Paper:

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.
http://dx.doi.org/10.1016/j.jeap.2017.12.006

Measuring the contribution of academic and general vocabulary knowledge to learners’ academic achievement

Abstract

The Academic Word List (AWL) (Coxhead, 2000) is widely used in preparing non-native speakers for academic courses, and it is thought that these words are essential for the understanding of English academic texts (Cobb & Horst, 2004). It is also thought the AWL is an infrequent and specialised list inaccessible from general language. These preconceptions are challenged in this study which demonstrates with reference to BNC/COCA word lists that the majority of the AWL fall within the most frequent 3,000 words in English, a grouping which Schmitt and Schmitt (2014) describe as highly frequent. Using a specially created test of the AWL and the XK-Lex test of overall vocabulary size (Authors, 2012), the study demonstrates that the learning of the AWL appears strongly influenced by the frequency of these words in general corpora and that the AWL test resembles very strongly a test of overall vocabulary size. When scores from these tests are related to a Grade Point Average (GPA) measure, it appears that knowledge of the AWL adds only marginally to the explanatory power of overall vocabulary size in explaining variance in GPA scores. This conclusion matches that of Townsend et al. (2012) although the tests in this study appear to have a greater explanatory power.

Keywords: Academic vocabulary, receptive knowledge, vocabulary size, Rasch model, test validity

1. Background

1.1. The academic word list and academic success

Horst and Cobb (2006), in considering the impact that recent vocabulary acquisition research has had on the English as a foreign language (EFL) world, suggest that most research on the lexicon is considered peripheral to the design and content of EFL programmes. They go on to point out, however, that Coxhead's (2000) Academic Word List (AWL) is an exception to this. The AWL is a list of the words that are important to communicating the concepts taught either in school, university or in English for academic purposes (EAP) programmes. There is support in the academic literature for this claim. Knowing these words is thought essential to improve L2 learners' comprehension of academic written text (Cobb & Horst, 2004; Dang & Webb, 2014). Gardner and Davies (2014) ascribe academic vocabulary a central role in school success for both native and non-native speakers. Not surprisingly therefore, the AWL has become central to the teaching of EAP.

Coxhead's AWL comprises 570 headwords drawn from a 3.5 million corpus of academic texts drawn from a wide range of academic genres. The criteria for the creation of the list are that these words should be:

a) *specialised*, so "…[t]he word families had to be outside the first 2,000 most frequently occurring words of English as represented by West's (1953) GSL" (Coxhead, 2000, p. 221).
b) *generic to academic discourse*, rather than specialist vocabulary items restricted in use to only limited subjects, and a *Range* criterion was used to ensure this was occurring.
c) *frequent*, so items on the list had to occur at least 100 times in the academic corpus.

In principle, therefore, the AWL represents a selection of vocabulary that is not drawn from the most frequent words (Kremmel & Schmitt, forthcoming) but appears to be thought of as something beyond, perhaps well beyond, basic levels of vocabulary knowledge (Coxhead, 2000) and which is unlikely to be accessed through general language exposure (Townsend, Filippini, Collins, & Biancarosa, 2012).

The rationale for the importance of the AWL comes primarily from the evidence of the contribution to coverage which the list provides. The AWL is generally thought to provide approximately 10% coverage of academic written texts (e.g., Chen & Ge, 2007; Cobb & Horst, 2004; Coxhead, 2000). This list, with the knowledge of the words in West's (1953) General Service List (GSL), gives about a 90% coverage of academic written text (Nation, 2004). Just how important the list is may vary from one subject domain to another. Coxhead's (2000, p. 222) figures suggest that the AWL is more useful in her Commerce sub-corpus than that in the Science sub-corpus, for example. In Commerce the AWL by itself contributed 12% to coverage and, combined with the GSL, comprised 88.8% of the sub-corpus. In Science, the AWL contributed 9.1% to coverage and, with the GSL, comprised 79.8% of the sub-corpus. In a recent study, Dang and Webb (2014) investigated the coverage of the AWL in academic spoken English by analysing the vocabulary in 130 lectures and 39 seminars from four sub-corpora of the British Academic Spoken English (BASE). Their findings suggest that the AWL accounts for 4.41% coverage of academic spoken English, and that its coverage in each sub-corpus varied from 3.82% to 5.21%. They conclude that with aid of the AWL and knowledge of proper nouns and marginal words, learners will need a vocabulary size of 3,000 and 8,000 word families to attain 95% and 98% coverage of academic spoken English, respectively.

Nonetheless, studies appear to support the importance of the AWL in academic texts across a variety of academic fields, such as Engineering (Mudraya, 2006), medical research (Chen & Ge, 2007), and applied linguistics (Chung & Nation, 2003). The use of the AWL sub-lists in setting goals for learning, is thought to be additionally useful in promoting a significant improvement in learners' overall vocabulary knowledge (Snow, Lawrence, & White, 2009).

1.2. The AWL and general vocabulary

One criticism that is repeatedly levelled at the AWL is the difficulty that exists in distinguishing the words it contains from frequent general vocabulary (Gardner & Davies, 2014). Coxhead's claim that the words of the AWL are specialised and distinct from general frequent vocabulary falls down when the words in the list are compared with word frequency list drawn from general language sources. A comparison of the AWL with Kilgarriff's (2006) lemmatised frequency lists and the BNC/COCA word family lists (Nation, 2012) is made in Table 1 which summarises the frequency distribution of words in the AWL.

[*Table 1 about here*]

Kilgarriff's (2006) BNC lists are lemmatised and are constructed slightly differently from the AWL which, like the BNC/COCA lists are based on word families. Base words in the AWL will include morphemic derivations which, in Kilgarriff's lists may be counted as two or more words. The BNC and BNC/COCA lists are also slightly different with a different distribution of words across the 1,000 word groups. Despite these differences, it is clear from both just how heavily weighted to the most frequent words the AWL is. In Kilgarriff's lists 369 AWL words, 65% of the total, fall within the first 3,000 words. In the BNC/COCA lists 484 AWL words, 84% of the total, fall within the most frequent 3,000 words, a grouping which Schmitt and Schmitt (2014) call high-frequency. But there is also a spread of frequency in the AWL with a small number of items appearing highly infrequent, and occurring at or beyond the 8,000 word level.

Viewed from this perspective, and taking the AWL as just a list of words, it is not clear how specialised the majority of the AWL really is. Estimates which link lexical size to hierarchies of performance such as the Common European Framework of Reference (CEFR) (e.g., Authors, 2009) suggest that learners at B2 level or above are likely to know most vocabulary that falls within the first 3,000 words of English and so are likely to have good knowledge at least of the words in the AWL. Given the frequency of so much of the AWL, and the spread of the remainder across the whole list and across the frequency bands, it seems very likely that it should be accessible in general language use, and it seems much more likely that its acquisition should be linked to frequency just as all other vocabulary appears to be.

This perspective on the AWL raises a further issue as to whether it is general vocabulary size or specialist vocabulary knowledge which is the most important criterion for academic success. While AWL words are arguably very important in the handling of academic discourse, the 2,508 word families of the GSL and AWL and 90% coverage they provide, are not generally considered sufficient to achieve good comprehension of any text. Laufer and Ravenhorst-Kalovsky (2010) suggest two coverage figures would be needed for something like the fluency needed to handle English texts in the context of formal study. They give a minimal coverage figure of 95% for adequate comprehension, which they suggest would not satisfy most educators, and an optimal figure of 98% coverage for significantly better comprehension associated with ‘functional independence in reading’ (p. 25). The minimal coverage figure requires knowledge of the most frequent 4,000 to 5,000 word families in English and the optimal coverage figure requires knowledge of the most frequent 8,000 word families. The implication of this is that learners are likely to need double or three times the volumes of vocabulary provided by the GSL and AWL in order to achieve the levels of knowledge, and the command in reading, needed for academic study.

