format: Validation evidence. *Language Testing, 28*(1), 105-126. https://doi.org/10.1177/0265532210373605

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55-88. https://doi.org/10.1177/026553220101800103

Smith, R. M. (2000). *SIMITEM: Rasch model data simulation program.* Maple Grove, MN: JAM Press.

Skjelde, K., & Coxhead, A. (2020). Mind the gap: Academic vocabulary knowledge as a predictor of English grades. *Acta Didactica Norden, 14*(3), Article 6.

Stenner, A.J., Smith, M., & Burdick, D.S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement, 20*, 305-315.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th edition). Erlbaum.

Thorndike, E.L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. Bureau of Publications, Teachers College, Columbia University.

Webb, S., Sasao, Y., & Balance, O. (2017). The updated vocabulary levels test. *ITL - International Journal of Applied Linguistics, 168*(1), 33-69 https://doi.org/10.1075/itl.168.1.02web.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 3*(2), 215-229.

Chi Duc Nguyen is a lecturer in Applied Linguistics and TESOL in the Faculty of English Language Teacher Education, VNU University of Languages and International Studies, Vietnam National University, Hanoi. He conducts research in second language acquisition, particularly incidental grammar and vocabulary learning through meaning-focused input, output and interaction activities. He now ventures into the field of language testing and assessment. His latest publications appeared in TESOL Quarterly, Language Teaching Research, RELC and Reading in a Foreign Language.

Hoang Thi Hanh is a lecturer at the Faculty of Linguistics and Cultures of English Speaking Countries, VNU University of Languages and International Studies, Vietnam National University, Hanoi. She earned her PhD of Applied Linguistics at the University of Queensland, Australia in 2013. She is interested in exploring the personal and cultural elements that shape different people's learning conditions and identities within and across societies.