A test of controlled productive knowledge of English academic vocabulary

Citation for the original published paper (version of record):
Pecorari, D., Malmström, H., Warnby, M. (2025). A test of controlled productive knowledge of English academic vocabulary. Acta Didactica Norden, 19(1): 1-32.
http://dx.doi.org/10.5617/adno.11584

Diane Pecorari
University of Leeds
Hans Malmström
Chalmers University of Technology
Marcus Warnby
University of Gothenburg

DOI: https://doi.org/10.5617/adno.11584

A test of controlled productive knowledge of English academic vocabulary

Abstract
Within the field of language education, language assessment is an important concern, for both pedagogical and research purposes. Vocabulary is a key aspect of language proficiency, underpinning the four skills of reading, writing, speaking and listening. In academic contexts, of particular importance is academic vocabulary, the part of the lexicon which is better represented in academic than in general discourse, and which constitutes a key dimension of students' academic literacy. The present paper details the design and initial validation of a new test of controlled productive knowledge of English academic vocabulary. The test design closely follows that of the Productive Vocabulary Levels Test (Laufer & Nation, 1999) but the new instrument – named the Productive Academic Vocabulary Test – uses a different basis for defining academic vocabulary and employs an updated set of scoring principles. The test was validated using scores from 232 participants at three Swedish universities. Findings indicate that the test can discriminate levels of knowledge of English academic vocabulary amongst the target population and that the scoring principles provide a nuanced measure of vocabulary knowledge. Pedagogical and research implications are discussed.

Keywords: academic vocabulary; vocabulary testing; language education; Productive Academic Vocabulary Test (PAVT); English for academic purposes (EAP)

Ett test av kontrollerad produktiv akademisk engelsk ordkunskap

Sammanfattning
Av både pedagogiska skäl och för forskningsändamål är testning och utvärdering av språkfärdighet ett viktigt område inom språkdidaktik. Ordkunskap är en central dimension av språkfärdighet eftersom den ligger till grund för samtliga fyra färdigheter: läsa, skriva, tala och lyssna. I akademiska sammanhang är det särskilt viktigt att ha ett akademiskt ordförråd, det vill säga ord som förekommer oftare i akademisk diskurs än

i allmänna sammanhang. Denna artikel beskriver utvecklingsprocessen för att skapa och validera ett nytt test för kontrollerad produktiv engelsk akademisk ordkunskap. Testformatet följer Productive Vocabulary Levels Test (Laufer & Nation, 1999), men detta nya test, benämnt Productive Academic Vocabulary Test, använder en annan definition av akademiskt ordförråd och ett mer nyanserat sätt att bedöma testsvar. Testet validerades med resultat från 232 deltagare vid tre svenska universitet. Resultaten visar att testet kan mäta olika grader av ordkunskap hos deltagare och att bedömningsprinciperna ger en nyanserad bild av ordkunskap. Implikationer för pedagogik och vidare forskning diskuteras.

Nyckelord: akademisk ordkunskap; testning av ordkunskap; språkdidaktik; Productive Academic Vocabulary Test (PAVT); engelska för akademiska ändamål (EAP)

Introduction

At the end of the 20th century, Laufer and Nation (1999) published the Productive Vocabulary Levels Test (PVLT), an instrument designed to measure test takers' controlled (rather than free) productive knowledge of general as well as academic English vocabulary. The objective of the PVLT was to offer testers (such as teachers and researchers) a means to establish stages in individuals' vocabulary development based on a "rounded picture of a learner's vocabulary knowledge" (1999, p. 34). In many ways, the PVLT has stood the test of time. While aspects of the test have been criticised (e.g., Fitzpatrick & Clenton, 2017; Webb, 2008), the test continues to be used in different contexts and for various purposes (e.g., Uchihara et al., 2022; Vu, 2024). The PVLT covers general high-frequency vocabulary across different frequency levels; in addition, a section tests knowledge of academic vocabulary, originally words from the University Word List (Xue & Nation, 1984) and in later versions, Coxhead's (2000) Academic Word List (AWL). In the present study, we focus specifically on this latter domain: English academic vocabulary, and measurement of productive knowledge of academic vocabulary for educational purposes.

Academic vocabulary knowledge is a crucial aspect of students' overall academic literacy and significantly enhances their academic performance (e.g., Aizawa & Rose, 2020; Masrai & Milton, 2021; Vu, 2024). For various reasons (e.g., for diagnostic purposes), it is useful to be able to test language users' knowledge of academic words, receptively as well as productively. Currently, several instruments exist that test receptive knowledge of English academic vocabulary (e.g., the Word Associate Test, Read, 1993; the academic section of the Vocabulary Levels Test, Nation, 1990; Schmitt et al., 2001; the Academic Vocabulary Test, Pecorari et al., 2019). By contrast, the PVLT is the only widely available validated instrument for measuring productive knowledge of academic vocabulary. As with other tests based on the AWL, the academic section of the PVLT (henceforth PVLT-Ac) can be regarded as somewhat dated, given that a

newer inventory of academic vocabulary, Gardner and Davies' (2014) Academic Vocabulary List (AVL) is available.

The objective of this paper is to detail exploratory work relating to the design and initial validation of a new test of controlled productive knowledge of English academic vocabulary. We do this much in the same spirit as did Meara and Fitzpatrick, i.e., as "a first step" (2000, p. 20) in the development of a test instrument, in the belief that it is beneficial to make this test available. This new instrument – the Productive Academic Vocabulary Test (PAVT) – follows the design of the PVLT, but the new instrument is based on the AVL. Additionally, it employs an updated set of scoring principles, allowing for a more sensitive measurement of written controlled productive academic vocabulary knowledge. The PAVT enables teachers and researchers to assess test takers' productive academic vocabulary, thereby supporting more targeted and effective language instruction and research.

Literature review

This section discusses the construct *academic vocabulary* and the rationale for measuring knowledge of academic vocabulary; provides a brief review of existing academic vocabulary measurement protocols; and presents an argument for a new test of controlled productive academic vocabulary knowledge.

Why measure productive knowledge of academic vocabulary?

Academic discourse contains specialized vocabulary that is not as commonly used in everyday language. A frequent distinction is between disciplinary academic vocabulary – words that have a specific affinity with a specific discipline – and general academic words, those which occur frequently in academic discourse regardless of the discipline (Coxhead, 2016). Debate has centred around the place *general* academic vocabulary holds in academic discourse, and while some scholars, like Hyland and Tse (2007), believe that no single, general academic vocabulary is equally useful across all disciplines, other researchers continue to find general academic vocabulary a useful concept; Gardner and Davies' (2014) AVL is just one of many examples.

General academic vocabulary is a key component of general academic literacy (in the same way that disciplinary academic vocabulary is central to disciplinary literacy) and mastering this vocabulary can significantly improve the ability to function in (e.g., read academic texts, listen to lectures, write assignments and speak in seminars) and thus be successful in an academic environment, as evidenced by several studies. For example, Masrai and Milton (2018; 2021) established that receptive knowledge of academic words is a strong predictor of students' academic success, accounting for a significant amount of variance in grade point averages. In a study by Warnby (2024), students' scores on an

academic reading task including multiple-choice questions as well as productive word gaps correlated highly with receptive academic word knowledge as measured by two tests, one based on the AWL (r=.72) and the other on the AVL (r=.80). Focusing on students' productive knowledge of academic vocabulary, Vu (2024) confirmed the important role played by knowing many academic words: productive vocabulary knowledge, as measured by Vietnamese students' scores on the PVLT, was a significant predictor of academic outcomes in their English-medium education setting. The importance of academic vocabulary for academic writing has also been confirmed (e.g., Alharbi, 2017).

Since understanding and using academic vocabulary is considered important, ways of measuring this knowledge are equally important, for example, to help identify learners with limited academic vocabulary who may struggle in academic environments. In this regard, Vu (2024) emphasizes that it is relevant to measure both receptive and productive academic vocabulary knowledge. Productive knowledge of academic English vocabulary becomes even more crucial as students advance to higher levels of education, where they face greater demands for producing academic work. Additionally, with English playing a more prominent role in many global educational settings, there is a heightened need for proficiency in English academic vocabulary at these advanced educational stages.

How is productive knowledge of academic vocabulary measured?

Productive knowledge of academic vocabulary can be assessed in two primary ways. A distinction is typically made between free productive and controlled productive measures:

> controlled indicates that the test is designed to elicit specific, predetermined vocabulary items, and free indicates that vocabulary produced by the test taker in a relatively unconstrained task will be measured.
>
> (Fitzpatrick & Clenton, 2017, p. 846)

The choice of approach obviously depends on what the objective of the assessment is and what kind of vocabulary knowledge claims one wants to make.

The research literature features two influential means of measuring free or controlled productive knowledge of academic vocabulary, neither of which is entirely unproblematic. A procedure that lends itself to assessing free academic vocabulary knowledge is lexical frequency profiling (LFP, Laufer & Nation, 1995); LFP involves analyzing the vocabulary levels and frequency bands present in a text. Specifically, LFP categorizes words in a text into different frequency bands based on corpus data and offers a profile of the text produced, detailing the number of words in the text, the number of different words used, and the proportion of words from each frequency band. Typically, the LFP output also details the proportion of academic words; in most cases this has been based on the AWL, but in principle any list of academic words can be used.