The importance of growing a large English lexicon for academic study through English has been explored by a body of research (e.g., Biemiller, 2001; Hart & Risley, 2003; Authors, 2013) and this suggests a positive correlation exists between the two. The significance of the lexicon in particular to academic success lies in its importance for overall language development. The possession of a lexicon of the right size and quality is essential for good language performance, and good language performance is essential for academic achievement (Authors, 2017). Thus, research points to the idea that moderate to high correlations exist between general vocabulary size measures and performance in the four skills as measured by tests of academic English such as IELTS. The correlations are typically between 0.6 and 0.8 so not surprisingly, overall vocabulary size alone is often capable of explaining over 50% of variance in scores in foreign language performance (e.g., Stæhr, 2008). This leads to the suggestion that vocabulary size is not just a major factor but is *the* major factor in explaining variation in language performance (e.g., Author, 2013; Stæhr, 2008).

1.3. What makes AWL words special?

If AWL vocabulary is not, as a list of words, specialised then a further case has to be made for the AWL and its relevance to teaching English for academic study. One argument mounted in its defence is the way many of the words it contains are polysemous and that AWL words, as Hyland and Tse (2007) point out, can have specialist meanings and usage relevant to academic language particularly, and may be used with specific and different meanings across the different academic disciplines. One meaning, probably a core meaning, will be known by learners even at a quite modest level of English performance but it is these specialised meanings which need to be taught for academic purposes. Gardner and Davies (2014) point out that this can also be said of some words in the GSL and cite the example of *interest* which has a major academic meaning in the domain of Commerce, in addition to a more general meaning. It is not at all clear how many AWL words this argument might apply to. However, defending a generic AWL by arguing that the important meanings which some of these words contain are highly subject specific, seems unsatisfactory. We would argue too that specialist meanings for at least some of these words, and often too the phrases they occur in, are accessible from the core meanings of these words. Nation (2001) points to the relevance of the AWL word *demand* to Commerce with a special meaning and usage that is not relevant to, for example, Biology. But an understanding of the word *demand*, even in its most general form, ought to allow Commerce related expressions like *demand curve* to be well understood.

Ward (1999) cuts across all this difficulty by arguing that general word lists, whether the GSL or the AWL, can only distract learners from the vocabulary which is needed for specialist subject study, by introducing words they will never encounter or need to use. He argues that a far lighter vocabulary burden than is usually assumed is needed in some subject areas. Thus, in his analysis of Engineering textbook material he concludes that the most frequent 2,000 words can provide over 95% coverage even among the more specialist sub-corpora he identifies. Nonetheless, Gardner and Davies (2014, p. 6) argue that "it is crucial to identify a statistically viable list of core academic words that can be focussed on in academic teaching and research".

A more convincing line to argument in favour of the AWL comes from studies which attempt to show that knowledge of the AWL makes a distinct and important contribution to success in academic performance, and that this contribution is separate from the contribution made by knowledge of other words as in general vocabulary knowledge. Roche and Harrington (2013) attempt to understand how vocabulary and academic performance link and results from their study suggest that vocabulary size can explain about 25% of the variance in students' academic performance as measured by GPA. Townsend et al. (2012) go further and attempt to disambiguate the contributions to GPA of general vocabulary knowledge and knowledge of the AWL specifically. Using regression analysis, they argue they can measure these contributions both separately and combined. They conclude that the explanatory power of vocabulary size as a whole is greater than that of academic word knowledge, and that general vocabulary size can explain between 26% and 43% of variance according to discipline. The contribution of AWL knowledge is substantial but lower than that of general vocabulary knowledge. However, academic word knowledge can contribute a unique variance to attainment across disciplines even when the overall breadth of vocabulary knowledge is controlled for. Knowledge of the AWL can add an additional 2% to 7%, depending on discipline, to this explanation of variance provided by general vocabulary knowledge. This appears to suggest that growing a suitably big vocabulary is most important for success but that knowledge of the AWL has some additional and marginal effect on academic performance.

Our interpretation of this paper is that the two tests Townsend et al. (2012) used should perform very similarly, and that scores from the two tests should give similar results. In principle, general vocabulary size and specialist academic vocabulary knowledge are two different constructs. In practice, however, and as the figures for the frequency of AWL words in Table 1 demonstrate, the AWL is a list of words which can be found in the frequent section of general word lists, and so both tests are likely to give a broad estimate of general vocabulary knowledge. Without a better understanding of the way these two test inter-relate, it is not possible to understand the separate contributions that these two types of vocabulary knowledge can make to academic success. Nonetheless, a testing approach does appear a very useful way of demonstrating the benefit of teaching and learning the AWL, if a benefit exists.

It is the intention of the current paper, therefore, to revisit the conclusions made by Townsend et al. (2012) and, using a different group of learners, to see whether knowledge of the AWL and general vocabulary size can be shown to make separate contributions of academic performance. Townsend et al. (2012) note the absence of a standard test of AWL knowledge and, consequently, an Academic Vocabulary Size Test (AVST) has been created for this paper. It is intended too, to compare learners' performance on both vocabulary tests to see whether the AWL is also testing vocabulary size, and whether the two constructs of general vocabulary size and specialist knowledge of the AWL can be meaningfully distinguished.

2. Aims and method

2.1. Aims

This study addresses four broad research questions:

1. Does the newly created AVST perform reliably?
2. Can AWL knowledge be shown to be related to word frequency derived from general corpora? The validity results of the AVST for the purpose it was devised are presented under this research question.
3. What are the contributions of AWL knowledge and general vocabulary size to academic success?
4. Are tests of the AWL and general vocabulary knowledge testing the same factor?

In order to answer these questions, data have been collected from users of English as a foreign language taking academic courses, using tests of both the AWL and general vocabulary size. The results of these tests are compared, as in Townsend et al (2012), with the GPA of these participants.

2.2. The Academic Vocabulary Size Test (AVST)

The test is designed to measure the written receptive academic vocabulary knowledge of non-native speakers of English of the 570 academic word families presented in the AWL (Coxhead, 2000), and is designed as a checklist test (see Author, 2009). According to Gyllstad, Vilkaite, and Schmitt (2015), a sampling rate of higher than 1:100 is probably needed in order to better represent the underlying population of words in the corresponding frequency bands in L2 tests. As the number of words in the AWL is small, a sampling rate of 1:5 was employed.

Applying this sampling rate, the AVST comprises 114 items divided into six equal frequency bands, each including a sample of 19 words. The words are a principled sample from the highest frequency words in the AWL to the lowest. The test, further, includes 19 control words which were assigned to a seventh column in the test to adjust for guesswork when calculating the final score, as in Authors' *X-Lex* (2003). The control words are very rare words, beyond the 25,000 word level in Thorndike and Lorge's (1944) word list and are taken from Goulden, Nation, and Read's (1990) test. Results from both Goulden et al.'s study and Authors' (2013) study suggest knowledge of these words in university populations is negligible and learners would not know them. As the test includes 114 real items and 19 control items, each real item is given a credit of 5 points to get a total of 570 and each control item is given 30 points to have an equivalent score of 570 when adjusting for guesswork.

The test-takers were presented with each test word in turn and were requested to make a decision whether they know each of these words. The test was designed to take no longer than 10 minutes to complete. This short duration encourages the participants to respond to all items in the test without being influenced by the test length. Two forms of the AVST, A and B, were created (see Appendices A and B). In these two forms, the frequency bands were organised from the most frequent band to the least frequent band (left to right).

In order to make a calculation out of 570 of the test-takers' academic vocabulary size, each *Yes* response to a real word in the test is given a score of 5 to form an unadjusted score, and each false *Yes* response to a control word deducts 30 points to give an adjusted score. This final, adjusted, score is expected to represent a learner's total academic vocabulary knowledge.