The PVLT, and more specifically the PVLT-Ac, is the only validated and commonly accessible instrument designed to test controlled productive knowledge of academic vocabulary, as other instruments for assessing controlled productive vocabulary knowledge (such as the Lex30 by Meara & Fitzpatrick, 2000, and the P Lex by Meara & Bell, 2001) focus on general rather than academic productive vocabulary knowledge. A partial exception is Paribakht and Wesche's (1993) Vocabulary Knowledge Scale, which provides a framework for describing knowledge of any target word, and has been used to test academic vocabulary (e.g., Freimuth, 2020). Laufer and Nation explain their use of the term "controlled productive" thus:

> We use the term "controlled productive ability" for the ability to use a word when compelled to do so by a teacher or researcher, whether in an unconstrained context such as a sentence-writing task, or in a constrained context such as a fill-in task where a sentence context is provided and the missing target word has to be supplied.
>
> (1999, p. 37)

This definition is consistent with other conceptualisations of controlled (as opposed to free) in the literature (see, e.g., Fitzpatrick & Clenton, 2017). Figure 1 provides an example of this controlled productive principle in a constrained context (item 17, version 1 of the PVLT-Ac). Helped by the prompt *hom*_____, the test taker is expected to supply the rest of the target *homogenous*, as indicated by the gap and blank line. The initial letters of the target word (always as few letters as possible, according to Laufer and Nation, 1999) are provided to exclude other options and elicit only the target word.

In a hom_____ class all students are of a similar proficiency.

Figure 1. Example item from Laufer and Nation's (1999) PVLT-Ac

According to Laufer and Nation (1999), the PVLT, including the academic section, measures the size of a test takers' orthographic vocabulary, based on words sampled from across word frequency bands (i.e., the 2,000, 3,000, etc. most frequent words in English). The original PVLT tested academic vocabulary with items from the University Word List (Xue and Nation 1984), while some studies have used items from the AWL. Each section of the PVLT contains 18 items of the kind shown in Figure 1, and the size of a test taker's vocabulary is determined by dividing the number of correctly answered items on each level by the maximum total score (18); the "percentage score" for a level on the test serves as "a very rough indication of the number of words known at that level" (1999, p. 41). Laufer and Nation (1999) say that a percentage score of 85-90% indicates "satisfactory mastery of a level." This conception of "mastery" has been challenged in later research (see for example McLean, 2021), but continues to be

used as a rough benchmark. Importantly, though, Fitzpatrick and Clenton (2017, p. 847) highlight the necessity of testing many rather than fewer words (from whatever underlying list is used) “if inferences are to be drawn about untested words.”

The PVLT is purported to be “a very practical instrument. . . easy to administer. . . completed in a short time [and] easy to mark” (Laufer & Nation, 1999, p. 41). Fitzpatrick and Clenton (2017, p. 845), however, caution that the apparent simplicity of the test format could be misleading and “belie the complexity of the construct they claim to measure” (and virtually the same argument can be raised against LFP). In this regard, Fitzpatrick and Clenton (2017) highlight the multidimensionality underlying vocabulary knowledge and remind us that tests like the PVLT only measure a fraction of productive vocabulary knowledge (in this case, controlled recall, because the item tested is cued).

Why is there a need for a new test of controlled productive academic vocabulary knowledge?

For many scholars and testers, the PVLT-Ac has been the go-to instrument for measuring controlled productive academic vocabulary knowledge for nearly three decades. It has been used for various testing purposes with different groups of participants in different geographical and academic contexts. Thus, for example, Zheng (2009) used the VLT and PVLT to assess the receptive and productive vocabulary knowledge among Chinese learners of English as a foreign language (EFL) at five word-frequency levels, including academic vocabulary; her focus was the size relationship between EFL learners’ receptive and productive vocabulary. The PVLT-Ac was also used by Yamamoto (2014) with first-year students in Japan to study lexical gains from vocabulary list learning activities. More recently, Kiliç (2019) made use of the PVLT to study the effect of academic and general productive vocabulary knowledge on the various productive academic skills (writing and speaking performance) of Turkish students attending an intensive English class in preparation for English-medium education.

Despite its longevity and status as a test of controlled productive academic vocabulary knowledge, the PVLT-Ac – built on the AWL – has some known issues. What is arguably a better source for target words is now available. In comparison to the AWL (Coxhead, 2000), Gardner and Davies’ (2014) AVL used a larger, more representative and contemporary corpus, employing lemmas instead of word families, and utilizing more sophisticated computational techniques to identify distinctly academic vocabulary across disciplines. It thus "can be seen as an advance on the AWL” (Therova, 2020, p. 7).

Another feature of the PVLT which has caused it to be criticized is its apparent inability to measure partial word knowledge (e.g., Webb, 2008). A response to a test item like the one in Figure 1 is typically scored as either correct or incorrect,

denying test takers the opportunity to display partial knowledge of the target word. In this way, PVLT scores lack the sensitivity to distinguish between a test taker who has no knowledge of the word at all, and a test taker who knows many aspects of the word, such as its meaning, but not all aspects, such as its correct spelling or how it should be inflected to fit in a particular grammatical context.

For these reasons, a new means of measuring productive academic vocabulary was deemed beneficial. In the remainder of this paper, we present a tool for this purpose, the Productive Academic Vocabulary Test (PAVT).

Methods

In this section we describe the construction of the PAVT and the process of piloting and validating it.

Purpose and context of the test

Following standards in test development, such as providing clear test specifications for intended purposes and contexts (Schmitt et al., 2020), the test presented here was developed in the context of a larger research project investigating the English-language proficiency of postgraduate students in Swedish higher education. For the purposes of that project there was a need to measure both receptive and productive academic vocabulary. For receptive purposes, the Academic Vocabulary Test (AVT, Pecorari et al., 2019) was used. The PAVT was developed to provide a comparable, parallel, productive measure.

Test construction

Because of the simplicity of use of the PVLT format, the PAVT is modeled on it. As noted above, this involves a context sentence in which a target item fits and is cued by some of the initial letters of the word, followed by a blank space in which the test taker can write the answer.

In constructing the target sentences, several principles were adopted. The first was a defining vocabulary principle, according to which the words in the context sentence should not be meaningfully less frequent than the target item, provided that using frequently occurring vocabulary did not interfere with the objective of providing clear, unambiguous and natural-sounding prompts. The purpose of this procedure was to ensure that failure to answer a given question would reflect the inability to recall the target, rather than an inability to understand the context sentence. This was done in the first instance impressionistically; subsequently, context words were profiled using the profiling function on the Compleat Lexical Tutor site. This confirmed that the context sentences were written in accessible vocabulary. The 1K level, i.e., the list of the 1,000 most frequent word families, covers 83.3% of the words in the context sentences, and the 1K-3K lists provide 98% coverage (all but ten words). By contrast, among the target items on the

PAVT, only two are at the 1K level, and the least frequent word is at the 13K level.

A second criterion was that each context sentence should be sufficiently specific that the target word or another very close synonym was called for; that is to say, other words which are unrelated in meaning to the target would be made inappropriate by the overall sense of the context sentence. The initial letters then served the purpose of cuing the target word rather than a close synonym.

Context sentences were also designed to sound natural and idiomatic when the target word was used. However, strong collocations with the target or formulaic expressions involving it were avoided. These could make it ambiguous whether the learner knew the target word in its own right, as opposed to knowing the formulaic unit of which it was part; alternatively the presence of the beginning of a formulaic expression could prime the test-taker to produce the target.

Following Laufer and Nation (1999), the cue consisted of the minimum number of letters necessary to eliminate alternative answers. In the initial phase of construction, this was done on a trial-and-error basis. Context sentences were constructed giving the first two letters of the target. When it became apparent, either during the initial process of writing items, or during piloting, that multiple answers were possible, additional letters were added, until an unambiguous prompt was arrived at. For example, in prompting the target *accuracy*, piloting showed that two letters were sufficient; pilot test takers either answered correctly or did not supply an answer but no plausible competing words beginning with *ac* were offered. However, in the sentence *It was the biggest (migration) of groups of people in European history*, providing the first two letters of the target, *migration*, led to numerous test takers answering *mistake*, so the third letter had to be added. The length of the blank space following the cue letters was constant in all questions, to avoid suggesting the length of the target word.

Because the impetus to the creation of the PAVT was to be able to compare test takers' scores with their scores on the AVT, which measures receptive recognition of academic words, items on the PAVT were drawn from Form 1 of the AVT. This means, in turn, that they come from Gardner and Davies' (2014) AVL, and are sampled from the range of frequencies on that list (see Pecorari et al., 2019 for details of their sampling procedures in developing the AVT). Items on the test (see Appendix 1 for the full version of the PAVT) appear in order of their frequency in the Corpus of Contemporary American English (COCA) at the time the AVL was produced, from most to least frequent. Ordering items according to frequency is based on the idea that, all other things being equal, the higher the frequency of a word, the more likely it is to be known, and for that reason, the test generally becomes more difficult as the test taker progresses through it.

During the process of developing and piloting the test, it became clear that some items on the AVT simply do not lend themselves to the controlled productive format, because disambiguating all alternatives would require

providing an unreasonably large number of initial letters. In the cases of five such target items (out of 57 on the AVT), no meaningful way of testing the items was identified. For instance, the target *subset* attracted a wide range of answers beginning with *sub*, e.g., *subgroup*, and *subsection*. To preclude *subgroup* and similar words as possible answers, it would have been necessary to provide four of the six letters in the target, and that still would not have precluded *subsection*. These five words were therefore eliminated, resulting in a test of 52 items, which is significantly more than the 18 items in the academic section of the PVLT (Laufer & Nation, 1999), providing greater robustness in measuring productive academic vocabulary knowledge. In a few additional cases, two closely related answers were judged to be acceptable, i.e., a small number of questions have two correct answers. An example is the target *ubiquity*, for which the answer *ubiquitousness* is also accepted as correct. Finally, one target word – *aid* – is cued with only the first letter, because this was deemed less problematic than the alternative, providing the first two letters and asking test-takers to supply only the final one. These less-than-ideal features of a small number of items were the concomitant cost of a test parallel to the receptive AVT.

Participants

In the early stages of construction, the PAVT items were repeatedly trialed on small convenience samples of educated (i.e., with at least one university degree) speakers of English with a range of first language (L1) backgrounds. Although there was no formal measure of their English proficiency, there is reason to believe that all were at least at B2 (i.e., upper intermediate) level according to the Common European Framework of Reference (Council of Europe, n.d.).

Following this development and trial stage, the test was administered to 232 participants. Of these, 102 were master's students in science (MSc). The remaining 130 were enrolled in higher education pedagogy courses, i.e., courses intended to confer skills in teaching at university. Such training is mandatory for doctoral students at Swedish universities, and as a result, a large proportion of the test takers recruited from the HE pedagogy courses (henceforth referred to as the HEP group) were PhD students. In Sweden, doctoral candidates are (in the majority of cases) simultaneously students and university employees who are assigned some teaching responsibilities. Amongst the HEP participants, some were pre-service teachers and others had previous or concurrent teaching experience. A small number of individuals in this group may have been university teachers in some other category.

Participants came from three universities. The HEP participants came from two universities of science and technology, and one comprehensive university. The participants from the comprehensive university were, however, attending an iteration of the higher education pedagogy course which had been designed for the natural sciences. The master's students all came from one of the universities of science and technology. As a whole, therefore, the participants were drawn

from the same broad disciplinary area. All three universities are regarded as prestigious in the Swedish context, and this may limit the extent to which participants can be regarded as representative of the student body in Swedish higher education more generally.

Due to constraints on the access to data collection provided, slightly different types of background information were available for the participants. Information about L1 was available for the HEP participants. They included a total of 29 different L1 backgrounds. Swedish (*n*=33, representing 25.4% of participants) was the most frequent L1, with the remainder ranging from Chinese (*n*=15) to Amharic (*n*=1). This linguistic diversity of languages is, broadly speaking, typical of the postgraduate student body at these universities. Data about L1 was not available for the MSc students; however, prior educational background was. Of the MSc students, 66 (64.7%) had earned their bachelor's degrees in Sweden, while 33 (32.4%) earned it outside of Sweden. In sum, while the single largest group of participants had a Swedish background either in terms of L1 or of prior education, they constituted a minority in a group reflecting the diversity of Swedish higher education.

All of the master's students were enrolled on full English-medium programmes; that is to say that all formal instructional activities, including lectures, assigned reading and assessments, were to be entirely in English, although informal interactions outside of class time might take place in Swedish or another language. The courses which the HEP participants were enrolled on were all taught through the medium of English. Although it is unknown in what language or languages they conduct their academic activity outside of the HEP course, the presence of English in Swedish higher education is strong, and especially so in the sciences and technology (cf. Malmström & Pecorari, 2022). This, in conjunction with the fact that only a minority of the HEP participants had Swedish as L1, makes it highly likely that for this group, much of their academic activity is conducted in English.

Participants were given a paper-and-pencil version of the test, administered during a meeting of a class in which they were enrolled, and to which the teacher allowed access. According to the prevailing regulations in Sweden, ethical approval was not required for this study; all participants were informed about the voluntary nature of participation and provided their informed consent in writing.

Scoring

Following Webb (2008), the test was scored in two ways. In the strict scoring condition, to earn a point for an item, the participant had to give an entirely correct answer, i.e., correctly spelled (though accepted regional variations in spelling were credited with a point), and in the correct grammatical form. In relaxed (or in Webb's term, *sensitive*) scoring, errors in spelling or grammatical form were allowed, provided that they did not create any ambiguity about whether the target item was intended. Thus, for the item in Figure 2, only *homogeneous* received a

point in strict scoring. In relaxed scoring, *homogenius* was also accepted for a point, but *homolog* was not, because, despite the fact that it fits grammatically into the sentence, it produces a sentence which is arguably meaningless, and ignores the strong implication of the context sentence that a word meaning the opposite of *diverse* is sought.

Crediting partial knowledge allows different aspects of productive word knowledge to be tested. A fully correct answer demonstrates productive word knowledge in form, meaning and use, while the relaxed scoring condition recognises that some test takers may understand the prompt and recall the word's meaning but be uncertain about its exact spelling or grammatical form. The additional sensitivity provided by this procedure gives a more nuanced measure of productive academic vocabulary knowledge, addressing a key limitation of earlier tools like the PVLT.

Is it better for society to be diverse or hom_____?

Figure 2. PVLT item 31

A record was kept of answers not accepted under strict scoring. A research assistant assigned them provisionally into two categories, as acceptable for a point under relaxed scoring or not, using the criteria described above. Thereafter, the first and second authors independently reviewed them and, where there was uncertainty, discussed answers until consensus was reached. All tests were then rescored to ensure that points under relaxed scoring were awarded consistently.

Validating the PAVT

Several measures were used in the analysis to arrive at an understanding of how well the test served its purpose. Descriptive statistics (produced using SPSS) were used to explore central tendencies (mean, median, mode) and variability (standard deviation), and an exploratory factor analysis (EFA) was performed on both scoring sets of the full sample using maximum likelihood estimation with promax rotation. While the test is intended to measure a single construct – productive academic vocabulary knowledge – promax rotation was chosen as part of the validation process to explore the possibility of multiple correlated sub-dimensions underlying the 52 test items (e.g., morphological or grammatical knowledge). This approach allowed us to assess the strength of unidimensionality and identify any secondary factors.

This analysis aimed to identify the underlying factor structure of the test items, checking for unidimensionality to ensure that all items collectively measure the same construct (productive academic vocabulary knowledge). The Kuder-Richardson Formula 20 (KR-20) was employed to investigate the reliability and

internal consistency of the test for both scoring methods in the full sample and within each subgroup.

Means for all items under strict scoring and relaxed scoring were calculated for both the full sample and each subgroup, facilitating the ranking of items based on mean difficulty. The items were ordered according to their frequency in COCA (COCA, n.d.), thus ranked from 1 to 52, corresponding to their order in the test. To investigate the assumption of a linear relationship between item frequency and difficulty, Spearman's *rho* was used to correlate the ranks of strict and relaxed scores with frequency ranks.

Then differences between the two scoring conditions and the two academic levels were explored. First, the differences between strict and relaxed mean scores were used to compute a mean increase for each item. This analysis provided insight into which items and sub-groups benefited most from the relaxed scoring approach. Items with large mean increases (>50%) were qualitatively analyzed to identify the linguistic difficulties that test-takers were relieved from when credited with a point under the relaxed scoring condition. Finally, to explore differences between the two educational levels under the two scoring conditions, *t*-tests were used.

Results

In this section we report the findings regarding the overall performance of the test, and the differences between the two scoring conditions and educational levels of the test takers.

Performance of the PAVT

The results of the PAVT under strict and relaxed scoring conditions for the full sample are presented in Table 1. For the full sample (*n*=232), the mean score was 22.63 (SD = 12.4) under strict scoring. Under relaxed scoring, this increased to 26.46 (SD = 11.55), representing a mean increase of 3.83 points This increase corresponds to an increase in scores of approximately 16.92%.