2.3. The XK-Lex test of general vocabulary size

To test for overall vocabulary size, the XK-Lex 10,000 word test of vocabulary knowledge (Authors, 2012) was used. This test takes a principled sample from across 10,000 word range and there are 10 test words from each 1,000 word band. The version chosen contains no words from the AWL. There are also 20 false words, used as a control for guesswork. In order to make a calculation out of 10,000 the *Yes* responses to real words are totalled and multiplied by 100 to give an unadjusted score, and each *Yes* response to the false words deducts 500 points from the unadjusted score to give a total adjusted score which, it is thought, captures learners' general vocabulary size best.

2.4. GPA as a measure of academic success

As in Roach and Harrington's (2013) study, a calculation for GPA was made for the students' end of year academic exams and this was used as a metric for academic achievement.

2.5. Participants and procedure

232 students participated in the current study.

96 undergraduate native Arabic speaking students (64 males and 32 females), taking English language courses at a university in Saudi Arabia, provided results which were used to investigate the relationship between academic vocabulary knowledge, general vocabulary size and academic achievement. These students' results from XK-Lex, form A of the AVST and details of their GPA scores were used in order to address the second research question.

Two further groups were used to check the performance of the AWL knowledge test, and to help disambiguate the general vocabulary size from knowledge of the AWL specifically in predicting GPA. The first group included 16 English native speakers, all Doctoral level, who took both forms of the test and whose results were used in a pilot application to check the practicality of the test and to check the difference in the responses to the real and control words. The second group comprised 120 non-native speakers (69 males and 51 females) who were enrolled in three levels of study (Bachelor, Master and Doctoral) at British universities. These students took both forms of the test.

2.6. Analyses

Test data were input into an Excel spreadsheet and then exported to WINSTEPS 3.73 (Linacre, 2011) and calibrated using the Rasch dichotomous model (Rasch, 1960). The Rasch model has several strengths, such as offering equal interval item and person statistics, item and person fit indices, item and person reliability measures, and detailed information regarding item and person dimensionality. The data were also processed into SPSS 22 for further reliability measures, correlational and regression analyses.

Since the AWL knowledge test is a new creation, a number of results are presented to show it is working satisfactorily. These include, a pilot test with native speakers to check the performance of the control items, and reliability calculations including Rasch analysis. Following this, the relationship between the two vocabulary knowledge tests and GPA is reported using correlations and regression analysis. Finally, to assess whether the two vocabulary tests are testing the same or different constructs, collinearity and factor analysis are reported.

3. Results

3.1. Does the AVST perform reliably?

The results obtained from the 16 English native speaking participants taking both forms of the AVST are as shown in Table 2. Hits are the number of *Yes* responses to real words and False Alarms (FA) are the number of *Yes* responses to the control words. All participants achieved the maximum score and identified both the real and false words correctly. This indicates the test is practical and appears to be performing as intended.

[*Table 2 about here*]

In order to assess the reliability of the AVST, mean scores from the 120 non-native speakers on both forms of the AVST, the parallel forms tests (A and B), and the two split-halves internal consistency test conducted on test A, are summarised in Table 3.

[*Table 3 about here*]

The parallel forms of the AVST test produce scores which are nearly identical. The difference between the means is not statistically significant (t = -.23, p = .815). Also, the mean scores in the split-half pairs 2 and 3, are, again, very similar and the differences between the mean scores are not statistically significant (*t* = .61, *p* = .543; *t* = .74, *p* = .46, respectively).

The reliability of the AVST was measured by comparing the scores obtained from the two parallel forms of the test, A and B. The results indicate a strong and statistically significant correlation (*r* = .96, *p* < .001). Internal consistency reliability, on the other hand, was further examined within test A. A split-halves method (e.g., Bachman, 2004) was utilised. Thus, form A was split into odds-evens and top-bottom pairs. The results show high levels of correlations between odds-evens (*r* = .86, *p* < .001) and top-bottom (*r* = .84, *p* < .001).

The final reliability measure for the AVST was to perform a Cronbach's alpha analysis and the results are summarised in Table 4.

[*Table 4 about here*]

The result shows high reliability indices for all the test pairs (all were above .90). DeVellis (2003) proposes that the Alpha score is very good when it falls between .80 and .90 and excellent when it is above .90. These findings provide support for the test's reliability to estimate test-takers' knowledge in a consistent manner. There is no obvious obstacle to using this test for the intended analyses.

3.2. Is AWL knowledge related to word frequency from general corpora?

In vocabulary size tests development, it is generally assumed that items at successively decreasing frequency levels will form a continuum (e.g., Beglar, 2010; Milton, 2009, Read, 2000). Similarly, in this study it was assumed that the AVST items would form a difficulty continuum based on their frequency in the AWL (Coxhead, 2000), given the probability that higher-frequency words will be learned than lower-frequency words. The relationship of scores from the AWL test items with frequency data is summarised in Table 5 and Figure 1.

[*Table 5 about here*]

[*Figure 1 about here*]

The results show that the frequency profile is observable at each frequency band examined and confirm the item hierarchy in terms of difficulty order, where more knowledge is observed in the high frequency bands than less frequent bands. A Friedman analysis confirms that the observable trend in the summary data is statistically significant ($\chi 2$ = 508.38, $p <$ .001). The results show a clear effect for frequency, as drawn from general language corpora, in the participants' knowledge of the AWL.

The Wright map in Figure 2 summarises the results from Rasch analysis and records the relationship between person measures and the item calibrations which are a product of word frequency.

[*Figure 2 about here*]

The results of Rasch analyses show very good reliability for both items and persons, at .92 and .95, respectively. This indicates the stability of the item difficulty and person ability hierarchies and also clearly shows the effect of frequency in the participants' responses. In the light of these results it can be suggested that the substantive aspect of construct validity of the AVST is met. Rasch analyses used in this study also suggest that the AVST has a good content aspect of construct validity, such as representativeness (the degree to which a test is sensitive to variations in the construct being measured) (Borsboom et al., 2004) and technical quality. These results are discussed in some detail in the discussion section.

Another validity measure of AVST performed in the study is the structural aspect of construct validity. The structural aspect of construct validity concerns the assumed dimensionality of the measured construct (Messick, 1989). The assumption was that the AVST measures a single primary construct, receptive academic vocabulary knowledge. This assumption was tested using a Principle Axis Factoring (PAF).

The PAF analysis indicates that there is only one Eigen value above 1, which explained some 78.19% of the total variance, and this suggests that the test is measuring only one construct. The factor loadings and Scree plot summarising this analysis are presented in Table 6 and Figure 3. The Scree plot clearly indicates a sharp drop between the first and the other five factors. The Scree plot and the amount of variance explained by the first factor offer evidence of a single construct underlying the test. The PAF extraction method of only one factor also reveals that all items show a high level of loading on this factor. All factor loadings were greater than .7, which is above the minimum baseline loading suggested by Brown (2006), indicating that only one factor underlies the scores and that the uniformity of loadings on this factor is high.

[*Figure 3 about here*]

[*Table 6 about here*]

3.3. *What are the contributions of academic and general vocabulary knowledge to academic achievement?*

The relationship between academic vocabulary size measured using the AVST, overall vocabulary size measured using XK-Lex, and learners' academic performance, as measured by GPA, is quantified using correlation and regression analyses. The results show a strong, positive correlation between learners' academic performance and receptive academic vocabulary knowledge (*r* = .73, *p* < .001), and overall vocabulary size and academic performance (*r* = .68, *p* < .001). The two vocabulary tests also correlate (*r* = .78, *p* < .001). The correlations are statistically significant.

[*Table 7 about here*]

Regression analysis, reported in Table 7, indicates that AVST and XK-Lex can be used to explain variance in GPA scores and that a model including both can be used to explain additional variation as performed in the Townsend et al.'s (2012) study. Although a regression analysis is generally not recommended for a case where the two tests correlated significantly, Townsend et al. have demonstrated that such an analysis may provide some insightful information. The results indicate a large effect size of academic and general vocabulary, when combined, on learners' academic achievement.