Table 1. PAVT scores by group and scoring condition

	Full sample (*n*=232)		**Master's (*n*=102)**		**HEP (*n*=130)**	
Scoring condition	*Strict*	*Relaxed*	*Strict*	*Relaxed*	*Strict*	*Relaxed*
Mean	22.63	26.46	18.39	22.86	25.96	29.28
Mean in %	43.52%	50.88%	35.37%	43.96%	49.92%	56.31%
95% CI *M* - Lower bound	21.03	24.96	16.47	20.96	23.67	27.17
95% CI *M* - Upper bound	24.24	27.95	20.32	24.77	28.26	31.38
5% Trimmed Mean	22.42	26.46	18.13	22.77	26.03	29.48
Median	21	25	18	22.5	26	30
Variance	153.84	133.32	96.00	93.84	174.92	147.04
Std. Deviation	12.40	11.55	9.8	9.69	13.23	12.13
Minimum	1	1	1	1	1	3
Maximum	51	51	45	45	51	51
Range	50	50	44	44	50	48
Interquartile Range	20	17	13	13	22	19
Skewness	0.26	0.08	0.42	0.19	-0.08	-0.19
Kurtosis	-0.86	-0.78	-0.26	-0.25	-1.08	-0.92
5th percentile	4	7	2.3	7	4	8.55
10th percentile	7.3	11.3	5.3	11	8	12.1
25th percentile	12	18	11	16	14.75	20.75
50th percentile (median)	21	25	18	22.5	26	30
75th percentile	32	35	24.25	29	36.25	40
90th percentile	41	43	32	35.7	42.9	45
95th percentile	44	46	36	39.85	47	48
Reliability KR-20	.95	.94	.92	.92	.96	.95

The exploratory factor analysis was conducted with the full sample data to determine the underlying factor structure of the test items, under both strict and relaxed scoring conditions. Overall, the analyses yielded positive results, indicating that the items perform well in testing productive knowledge of academic English vocabulary.

EFA was conducted for both the strict and the relaxed scoring conditions. Under strict scoring, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was .92, indicating that the data were suitable for factor analysis. Bartlett's Test of Sphericity was significant ($\chi^2 = 4696.69$, df = 1326, $p < .001$), further supporting the factorability of the correlation matrix.

The initial eigenvalues indicated that the first factor accounted for 28.63% of the variance, while the second factor explained 4.88%. The first factor was thus almost six times larger in explaining variance than the second factor, as illustrated

by the scree plot[1] (Figure 3). This suggests that the items predominantly load on a single factor, supporting the assumption of unidimensionality.

The large eigenvalue of the first factor (strict scoring: ~15.0; relaxed scoring: ~13.5) and the steep drop-off to the second factor (strict scoring: ~2.5; relaxed scoring: ~2.5) strongly suggest that the test is predominantly unidimensional. While secondary factors accounted for minor variance, their relatively small eigenvalues indicate that they represent subtle nuances within the overall construct rather than distinct sub-dimensions. These findings align with the test's intended purpose of measuring a single construct.

The factor loadings showed that all items loaded on the first factor with a mean loading of .44 (see Appendix 2). One item had poor loading (.16) but still a positive relationship between the item and the factor. Since most items showed substantial loadings on the first factor, the unidimensionality of the test was further reinforced.

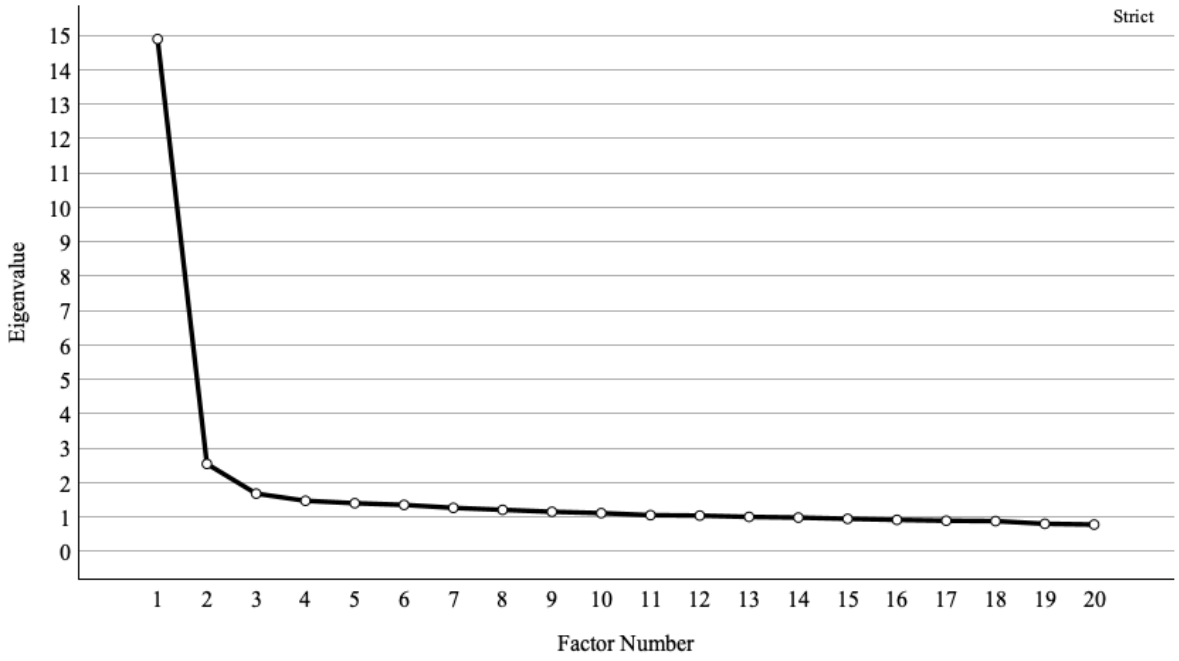

Figure 3. Scree plot, strict scoring

[1] The scree plot provides a visual representation of the eigenvalues, helping to determine the number of factors to retain. A large bend in the "elbow" (i.e., the point in the curve where the slope starts flatten out), indicates a substantial reduction in the amount of variance explained by subsequent factors.

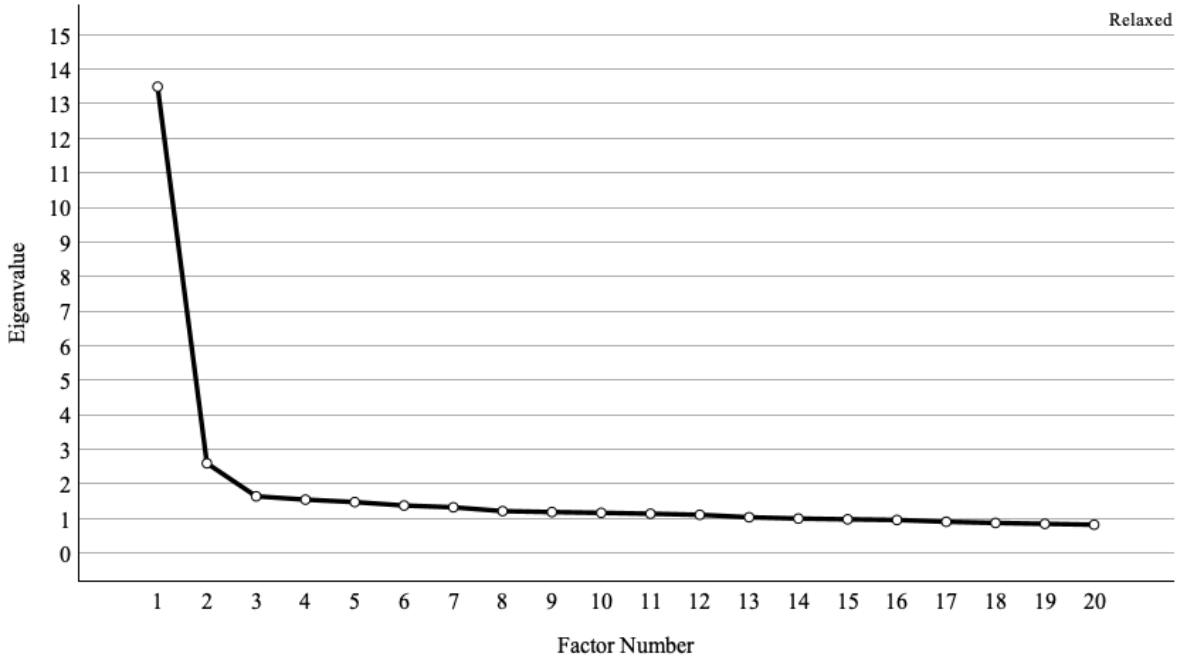

Figure 4. Scree plot, relaxed scoring

Under relaxed scoring, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy was .90, again indicating that the data were suitable for factor analysis. Bartlett's Test of Sphericity was significant ($\chi^2 = 4317.652$, df = 1326, $p < .001$), further supporting the factorability of the correlation matrix.

The initial eigenvalues indicated that the first factor accounted for 25.95% of the variance, while the second factor explained 4.99%. Thus, the first factor was more than five times larger in explaining variance than the second factor, as illustrated by Figure 4. This suggests that the items predominantly load on a single factor, supporting the assumption of unidimensionality.

The factor loadings yielded a mean loading of .42 across all items on the first factor (See Appendix 2). Three items had poor loadings (.16 - .19) but were still positively correlated with the first factor. Most items showed substantial loadings on the first factor, consistently affirming the unidimensionality of the test. Despite their relatively low values, the three items with poor loadings were retained because they showed a positive relationship with the primary factor and contributed to the conceptual coverage of the construct. Removing these items would have narrowed the scope of the test, and their inclusion did not adversely affect the test's reliability.

The range of scores remained similar across scoring conditions for both groups, with a slight reduction in variability under relaxed scoring. Both skewness and kurtosis values suggest a slight normalization of score distribution under relaxed scoring conditions for both groups. Reliability analysis using KR-20 indicated high internal consistency for the test under both scoring conditions, with values ranging from .92 to .96 (See Table 1)

As noted above, items on the test span the frequencies on the AVL, from *colleague*, with a frequency in COCA of 26,543, to *ubiquity*, which was found only 309 times in COCA at the time of the development of the AVL. To the extent that frequency is an important factor contributing to the amount of exposure any learner may have to a word, it would be expected that, on a general level, more frequent items would be answered more successfully than the less frequent items. To test this, the scores for the full sample were correlated with the item frequency ranks. Rank, rather than raw frequency, was used following Hashimoto's (2021) finding that the former gives a better indication of the extent to which frequency explains variation in scores. The resulting Spearman's rank-order correlation was found to be moderate for both the strict (*rho*=0.54) and relaxed (*rho*=0.51) conditions (see Table 2).

Table 2. Spearman's rank-order correlation between frequency and item difficulties

	COCA item ranks	Item difficulty - strict ranks
Item difficulty - strict ranks	.54**	-
Item difficulty - relaxed ranks	.51**	.90**

Note: *n*=52, representing the number of items and their respective mean score ranks from the full sample.
** p < .01

When plotting the item mean difficulties for each group (Figure 5), the analysis does indeed demonstrate increasing difficulty, with mean item scores generally decreasing through the test. This trend was consistent across both scoring methods and participant groups, as indicated by the downward slopes in the trend lines (Figure 5). However, the R^2 values (ranging from .22 to .29) suggest that item position only partially (approximately 26%) explains the variance in mean scores, indicating that other factors also influence item difficulty.

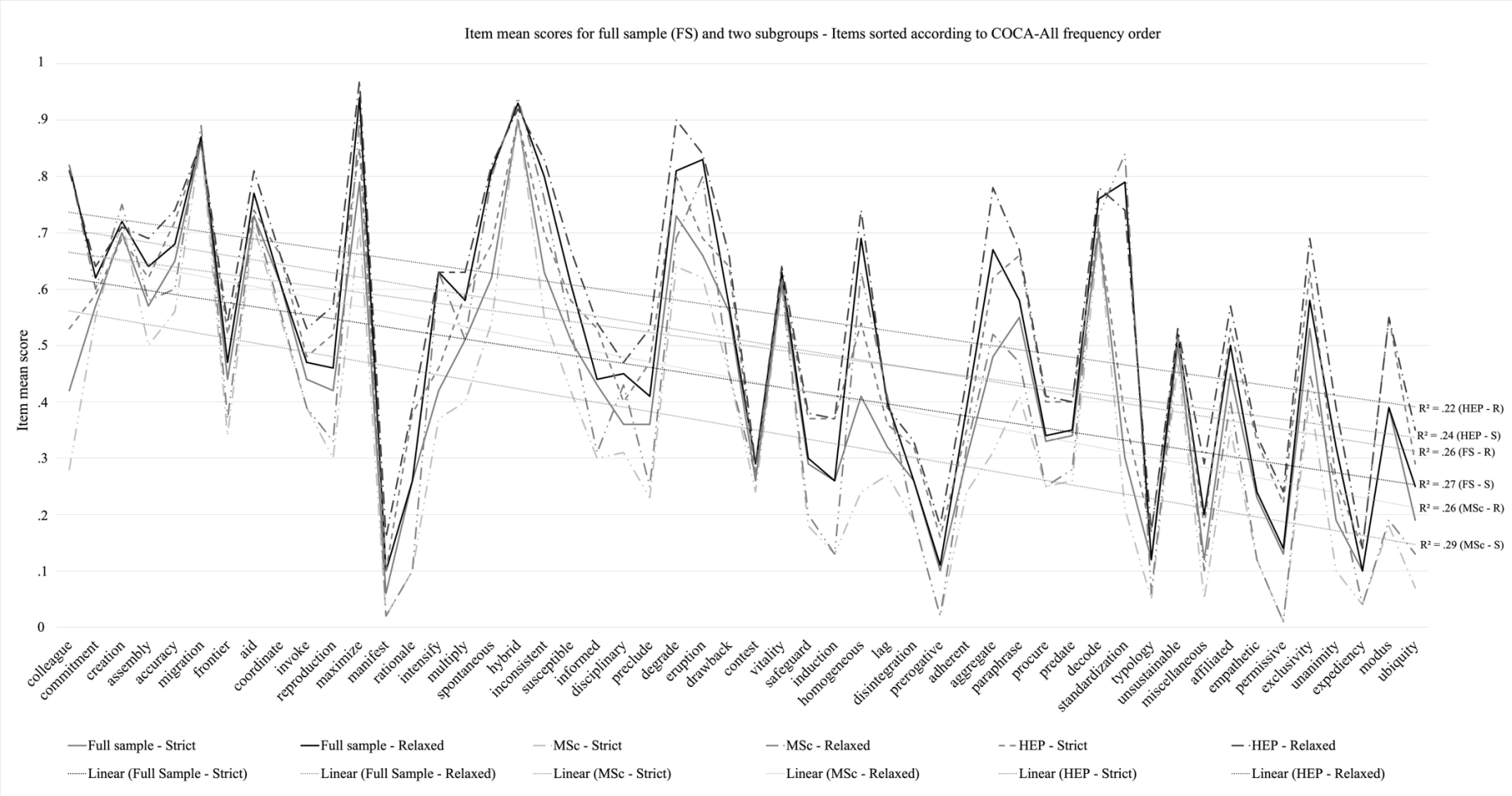

Figure 5. Item mean difficulty with trend lines

This is in line with earlier findings indicating that while frequency is an important factor in vocabulary learning and vocabulary knowledge, it is far from the only one (e.g., Hashimoto, 2021).

Scoring conditions and academic level

As noted above, and entirely as would be expected, scores increased across the board when relaxed scoring was applied. Interestingly, the master's students (*n*=102) benefitted more from the relaxed scoring condition. Their mean score under strict scoring was 18.39 (*SD* = 9.8), and this increased by 4.47 points to 22.86 (*SD* = 9.69) under relaxed scoring (an increase of 24.31%). By contrast, the HEP participants (*n*=130) had a mean score of 25.96 (*SD* = 13.22) under strict scoring, and this increased to 29.28 (*SD* = 12.13) under relaxed scoring, representing a mean increase of 3.32 points, i.e., an increase of 12.79%. This increase corresponds to a change of 24.31%, calculated as the relative change based on the strict scoring mean. To further analyze these observed differences, *t*-tests were performed and the results were visualized with boxplots.

Levene's test for equality of variances indicated that the variances between the master's students and HEP participants were not equal (strict scoring: $F = 15.763$, $p < .001$; relaxed scoring: $F = 10.316$, $p = .002$). To account for this, we used the *t*-test results that assume unequal variances. This adjustment provides a more accurate comparison of the group means under conditions where the variance assumption is not met.

Under the strict scoring condition, master's students ($M = 18.39$, $SD = 9.798$) scored significantly lower than HEP participants ($M = 25.96$, $SD = 13.226$), $t(229.294) = -5.006$, $p < .001$, Cohen's $d = -0.639$, indicating a moderate effect size. The boxplot in Figure 6 illustrates the group difference.

Similarly, in the relaxed scoring condition, master's students ($M = 22.86$, $SD = 9.687$) scored significantly lower than HEP participants ($M = 29.28$, $SD = 12.126$), $t(229.917) = -4.479$, $p < .001$, Cohen's $d = -0.577$, indicating a moderate effect size (see Figure 7).