3.4. Are tests of the AWL testing general vocabulary knowledge?

Given the frequency distribution of AWL words it is questioned whether a test of AWL knowledge can test something different from overall vocabulary size. In this study vocabulary size has been measured using the XK-Lex test. The correlation between AVST and XK-Lex (*r* = .78, *p* < .001) is high and is statistically significant. This might suggest the probability of multicollinearity and consequently collinearity diagnostics were performed. The result, however, suggests no presence of collinearity (tolerance > .02 and value for VIF is <5). The absence of collinearity is taken to support the use of regression analysis using results from these two tests notwithstanding their high correlation.

Factor analysis results, examining whether different factors can be discerned in the two sets of results, are summarised in the Scree plot in Figure 4 and the component matrix in Table 8.

[*Figure 4 about here*]

[*Table 8 about here*]

There appears to be only one component extracted with an Eigen value above 1 and it is concluded that AVST and XK-Lex are measuring the same construct.

It is assumed that if there is a significant correlation between AVST and XK-Lex scores, i.e., if the two tests are measuring the same knowledge, then the AVST should be able to distinguish between learners at different levels of general proficiency. Highly proficient test-takers (PhD students) will score significantly better on AVST than mid group (MA students) and the mid group would, in turn, score significantly higher than low group (BA students). The results for the performance of the three groups are summarised in Table 9.

[*Table 9 about here*]

An ANOVA confirms that the difference between the groups' means scores is statistically significant, $F(2, 117) = 23.79$, $p < .001$, and Dunnett's T3 post-hoc indicated that all pairwise comparisons were significant ($p < .05$). This suggests that the AVST is performing as general vocabulary size measures are able to do, in discriminating between learners with different levels of general lexical knowledge.

4. Discussion

4.1. The performance of the AVST

An academic vocabulary size test, which works well and for which there are normalised scores for attainment, ought to be pedagogically useful. This information can be used to set and monitor progress towards learning goals in a crucial area of language knowledge as is demonstrated with the information in Table 9. The results gained from the AVST used in this study suggest that our EFL learners recognise, on average, about 400 words from the AWL, our MA students about 450 words, our PhD students about 500 words, and our native speaker research students recognise all 570 words from the list. Given the distribution of AWL, shown in Table 1, complete knowledge of the AWL among native speakers, or something very like it, should be expected.

Native speakers appear to experience no difficulty in discriminating between AWL words and other English words of very low frequency which they are unlikely to know.

These scores are only useful if the test can be demonstrated to be reliable and valid. The results suggest the test is working well and that the scores they produce appear believable. The reliability measures from Cronbach's alpha scores and Rasch analysis, appear excellent. The alpha score for the odd and even calculation is higher than for the top-bottom halves, and this is the first suggestion that the learning of words from the AWL is related to these words frequency in general language corpora. The Rasch analyses also suggest that the test is working well both in terms of its representativeness and its technical quality.

As indicated in Figure 2, 114 items were almost sufficient to assess the test-takers academic vocabulary knowledge, since the items difficulty ranged from -1.30 to 5.72 logits. The results show no flooring effect but there appears to be a few test-takers (*N* = 6) maxing out the test that indicates some level of ceiling effect, which most likely caused by the presence of highly proficient non-native speakers of English pursing Masters or Doctoral programmes in an English medium university. However, because the test form includes a large number of items, the person ability estimates show good level of precision as indicated by the mean *SE* of .37 logits (*SD* = .12). There appears to be no significant gaps in the item hierarchy until the higher level. The technical quality of the test items was assessed by inspecting the items' fit to the Rasch model applying the Rasch Infit mean-square (Infit MNSQ) statistics. The results showed no overfit to the model, but indicated that two items slightly underfit the model (Infit MNSQ over 1.4). However, these two items represent only a 1.75% misfit rate (i.e., 2 out of 114 items); thus, the AVST items indicated very good fit to the Rasch model overall.

4.2. What is the contribution of AWL to academic success?

The regression analysis results, in Table 7, indicate that both academic word knowledge and overall vocabulary size contribute significantly to an explanation of variance in GPA scores. The AVST scores are the better predictor explaining about 53% of variance, and general vocabulary size explains about 47% of variance. When the two factors are built into a single model then the R^2 increases and AVST and XK-Lex combined can explain about 55% of the variance in GPA. These scores confirm the observation in Townsend et al. (2012) that vocabulary appears to be a major factor in explaining the variation in GPA scores that non-native speakers attain. The interesting feature in the data in this study is the way the AVST has better explanatory power than the overall vocabulary size measure. This may be a product of the kind of variation that might be expected from two tests of the same quality but which, in this case, use different sets of words, even when they test the same quality well and reliably. The AVST has some qualities, including a vastly superior sampling rate, which might account for this.

The explanatory power of both general vocabulary size and academic word knowledge appears greater in this study than in Townsend et al. (2012). However, this is not an exact replication of Townsend et al.'s study and it uses different test instruments applied to a different set of students. The construction of the AVST test in this study, which controls for frequency in the AWL, may account for some of the differences. This information is at the heart of the final element of the analysis in this study. Given the high frequency of most items in the AWL, and spread of frequencies among the AWL items, it is suggested that any test of the AWL must also function as a test of general vocabulary size and may actually be a test of general vocabulary size rather than a test of specialist vocabulary knowledge.

4.3. The AWL, frequency and collinearity with general vocabulary size scores

Our interpretation of the frequency profile (Figure 1 and Table 5) and the degree of variation seen in the Rasch model scores (Figure 2) is that there is a frequency effect in the students' responses to the AWL test. This is the sort of effect, and the sort of variation, that should be seen in any test of vocabulary which assesses vocabulary size and draws its items from a range of frequency bands. This interpretation is in contrast to that of Schmitt, Schmitt, and Clapham (2001) who regard such variation as problematic. In Schmitt et al. the test items drawn from the AWL are viewed as a coherent group of items which should therefore share the same performance characteristics: they are all relatively infrequent, they are not acquired from normal text but probably as a group of specialist academic words. In revising the Vocabulary Levels Test (VLT) they used this explanation to exclude items which performed, as they saw it, aberrantly. However, the frequency profile, depicted in Figure 1, indicates that the AWL is learned not as a coherent group, but words in the AWL are learned relatively sequentially in relation to the frequency of these words in texts. The Wright map (Figure 2) in this study clearly shows that items in the AWL will perform differently, and relatively predictably, in clear relation to their frequency.

Since the learning of the AWL and other general vocabulary items appear to progress so similarly in relation to frequency, this begs the question whether the AWL is a meaningfully different set of vocabulary which can make an important difference in academic performance. The high correlation between AVST and XK-Lex scores suggests these two tests may well be fulfilling the same function. The factor analysis (Figure 4 and Table 7) confirms that the two tests do not appear to be testing different qualities of knowledge but are testing the same thing. Our interpretation of this is that the quality being tested here is vocabulary size: the XK-Lex test contains no items from the AWL so it cannot directly test for this knowledge, but the AVST is testing across a range of word frequencies like the XK-Lex and so conceivably is a, slightly odd, test of vocabulary size. Like general vocabulary size, therefore, the test of AWL knowledge is able to discriminate between learners of different levels of ability. It is possible that this general size facet of the AVST is what lies at the heart of the way it is such a good predictor of GPA.

However, alongside the information from factor analysis should be set the size of the correlation between the two sets of scores and the question of collinearity. While the correlation between the two tests is high, it is not sufficiently high for us to interpret that the two tests are identical. Also, the values of the collinearity diagnostics confirm that the tests are not collinear. Our interpretation is that while the AVST contains a lot of general size information in the scores it produces it is, nonetheless, testing something qualitatively and quantitatively different from general vocabulary size. Because the explanatory power of the AVST is higher than general vocabulary scores in explaining GPA in this study, it appears that this is something additional to general vocabulary knowledge.