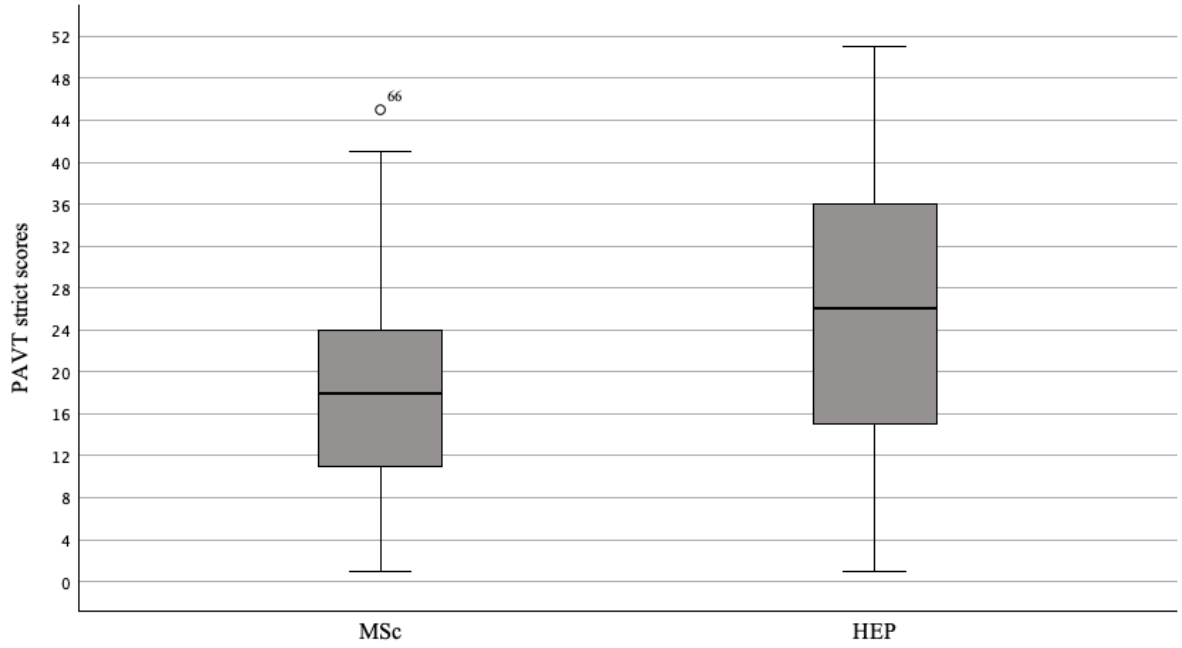

Figure 6. Boxplots, strict scoring

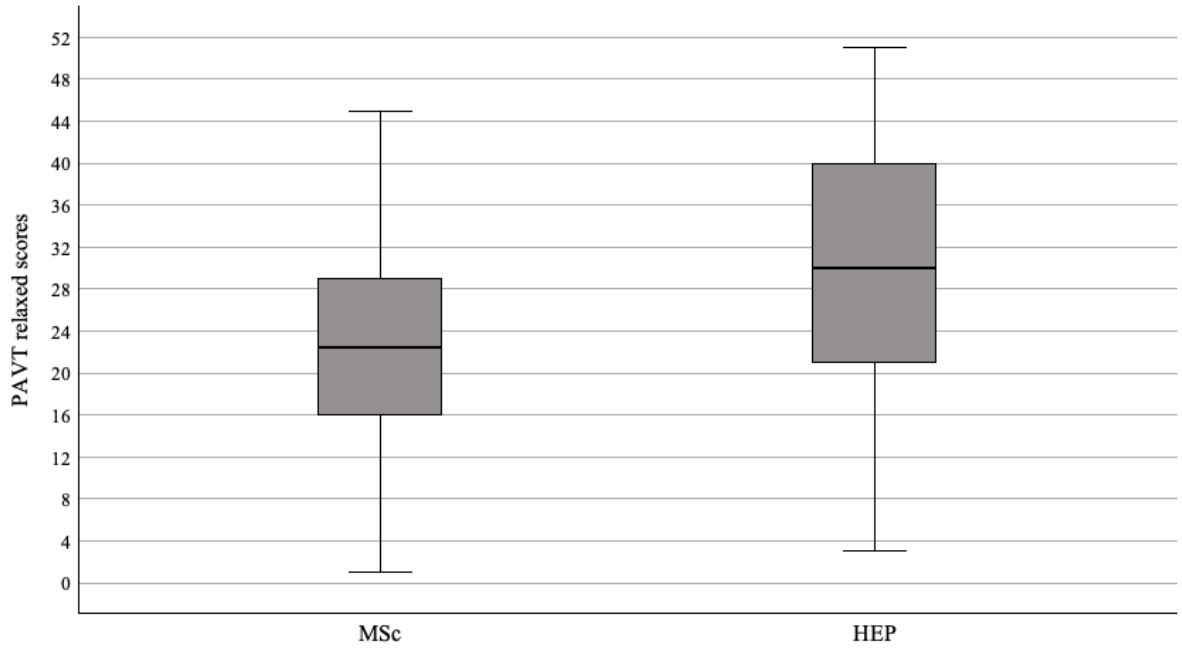

Figure 7. Boxplots, relaxed scoring

Discussion and conclusion

This paper has presented the development and validation of a productive test of academic vocabulary, the PAVT. The test is designed to provide insights into the

controlled productive knowledge of items from the AVL (Gardner & Davies, 2014). Results from the EFA and KR-20 demonstrate the reliability and validity of the test for this purpose, at least with respect to the type of participants to whom it was administered.

While frequency is, as expected, a contributor to the ease or difficulty of items on the test, it is clear that other factors play a role, and the approximately 26% explained variance from frequency corroborated the results of Hashimoto (2021) who found ranked frequency to account for 25% of the variance. In validating the PVLT, Laufer and Nation (1999) found a significant difference in performance between the vocabulary levels; that is, the scores for the items representing the 1,000 most frequent words in English were significantly higher than scores for the 2,000 most frequent words and so on. However, their method of analysis in considering frequency does not permit a direct comparison with the present data. A closer comparison is possible with the findings of Pecorari et al. (2019), who validated the AVT with a group of students in many ways similar to those in the present group, Swedish university students. In that study, the correlation between frequency and performance was found to be moderate with a Pearson's correlation of $r = 0.54$, which corresponds to the Spearman correlation in the present study between frequency and strict (*rho*=.54) and relaxed (*rho*=.51) scores. The fact that they found this similar correlation for essentially the same words (114 words are tested on the two forms of the AVT, of which 52 appear on the PAVT and the remainder were sampled according to the same principles), provides a validity argument regarding the performance of the PAVT.

A matter of interest is that the HEP participants scored significantly higher on the test than the master's students. In one sense, this may appear to be a logical and expected finding, given that in general terms people at a higher academic level would be expected to have larger vocabularies than those at a lower level. In Laufer and Nation's (1999) study, participants were exposed to English primarily in the English classroom for a small number of hours per week, and scores on the VLT did indeed increase from one year of school to the next. The participants in the present study, however, were not in or from environments in which exposure to English was as controlled or easy to estimate. Although information about their prior educational backgrounds is not available, they were a diverse and international group, and while some may have previously been educated in contexts with English as a medium of instruction, there is no reason to assume that all or most had. It is, however, a fact that the HEP participants had been in academia longer than the master's students. We tentatively interpret the higher scores for HEP participants as a byproduct of the status of English as a global academic *lingua franca* (e.g., Mauranen et al., 2016). Simply stated, regardless of the formal medium of instruction or university language policy, it is difficult to engage in academic activity without some exposure to academic English.

Not only did the mean scores for the master's students and HEP participants differ, there were differences in the extent to which the relaxed scoring condition benefitted them. This is in many ways unsurprising; productive vocabulary knowledge is not a binary (yes-no) condition. Test takers who are unable to satisfy the strict condition by providing an entirely correct answer may still be able to show some knowledge of the target word although with some inaccuracy. On a test of a closed set of words, the more words a test taker can demonstrate full knowledge of, the less scope there is to demonstrate partial knowledge. It is possible that this difference would not be observed if a free productive measure were used instead.

Nonetheless, this observed difference does underscore the importance of using both approaches to scoring (Webb, 2008), as a strict approach alone lacks the sensitivity to distinguish between partial productive knowledge of a word and no knowledge at all. It also demonstrates that for some groups, the additional sensitivity provided by this method of scoring is particularly beneficial. In other words, if only strict scoring had been used, a less full picture of test takers' vocabulary knowledge would have emerged, but the loss of information would have been particularly great in the case of the master's students.

This finding regarding the lower proficiency level of master’s students can also be interpreted, if tentatively, as having pedagogical implications for the development of academic vocabulary. In the context of assessment activities such as reports and case studies which ask students to produce running text, and to the extent that aspects of language use are assessed (as opposed to content knowledge alone), it is important to recognise and reward partially successful attempts to use words which are still trying to gain a foothold as part of the student's productive vocabulary.

It should be acknowledged that this test provides insights into a key dimension of vocabulary knowledge, which is only a portion of students' academic literacy needs and abilities. Academic vocabulary, such as the AVL from which the items on the PAVT are drawn, is a critical component of the vocabulary landscape, albeit one that Gardner and Davies (2014) found to account for just under 14% of the academic subsection of the COCA and the British National Corpus (BNC). The academic texts in COCA and BNC are primarily published works, i.e., written by relatively experienced academics. Malmström et al. (2018) found that the use of AVL items in student writing was even lower. Similarly, there is a limit to what can be extrapolated from a controlled productive measure to the ability to use words in authentic settings, which, in the case of the participants in this study, would mean activities such as writing assessment genres (e.g. lab reports and doctoral theses) and other academic genres (e.g. conference papers and research articles).

Beyond these points, which would apply to any controlled productive measure of vocabulary knowledge, the present study has a number of limitations. It was carried out with the participation of groups who are reasonably homogeneous in

that they have chosen to pursue postgraduate degrees in one country, Sweden, and have satisfied the same broad university admissions standards regarding English language proficiency in order to be able to do so. This means that it is not possible to assume that the PAVT, which performs well with this group, would necessarily do so with other groups.

In terms of L1 backgrounds, the participants were quite diverse, and this makes it impossible to even speculate about, much less analyse, the potential effect of cognacy of the test items with participants' L1s. Within the framework of the larger study for which this test was created, the ecological validity which the diversity of L1s conferred was desirable, making the inability to examine a cognate effect a reasonable trade-off. However, as Schmitt et al. (2020) rightly point out, the purpose of a test and the nature of the test takers are among a number of important considerations in test development, and it should not be assumed that tests have a one-size-fits-all property. While there is reason to believe that this test may be useful in contexts other than the one in which it was developed – and in principle, in any setting in which academic English is taught, or English is the medium of instruction – investigating the extent to which this is the case remains a question for future research.

Schmitt et al. (2020) also advocate for test developers to take a progressive, rather than a one-and-done approach; that is, to revisit and refine existing tests rather than treating them as carved in stone once they are developed. The PAVT was designed to be used in parallel with the receptive Academic Vocabulary Test (Pecorari et al., 2019). During the development of the PAVT, it became clear that some AVL items are not ideal for the format used in the PAVT (and the PVLT on which it is based). This has implications for future development. If alignment with the receptive test is not required for a given purpose, then a form of the PAVT with alternative AVL items added (ones which are more suitable for the format) would be advantageous. Conversely, for cases requiring parallel versions, alternative and aligned forms of the AVT and PAVT would be desirable. These are directions for future research, and one to which this paper, as a first step in the development of a new productive academic vocabulary test, can contribute.

Despite the recognition of the limitations on this study and possible avenues for future development, the PAVT provides a dual-faceted means of measuring productive knowledge of English academic vocabulary, using a strict scoring approach for fully correct answers and a relaxed scoring approach to credit words that are correct in meaning but may lack precision in form and use. This is only part of the set of language skills needed by university students (and academics), but it is an important one, and for a range of purposes it can be useful for researchers and educators to be able to measure productive knowledge of academic vocabulary. The present study represents a step in the development of a tool for this purpose.

Acknowledgements

We gratefully acknowledge the financial support from the Swedish Research Council (Vetenskapsrådet, under grant number 2013-2373) in the early stages of the project, and from City University of Hong Kong in later stages. We thank research assistants Christine Kan, HUI Chi Kin Ethan and Magnolia Manningo.

About the authors

Diane Pecorari is Professor of TESOL in the School of Education at the University of Leeds. Her research interests include English for academic purposes, educational linguistics, and research ethics. She co-edits (with Hans Malmström) the *Journal of English-Medium Instruction* and is the author of *Introducing English for Academic Purposes* (2nd ed., with Averil Coxhead).
Institutional affiliation: School of Education, University of Leeds, Leeds, LS2 9JT, United Kingdom
Email: D.Pecorari@leeds.ac.uk

Hans Malmström is Professor of Communication and Learning in Science at Chalmers University of Technology. His research is primarily concerned with aspects of second language acquisition (especially vocabulary) and the integration of communication and language, typically in contexts of English-medium instruktion. He co-edits the *Journal of English-Medium Instruction* (with Diane Pecorari).
Institutional affiliation: Department of Communication and Learning in Science, Chalmers University of Technology, 412 96 Gothenburg, Sweden.
Email: mahans@chalmers.se

Marcus Warnby is a senior lecturer in language education and deputy head of the Department of Education and Special Education at the University of Gothenburg. His research includes academic literacy, educational assessment, language testing, curriculum studies, and multilingualism. His experience includes certified language teaching for upper secondary school, lower secondary school principal, national language test development and teacher educator.
Institutional affiliation: Department of Education and Special Education, University of Gothenburg, Box 300, 40530, Sweden.
Email: marcus.warnby@ped.gu.se

References

Aizawa, I., & Rose, H. (2020). High school to university transitional challenges in English medium instruction in Japan. *System*, *95*, Article 102390. https://doi.org/10.1016/j.system.2020.102390

Alharbi, N. S. M. (2017). *An investigation into the academic writing: Difficulties of Saudi postgraduate students* [Doctoral dissertation, University of Exeter]. http://hdl.handle.net/10871/33113

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238. https://doi.org/10.2307/3587951

Coxhead, A. (2016). Reflecting on Coxhead (2000), "A new academic word list". *TESOL Quarterly*, *50*(1), 181–185. https://doi.org/10.1002/tesq.287

Fitzpatrick, T., & Clenton, J. (2017). Making sense of learner performance on tests of productive vocabulary knowledge. *TESOL Quarterly*, *51*(4), 844–867. https://doi.org/10.1002/tesq.356

Freimuth, H. (2020). Are academic English words learned incidentally? A Canadian case study. *BC TEAL Journal*, *5*(1), 32–43. https://doi.org/10.14288/BCTJ.V5I1.344

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305–327. https://doi.org/10.1093/applin/amt015

Hashimoto, B. J. (2021). Is frequency enough? The frequency model in vocabulary size testing. *Language Assessment Quarterly*, *18*(2), 171–187. https://doi.org/10.1080/15434303.2020.1860058

Hylan, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, *41*(2), 235–253.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Kiliç, M. (2019). Vocabulary knowledge as a predictor of performance in writing and speaking: A case of Turkish EFL learners. *Pasaa*, *57*(1), 133-164.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33–51. https://doi.org/10.1177/026553229901600103

Malmström, H., Pecorari, D., & Shaw, P. (2018). Words for what? Contrasting university students' receptive and productive academic vocabulary needs. *English for Specific Purposes*, *50*, 28–39. https://doi.org/10.1016/j.esp.2017.11.002

Malmström, H., & Pecorari, D. (2022). *Språkval och internationalisering: Svenskans och engelskans roll inom forskning och högre utbildning*. Språkrådet.

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

Masrai, A., & Milton, J. (2021). Vocabulary knowledge and academic achievement revisited: General and academic vocabulary as determinant factors. *Southern African Linguistics and Applied Language Studies*, *39*(3), 282–294. https://doi.org/10.2989/16073614.2021.1942097

Mauanen, A., Hynninen, N., & Ranta, E. (2016). English as the academic lingua franca. In K. Hyland & P. Shaw (Eds.), *The Routledge handbook of English for academic purposes* (pp. 44–55). Routledge.

McLean, S. (2021). The coverage comprehension model, its importance to pedagogy and research, and threats to the validity with which it is operationalized. *Reading in a Foreign Language*, *33*(1), 126–140.

Meara, P. & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16*(3), 323-337.

Meara, P., & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, *28*(1), 19–30. https://doi.org/10.1016/S0346-251X(99)00058-5

Paribakht, T. S., & Wesche, M. B. (1993). Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada Journal*, *11*(1), 9-29. https://doi.org/10.18806/tesl.v11i1.623

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

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109–120. https://doi.org/10.1017/S0261444819000326

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

The Open University, & Therova, D. (2020). General word lists: Overview and evaluation. *Vocabulary Learning and Instruction*, *9*(1), 51–61. https://doi.org/10.7820/vli.v09.1.therova

Uchihara, T., Eguchi, M., & Clenton, J. (2022). The contribution of guessing from context and dictionary use to receptive and productive vocabulary knowledge: A structural equation modeling approach. *Language Teaching Research*, Article 136216882211221. https://doi.org/10.1177/13621688221122138

Vu, V. D. (2023). Predictors of English medium instruction academic success in Vietnamese higher education. *Language Learning in Higher Education*, *13*(2), 411–430. https://doi.org/10.1515/cercles-2023-2029

Warnby, M. (2024). Relating academic reading with academic vocabulary and general English proficiency to assess standards of students' university-preparedness – the case of IELTS and CEFR B2. *Scandinavian Journal of Educational Research*. Advance online publication. https://doi.org/10.1080/00313831.2024.2318434

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, *30*(1), 79–95. https://doi.org/10.1017/S0272263108080042

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

Yamamoto, Y. (2014). Multidimensional vocabulary acquisition through deliberate vocabulary list learning. *System*, *42*, 232–243. https://doi.org/10.1016/j.system.2013.12.005

Zheng, Y. (2011). Exploring Chinese EFL learners' vocabulary depth knowledge: The role of L1 influence. *The Journal of Asia TEFL*, *8*(3), 191–219.

Appendix 1: PAVT and scoring guide

Productive Academic Vocabulary Test

In each sentence below, you are asked to fill in one missing word. The first letter or letters of the word are provided. Work as quickly as you can, but take as long as you need on each question. There is no penalty for guessing. Please note that the length of the line is not an indication of the length of the target word.

EXAMPLE

This is a t est____ of English academic vocabulary.