It must be emphasised that the AVST appears, at root, to be a test of vocabulary size. But the findings may indicate, too, that the AWL words do fulfil some important function in addition to having a vocabulary of the right general size, that allows the possessors of this vocabulary to gain better GPAs. Our best interpretation of the data, therefore, is to confirm Townsend et al.'s (2012) conclusion that knowledge of academic words contributes some unique, albeit marginal, variance to general academic attainment. A focus on the AWL in teaching, within this interpretation, appears a useful element of any EAP course provided it is done within the context of an overall program of vocabulary development for learners to attain the size of lexicon needed for real fluency.

5. Conclusions

The first conclusion to be drawn relates to the status the AWL has as a specialist list uniquely, and indispensably, useful to academic success. The AWL may not be as specialist, and as distinct from other vocabulary, as is assumed. It appears the AWL covers a range of frequencies in general corpora such as Kilgarriff (2006) and BNC/COCA but is heavily concentrated in the 3,000 most frequent words. They appear overwhelmingly to be words that are generally accessible to learners from general language and are likely to be known, in some form, by learners with fairly modest, intermediate level knowledge of English. The 10% contribution to coverage in academic texts claimed for the AWL can at least be questioned and may be exaggerated by the choice of the GSL for assumed knowledge of general and frequent words. It has proved very difficult in the work reported in this paper to separate out knowledge of the AWL from knowledge of English vocabulary generally. From this standpoint, the AWL looks like a selection of general words which are learned in relation to frequency just like other words appear to be. Having suggested this interpretation, however, it must be acknowledged that the validity of the AWL is dependent on further empirical study. We need to determine exactly how many and what percentage of the AWL words have specialised meanings relevant to academic study, and how many and what percentage of the high-frequency words in the AWL do not and might, perhaps, be excluded from the list. Ha and Hyland (2017) carry out such an analysis in relation to a specialist Finance word list.

This is not to conclude, however, that the AWL is entirely redundant and this study, like Townsend et al.’s (2012), suggests there is some marginal advantage to be had in learning it.

Our best interpretation of the data from this study is that general vocabulary size can explain a very sizeable proportion of the variance in GPA scores but the explanatory power of vocabulary knowledge can be improved slightly when knowledge of the AWL is also factored in. Knowledge of the AWL by itself is not a short-cut or an easy route to academic success, no serious teacher or academic will have thought this anyway, but it does appear useful if built into the far lengthier and more difficult process of mastering the scale of lexicon needed to handle academic discourse easily.

References

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Biemiller, A. (2001). Teaching vocabulary: Early, direct, and sequential. *The American Educator, 25*(1), 24-28.

Borsboom, D., Mellenbaugh, G.J., & van Heerden, J. (2004). The concept of a validity. *Psychological Review, 111*, 1061–1071.

Brown T. A (2006). Confirmatory factor analysis for applied research. New York: Guilford Press.

Chen, Q., & Ge, C. (2007). A corpus-based lexical study on frequency and distribution of Coxhead’s AWL word families in medical research articles. *English for Specific Purposes*, *26*, 502–514.

Chung, M., & Nation, P. (2003). Technical vocabulary in specialized texts. *Reading in a Foreign Language*, *15*, 103–116.

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

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

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes, 33*, 66-76.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, California: Sage Publications, Inc.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics,* *35*(3), 305-327.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics, 11*(4), 341-363.

Gyllstad, H., Vilkaite, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL International Journal of Applied Linguistics, 166*, 276–303.

Ha, A. Y. H., & Hyland, K. (2017). What is technicality? A Technicality Analysis Model for EAP vocabulary. *Journal of English for Academic Purposes, 28,* 35-49.

Hart, B., & Risley, T. (2003). The early catastrophe: The 30 million word gap. *The American Educator, 27,* 4-9.

Horst, M. & Cobb, T. (2006). Editorial: Second language acquisition. *The Canadian Modern Language Review*, *63*(1), 1-12.

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

Kilgarriff, A. (2006). BNC database and word frequency lists. Retrieved May, 2016, from http://www.kilgarriff.co.uk/bnc-readme.html#lemmatised.

Kremmel, B., & Schmitt, N. (forthcoming). Vocabulary levels test. In Liontas, J. I., DelliCarpini, M., & Riopel, J. C. (Eds.), *The TESOL Encyclopedia of English Language Teaching*. John Wiley & Sons, Inc.

Laufer, B. & Ravenhorst-Kalovsky, G. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15-30.

Linacre, J.M. (2011). *A user's guide to WINSTEPS Ministeps*; Rasch-model Computer Program. Program Manual 3.71.0.

Authors. (2012). The vocabulary knowledge of University students in Saudi Arabia. *TESOL Arabia Perspectives, 19*(3), 13-20.

Authors. (2017). Recognition Vocabulary Knowledge and Intelligence as Predictors of Academic Achievement in EFL Context. *TESOL International Journal, 12*(1), 128-142.

Authors. (2003). *The Swansea levels test*. Newbury: Express.

Author. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Author. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist and B. Laufer (eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 57-78).

EUROSLA monograph *2*.

Authors. (2009). Vocabulary size and the Common European Framework of Reference for Languages. In B. J. Richards, H. M. Daller, D. Malvern, P. Meara, J. Milton & J. Treffers-Daller (Eds.), *Vocabulary studies in first and second language acquisition: The interface between theory and application* (pp. 194-211). Basingstoke: Palgrave.

Authors. (2013). Vocabulary size revisited: the link between vocabulary size and academic achievement. *Applied Linguistics Review, 4*(1), 151–172.

Messick, S. (1989). Validity. In: R.L. Linn (Ed.), *Educational measurement*. 3rd edition (pp. 13–103). New York: Macmillan.

Mudraya, O. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes, 25*, 235–256.

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

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 3–14). Amsterdam: John Benjamins.

Nation, I.S.P. (2012). The BNC/COCA word family lists (17 September 2012). Unpublished paper. Available at: http://www.victoria.ac.nz/lals/about/staff/paul-nation (June 2016).

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedogogiske Institute.

Roche, T., & Harrington, M. (2013). Recognition vocabulary knowledge as a predictor of academic performance in an English as a foreign language setting. *Language Testing in Asia, 3*(1), 1-13.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484-503.

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.

Snow, C., Lawrence, J., & White, C. (2009). Generating knowledge of academic language among urban middle school students. *Journal of Research on Educational Effectiveness, 2*, 325–344.

Stæhr, L.S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal, 36* (2), 139-152.

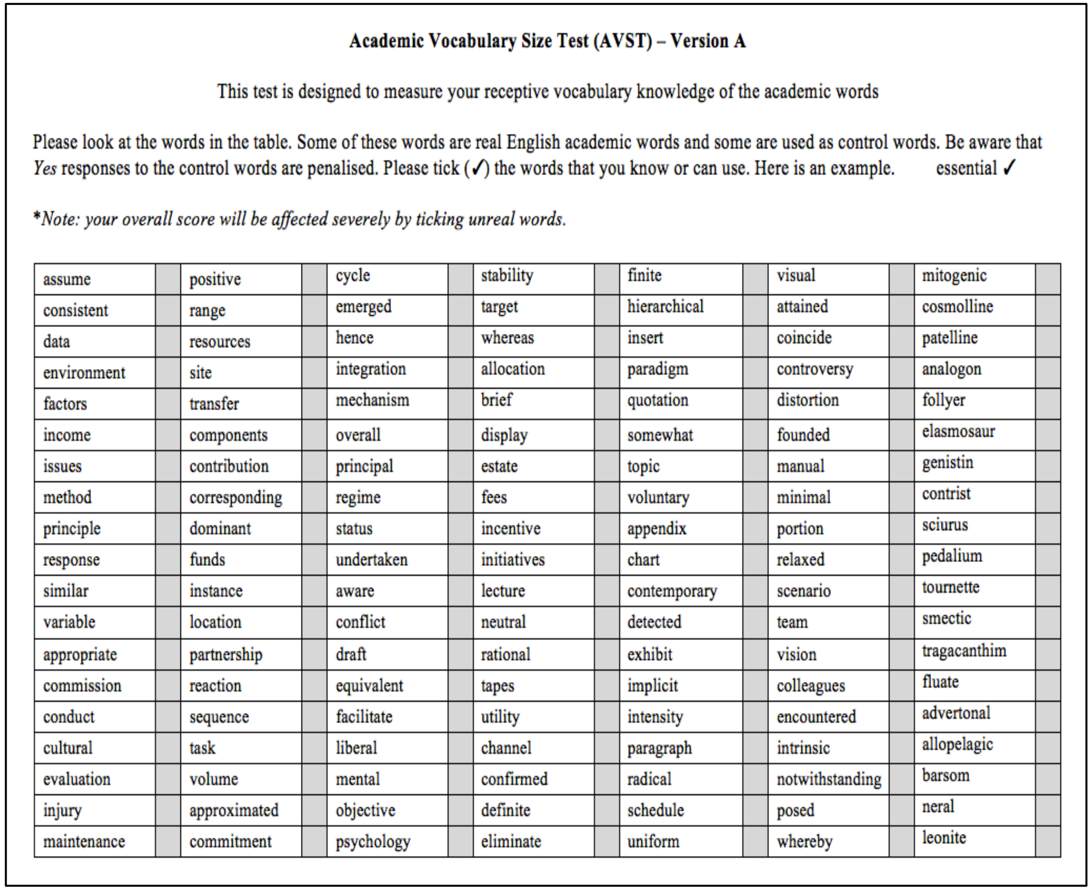
Thorndike, E.L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Bureau of Publications, Teachers College, Columbia University.