1. She is lucky to work with pleasant col_____.
2. Throughout his career he showed the qualities of dedication and comm_____.
3. After the defeat of the dictator came the cre_____of a democratic state.
4. The car is made on a modern asse_____line.
5. We have confidence in the ac_____of the statistics.
6. It was the biggest mig_____of groups of people in European history.
7. Humans have spread across earth; space is the final fr_____.
8. Fortunately, a nurse came to his ai_____.
9. When different people work on a project, someone needs to coo_____their efforts.
10. This is another argument which you can invo_____in support of the rule.
11. This isn't the original table; it's a repr_____.
12. The store raised prices to max_____profits further.
13. His suitability for the job was man_____; everyone recognized it.
14. He explained the ra_____behind the change.
15. She needed stronger medicine as the pain in her arm int_____.
16. Wash your hands regularly to keep germs from mul_____.
17. The trip was spon_____, not carefully planned.
18. Hy_____vehicles, which run on two kinds of fuel, are becoming more common.
19. The politician's statement about increased popularity is incons_____with the polls, which show decreasing popularity.
20. Small children are more sus_____to the disease.
21. Without more facts, we cannot make an inf_____decision.
22. The soldier faces dis_____action after disappearing without permission.
23. Eating your cake now precl_____eating it tomorrow.
24. The increase in cars has degr_____air quality in the city.
25. The volcano's sudden er_____made people leave the area.
26. There are advantages to that approach, but there are also dr_____.
27. Although they were against the decision, they had no way to cont_____it.
28. They were tired when the vacation began, but afterwards they had new vit_____.
29. Wearing a hat can saf_____against too much exposure to sunlight.
30. On the first day, new employees attended an ind_____.
31. Is it better for a society to be diverse or hom_____?
32. The second runner la_____behind the first by several minutes.
33. If other countries leave the union, steps must be taken to prevent its disi_____.
34. Higher education is often still the prer_____of the rich.

35. She is an adh_____of socialist views.
36. By aggr_____the numbers they got a larger data set.
37. You can either quote or pa_____the things you've read.
38. Strict purchasing rules apply when a government agency wants to proc_____goods and services.
39. This early ball game pred_____soccer and rugby, possibly by as much as 1,000 years.
40. The message looked like random letters until experts deco_____it.
41. Everyone must follow the same steps to ensure stan_____of our procedures.
42. Researchers have developed a new ty_____to group them into categories.
43. We can't keep on spending more than we earn; the situation is unsu_____.
44. The bag contains a mis_____assortment of candy, so there is something for every taste.
45. It's a small company, but it's af_____with one of the world's largest.
46. Defense attorneys want jury members to be em_____, understanding people.
47. In earlier times parents had many strict rules, but modern parents are pe_____.
48. The expensive furniture and carpets created an air of excl_____.
49. After discussion, the group has almost reached una_____on this issue.
50. The poor decision was the result of political exped_____.
51. The police are aware of the criminal's mo_____operandi.
52. The ub_____of mobile phones is remarkable; you see them everywhere.

Key

A	B	C
Question	**Strict** marking: **correct** answers	**Relaxed** marking: additional **correct** answers (examples of answers, not an exhaustive list)
1. She is lucky to work with pleasant **colleagues**.	colleagues	collauges, colleage, colleages, colleague, colleauges, collegue, collegues, colligues
2. Throughout his career he showed the qualities of dedication and **commitment**.	commitment	commitement, committement, committment
3. After the defeat of the dictator came the **creation** of a democratic state.	creation creator	create
4. The car is made on a modern **assembly** line.	assembly	assemble, assembley
5. We have confidence in the **accuracy** of the statistics.	accuracy	accurace, accuraty, acurusy
6. It was the biggest **migration** of groups of people in European history.	migration	
7. Humans have spread across earth; space is the final **frontier**.	frontier	frontaire
8. Fortunately, a nurse came to his **aid**.		aide, aids

9. When different people work on a project, someone needs to **coordinate** their efforts.	coordinate	
10. This is another argument which you can **invoke** in support of the rule.	invoke	invoque
11. This isn't the original table; it's a **reproduction**.	reproduction	reproduced, reproducible, reproductive, reprouduct
12. The store raised prices to **maximize** profits further.	maximise maximize	maximum
13. His suitability for the job was **manifest**; everyone recognized it.	manifest	manifested
14. He explained the **rationale** behind the change.	rationale	rationalization
15. She needed stronger medicine as the pain in her arm **intensified**.	intensified	intense, intensifies, intensities, intinsive
16. Wash your hands regularly to keep germs from **multiplying**	multiplying	multiplation, multiply, multiplication, multipling
17. The trip was **spontaneous**, not carefully planned.	spontaneous	sponatnous, spontamous, spontanious, spontaineous, spontinious, spontaneously
18. **Hybrid** vehicles, which run on two kinds of fuel, are becoming more common.	hybrid	hybride, hybridization
19. The politician's statement about increased popularity is **inconsistent** with the polls, which show decreasing popularity.	inconsistent	inconscistent, inconsistant, inconsisent, inconsistence, inconsisstence, inconssitant, incosisted
20. Small children are more **susceptible/susceptive** to the disease.	susceptible susceptive	susceptable, susceptibles, suscetible, suscipable, susseptible susuptable
21. Without more facts, we cannot make an **informed** decision.	informed	inform, informated
22. The soldier faces **disciplinary** action after disappearing without permission.	disciplinary	disiplinary, displinary, dissiplinairy, disipline
23. Eating your cake now **precludes** eating it tomorrow.	precludes	preclude, precluding
24. The increase in cars has **degraded** air quality in the city.	degraded	degrad, degradated, degrade degrading
25. The volcano's sudden **eruption** made people leave the area.	eruption	erapted, errupted, erruption, eruptation, erupted
26. There are advantages to that approach, but there are also **drawbacks**.	drawbacks	drawback
27. Although they were against the decision, they had no way to **contest/contradict** it.	contest contradict	

28. They were tired when the vacation began, but afterwards they had new **vitality**.	vitality	vitalisation
29. Wearing a hat can **safeguard** against too much exposure to sunlight.	safeguard	safguard
30. On the first day, new employees attended an **induction**.	induction	
31. Is it better for a society to be diverse or **homogeneous**?	homogeneous	homogeneon, homogenious, homogenius, homogenous, homogineous
32. The second runner **lagged** behind the first by several minutes.	lagged	lag, lagging
33. If other countries leave the union, steps must be taken to prevent its **disintegration**.	disintegration	disintigration
34. Higher education is often still the **prerogative** of the rich.	prerogative	prerogertive
35. She is an **adherent/adherer** of socialist views.	adherent adherer	adhearent, adhere, adherend, adherrent,adherant
36. By **aggregating** the numbers they got a larger data set.	aggregating	aggregate, aggregate, aggregation, aggrigate, aggrigatting, aggrigation
37. You can either quote or **paraphrase** the things you've read.	paraphrase	parafrace
38. Strict purchasing rules apply when a government agency wants to **procure** goods and services.	procure	
39. This early ball game **predated**/**predates** soccer and rugby, possibly by as much as 1,000 years.	predated predates	
40. The message looked like random letters until experts **decoded** it.	decoded	decode
41. Everyone must follow the same steps to ensure **standardisation** of our procedures.	standardisation standardization	standard, standardation standards
42. Researchers have developed a new **typology** to group them into categories.	typology	
43. We can't keep on spending more than we earn; the situation is **unsustainable**.	unsustainable	unsusteinable
44. The bag contains a **miscellaneous** assortment of candy, so there is something for every taste.	miscellaneous	miscelaneous, miselinious, misellonous
45. It's a small company, but it's **affiliated** with one of the world's largest.	affiliated	affiliate, afiliated

46. Defense attorneys want jury members to be **empathetic** understanding people.	empathetic	
47. In earlier times parents had many strict rules, but modern parents are **permissive**.	permissive	
48. The expensive furniture and carpets created an air of **exclusivity/exclusiveness**.	exclusiveness exclusivity	
49. After discussion, the group has almost reached **unanimity** on this issue.	unanimity	unanimity, unanimous
50. The poor decision was the result of political **expediency/expedience**.	expedience expediency	
51. The police are aware of the criminal's **modus** operandi.	modus	
52. The **ubiquity/ubiquitousness** of mobile phones is remarkable; you see them everywhere.	ubiquity ubiquitousness	ubicuosnes, ubiguity, ubquety, ubiquitous, ubquitiousness

Appendix 2. Factor loadings on the first factor, full sample, factor 1

PAVT item number	Strict loadings	Relaxed loadings
1	0.367	0.238
2	0.290	0.333
3	0.342	0.335
4	0.342	0.373
5	0.417	0.410
6	0.208	0.211
7	0.489	0.509
8	**0.160**	0.233
9	0.270	0.316
10	0.349	0.409
11	0.382	0.399
12	0.306	**0.186**
13	0.452	0.439
14	0.672	0.726
15	0.440	0.361
16	0.401	0.428
17	0.312	0.244
18	0.209	**0.190**
19	0.441	0.285
20	0.497	0.447
21	0.566	0.588
22	0.369	0.401
23	0.621	0.610
24	0.394	0.358
25	0.328	0.314
26	0.278	0.314
27	0.268	0.234
28	0.378	0.403
29	0.532	0.552

30	0.396	0.392
31	0.454	0.252
32	0.430	0.412
33	0.495	0.474
34	0.525	0.602
35	0.498	0.485
36	0.502	0.438
37	0.470	0.463
38	0.503	0.528
39	0.396	0.456
40	0.339	0.300
41	0.511	**0.157**
42	0.494	0.511
43	0.451	0.396
44	0.526	0.554
45	0.512	0.455
46	0.560	0.596
47	0.612	0.576
48	0.509	0.489
49	0.616	0.544
50	0.403	0.462
51	0.687	0.703
52	0.677	0.637