Townsend, D., Filippini, A., Collins, P., & Biancarosa, G. (2012). Evidence for the importance of academic word knowledge for the academic achievement of diverse middle school students. *The Elementary School Journal*, *112*(3), 497-518.

Ward, J. W. (1999). How large a vocabulary do EAP engineering students need? *Reading in a Foreign Language*, 12(2), 309–324.

West, M. (1953). *A general service list of English words*. London: Longman.

Appendix A. AVST version A



Academic Vocabulary Size Test (AVST) – Version A

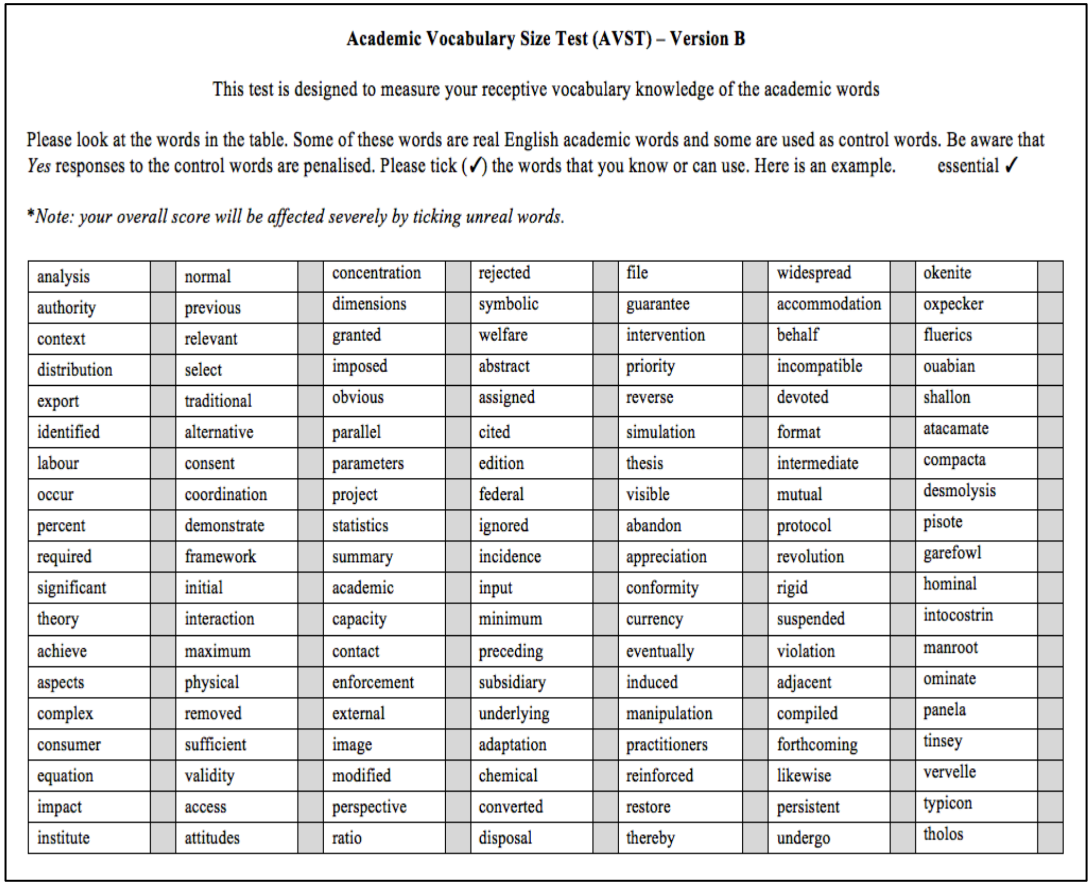
This test is designed to measure your receptive vocabulary knowledge of the academic words

Please look at the words in the table. Some of these words are real English academic words and some are used as control words. Be aware that *Yes* responses to the control words are penalised. Please tick (✓) the words that you know or can use. Here is an example.   essential ✓

**Note: your overall score will be affected severely by ticking unreal words.*

assume		positive		cycle		stability		finite		visual		mitogenic	
consistent		range		emerged		target		hierarchical		attained		cosmolline	
data		resources		hence		whereas		insert		coincide		patelline	
environment		site		integration		allocation		paradigm		controversy		analogon	
factors		transfer		mechanism		brief		quotation		distortion		follyer	
income		components		overall		display		somewhat		founded		elasmosaur	
issues		contribution		principal		estate		topic		manual		genistin	
method		corresponding		regime		fees		voluntary		minimal		contrist	
principle		dominant		status		incentive		appendix		portion		sciurus	
response		funds		undertaken		initiatives		chart		relaxed		pedalium	
similar		instance		aware		lecture		contemporary		scenario		tournette	
variable		location		conflict		neutral		detected		team		smectic	
appropriate		partnership		draft		rational		exhibit		vision		tragacanthim	
commission		reaction		equivalent		tapes		implicit		colleagues		fluate	
conduct		sequence		facilitate		utility		intensity		encountered		advertonal	
cultural		task		liberal		channel		paragraph		intrinsic		allopelagic	
evaluation		volume		mental		confirmed		radical		notwithstanding		barsom	
injury		approximated		objective		definite		schedule		posed		neral	
maintenance		commitment		psychology		eliminate		uniform		whereby		leonite	

Appendix B. AVST version B



Academic Vocabulary Size Test (AVST) – Version B

This test is designed to measure your receptive vocabulary knowledge of the academic words

Please look at the words in the table. Some of these words are real English academic words and some are used as control words. Be aware that *Yes* responses to the control words are penalised. Please tick (✓) the words that you know or can use. Here is an example. essential ✓

**Note: your overall score will be affected severely by ticking unreal words.*

analysis		normal		concentration		rejected		file		widespread		okenite	
authority		previous		dimensions		symbolic		guarantee		accommodation		oxpecker	
context		relevant		granted		welfare		intervention		behalf		fluerics	
distribution		select		imposed		abstract		priority		incompatible		ouabian	
export		traditional		obvious		assigned		reverse		devoted		shallon	
identified		alternative		parallel		cited		simulation		format		atacamate	
labour		consent		parameters		edition		thesis		intermediate		compacta	
occur		coordination		project		federal		visible		mutual		desmolysis	
percent		demonstrate		statistics		ignored		abandon		protocol		pisote	
required		framework		summary		incidence		appreciation		revolution		garefowl	
significant		initial		academic		input		conformity		rigid		hominal	
theory		interaction		capacity		minimum		currency		suspended		intocostrin	
achieve		maximum		contact		preceding		eventually		violation		manroot	
aspects		physical		enforcement		subsidiary		induced		adjacent		ominate	
complex		removed		external		underlying		manipulation		compiled		panela	
consumer		sufficient		image		adaptation		practitioners		forthcoming		tinsey	
equation		validity		modified		chemical		reinforced		likewise		vervelle	
impact		access		perspective		converted		restore		persistent		typicon	
institute		attitudes		ratio		disposal		thereby		undergo		tholos	

List of tables

Table 1. Frequency distribution of words in the AWL (Authors, forthcoming)

Frequency band	BNC (Kilgarriff, 2006)	BNC/COCA (Nation, 2012)
1k	94	19
2k	151	136
3k	124	329
4k	81	62
5k	51	15
6k	35	6
7k	9	1
Off list	25	2

Table 2. Results from 16 native speakers using AVST

	N	Hits	FA	Mean AWL size	Max. possible score
Form A	16	114	0	570	570
Form B	16	114	0	570	570

Table 3. Paired sample statistics of the AVST scores in parallel forms and split-halves

		Max. possible score	Mean	*N*	*SD*	*SE*
Pair 1	Form A	570	440.00	120	79.01	7.21
	Form B	570	440.54	120	88.01	8.03
Pair 2	A Odds	285	220.75	120	41.68	3.80
	A Evens	285	219.29	120	40.19	3.66
Pair 3	A Top-half	285	221.16	120	38.28	3.49
	A Bottom-half	285	219.83	120	44.50	4.06

Table 4. Cronbach’s alpha reliability indices for the parallel forms and split-halves

Pair	*Cronbach’s Alpha*
A-B	.97
Odds-Evens	.92
Top-Bottom	.91

Table 5. Mean ranks for hierarchy bands

Frequency bands	Mean Ranks
Band 1	5.45
Band 2	5.00
Band 3	4.08
Band 4	3.04
Band 5	2.34
Band 6	1.10

Table 6. Factor loadings

Variable	Level 4	Level 5	Level 3	Level 2	Level 1	Level 6
Factor loading	.95	.92	.90	.88	.82	82

Note: Level = frequency level.

Table 7. Regression model summary (individual and combined predictors)

Model	R	R Square	Adjusted R Square	SE of Estimate
1 AVST	.73	.53	.52	.54
2 XK-Lex	.68	.47	.46	.58
3 AVST and XK-Lex (combined)	.75	.55	.55	.53

Table 8. Component matrix[a] from AVST and XK-Lex data

	Component
	1
XK-Lex Score	.944
AVST Score	.944

Note. Extraction Method: Principal Component Analysis; a = 1 components extracted.

Table 9. Descriptive statistics for the performance of three groups of students in AVST

Group	Mean	*N*	*SD*	Min	Max	*SE*
Low	396.00	55	78.43	210	505	10.57
Mid	462.22	36	65.01	340	560	10.83
High	495.86	29	41.36	395	540	7.68

List of figures

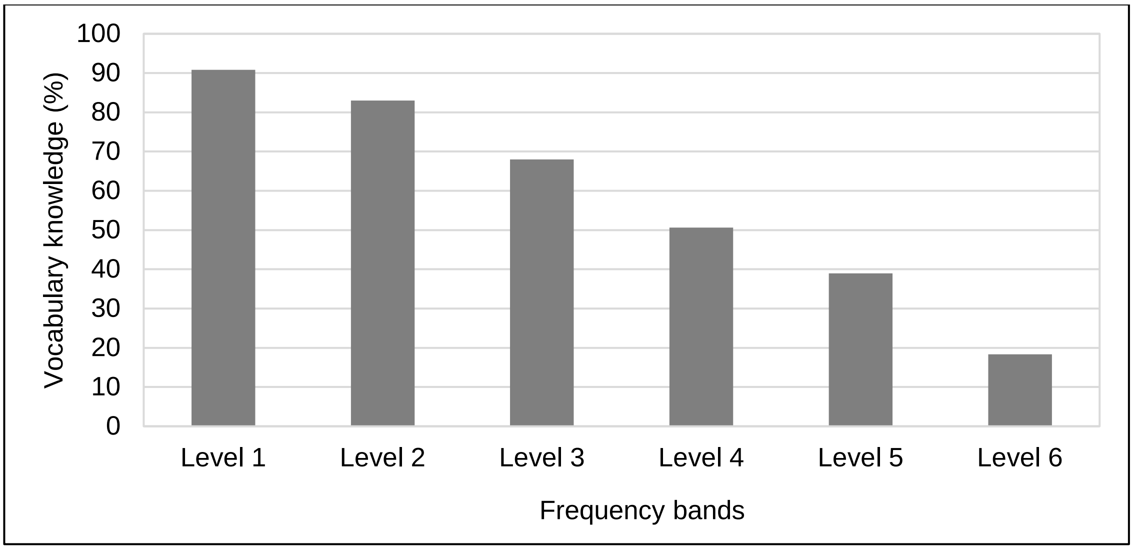
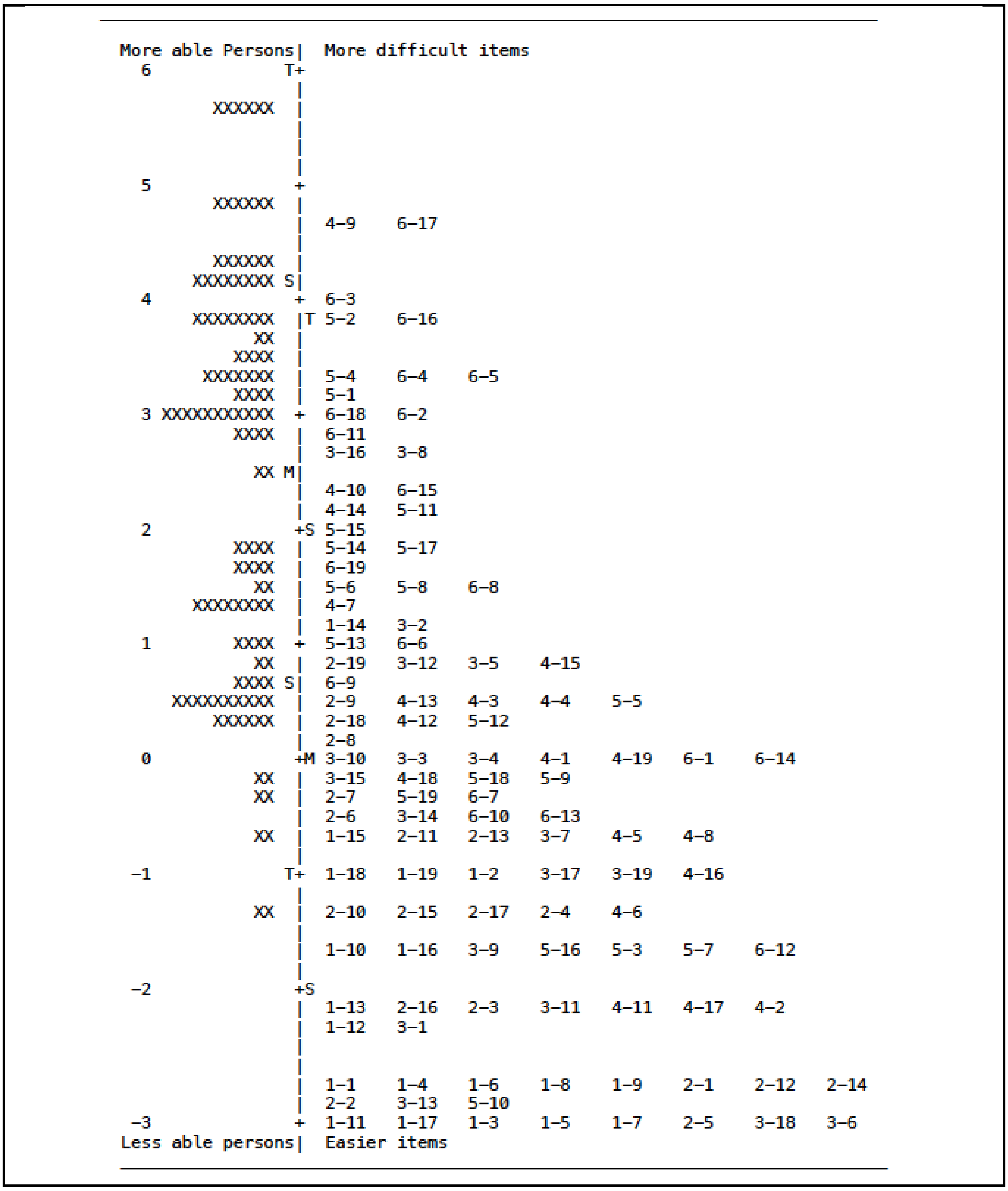


Figure 1. Learners' academic words frequency profile



```
More able Persons|  More difficult items
  6             T+
                 |
         XXXXXX  |
                 |
                 |
                 |
  5              +
         XXXXXX  |
                 |  4-9    6-17
                 |
         XXXXXX  |
       XXXXXXXX S|
  4              +  6-3
       XXXXXXXX  |T 5-2    6-16
             XX  |
           XXXX  |
        XXXXXXX  |  5-4    6-4    6-5
           XXXX  |  5-1
  3 XXXXXXXXXXX  +  6-18   6-2
           XXXX  |  6-11
                 |  3-16   3-8
             XX M|
                 |  4-10   6-15
                 |  4-14   5-11
  2              +S 5-15
           XXXX  |  5-14   5-17
           XXXX  |  6-19
             XX  |  5-6    5-8    6-8
       XXXXXXXX  |  4-7
                 |  1-14   3-2
  1        XXXX  +  5-13   6-6
             XX  |  2-19   3-12   3-5    4-15
           XXXX S|  6-9
     XXXXXXXXXX  |  2-9    4-13   4-3    4-4    5-5
         XXXXXX  |  2-18   4-12   5-12
                 |  2-8
  0              +M 3-10   3-3    3-4    4-1    4-19   6-1    6-14
             XX  |  3-15   4-18   5-18   5-9
             XX  |  2-7    5-19   6-7
                 |  2-6    3-14   6-10   6-13
             XX  |  1-15   2-11   2-13   3-7    4-5    4-8
                 |
 -1             T+  1-18   1-19   1-2    3-17   3-19   4-16
                 |
             XX  |  2-10   2-15   2-17   2-4    4-6
                 |
                 |  1-10   1-16   3-9    5-16   5-3    5-7    6-12
                 |
 -2              +S
                 |  1-13   2-16   2-3    3-11   4-11   4-17   4-2
                 |  1-12   3-1
                 |
                 |
                 |  1-1    1-4    1-6    1-8    1-9    2-1    2-12   2-14
                 |  2-2    3-13   5-10
 -3              +  1-11   1-17   1-3    1-5    1-7    2-5    3-18   3-6
Less able persons|  Easier items
```

Figure 2. Wright map of person measures and item calibrations.

Note. Each ‘X’ is 1 person; M = Mean; S = 1 *SD*; T = 2 *SD*.

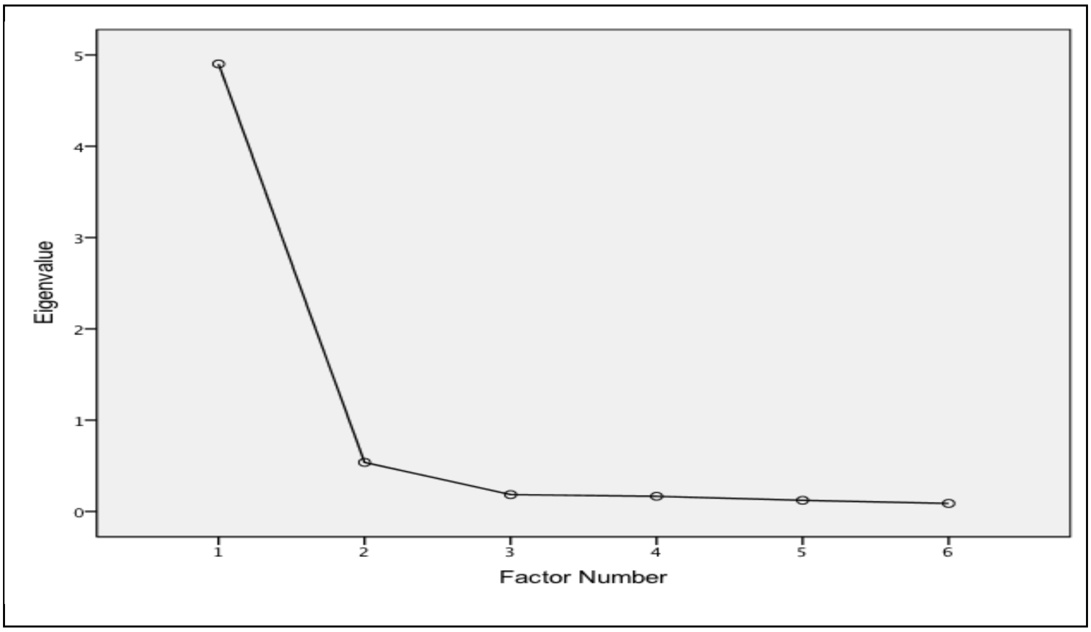


Figure 3. Scree plot from AVST data

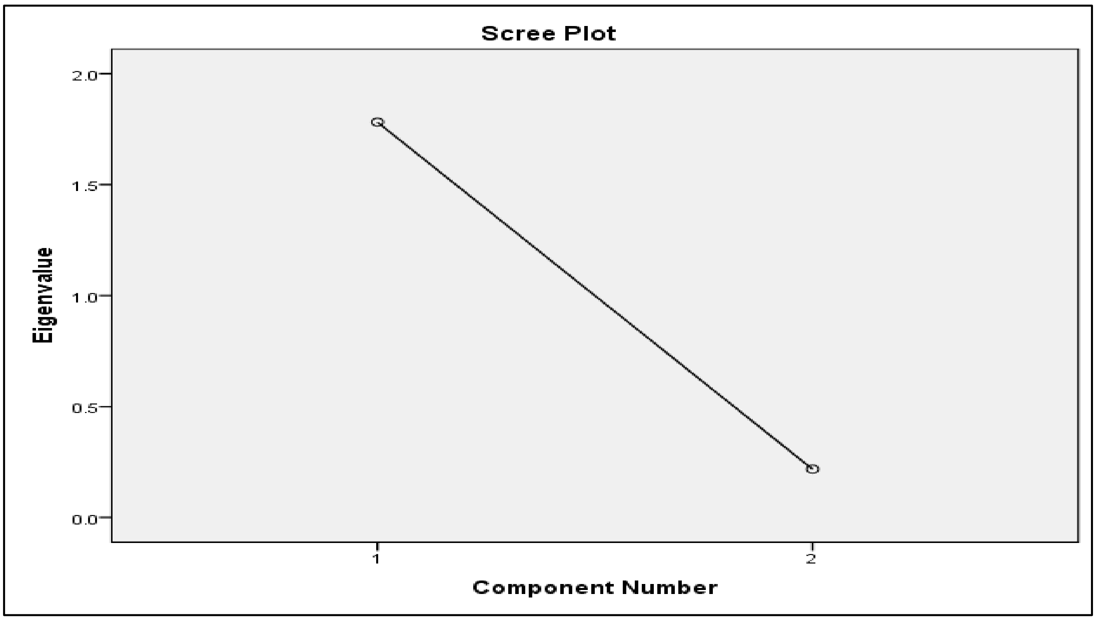


Figure 4. Scree plot from AVST and XK-Lex data