Linking scores from two written receptive English academic vocabulary tests—The VLT-Ac and the AVT

Article

LANGUAGE TESTING

Language Testing
1–28


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/02655322221145643
journals.sagepub.com/home/ltj

Linking scores from two written receptive English academic vocabulary tests—The VLT-Ac and the AVT

Marcus Warnby
Stockholm University, Sweden

Hans Malmström
Chalmers University of Technology, Sweden

Kajsa Yang Hansen
University of Gothenburg, Sweden

Abstract

The academic section of the Vocabulary Levels Test (VLT-Ac) and the Academic Vocabulary Test (AVT) both assess meaning-recognition knowledge of written receptive academic vocabulary, deemed central for engagement in academic activities. Depending on the purpose and context of the testing, either of the tests can be appropriate, but for research and pedagogical purposes, it is important to be able to compare scores achieved on the two tests between administrations and within similar contexts. Based on a sample of 385 upper secondary school students in university-preparatory programs (independent CEFR B2-level users of English), this study presents a comparison model by linking the VLT-Ac and the AVT using concurrent calibration procedures in Item Response Theory. The key outcome of the study is a score comparison table providing a means for approximate score comparisons. Additionally, the study showcases a viable and valid method of comparing vocabulary scores from an older test with those from a newer one.

Keywords

Academic literacy, concurrent calibration, item response theory, linking, receptive academic vocabulary, testing, upper secondary school EFL learners

Corresponding author:
Marcus Warnby, Department of Teaching and Learning, Stockholm University, Frescativägen, 114 19 Stockholm, Sweden.
Email: marcus.warnby@su.se

Introduction

The present study presents a means for comparing scores from two tests of academic vocabulary knowledge, the academic section of the Vocabulary Levels Test (Schmitt et al., 2001) and the recently published Academic Vocabulary Test (Pecorari et al., 2019). The study was originally prompted by research in Sweden with English-as-a-foreign-language (EFL) upper secondary school students enrolled in university-preparatory programs. Students in this context must develop a vocabulary that is appropriate for tertiary education, specifically to enable them to engage with a substantial amount of English reading at university (Malmström & Pecorari, 2022; Pecorari et al., 2011). To this end, we needed (*i*) to broadly measure students' academic vocabulary knowledge and (*ii*) to make comparisons of academic vocabulary knowledge over time within and between similar pre-academic contexts.

There is wide agreement in the literature that student engagement in English language activities at university requires knowledge of frequent general vocabulary as well as words with a particular affinity to the academic and disciplinary context (e.g., Charles & Pecorari, 2015; Coxhead, 2016; Hyland & Tse, 2007; Vilkaitė-Lozdienė & Schmitt, 2019). Knowledge of academic vocabulary, that is, "vocabulary that occurs across a range of academic subject areas" (Coxhead, 2016, p. 177), has been identified as particularly important and an "indispensable component of academic reading abilities" (Gardner & Davies, 2014, p. 305). In certain academic or "pre-academic" contexts, and for various pedagogical purposes, it is important to be able to estimate students' knowledge of academic words, using means of measurement that are appropriate to the context and to the purpose(s)).

In 2010, Lin and Morrison (2010, p. 257) asserted that "there is no one commonly accepted standard test of academic vocabulary". While there is still no recognized "standard" test available, the Vocabulary Levels Test (VLT; Nation, 1983)—specifically the most recent version of it with its updated academic section (VLT-Ac; Schmitt et al., 2001)—has become widespread (despite some criticism brought against the test as a whole, e.g., Schmitt et al., 2020; Stoeckel et al., 2021; Webb et al., 2017). Using a matching format, the VLT-Ac tests knowledge at the meaning-recognition level of 30 words sampled from Coxhead's Academic Word List (AWL; Coxhead, 2000) and is "designed to give an estimate of vocabulary size for second language (L2) learners of general or academic English [and] to inform decisions concerning whether an examinee is likely to have the lexical resources necessary to cope with certain language tasks, such as reading authentic materials" (Schmitt et al., 2001, pp. 55–56). On the face of it, therefore, the VLT-Ac is a suitable test to administer to upper-secondary students preparing for university. However, in our context of Nordic upper-secondary schooling, ceiling effects have recently been observed when the VLT-Ac was administered in testing (e.g., Edgarsson, 2018; Skjelde & Coxhead, 2020), calling into question the utility of the VLT-Ac for the purpose of capturing students' academic vocabulary knowledge on its full scale and affecting its predictions about other related variables. Clearly, the VLT-Ac is problematic for measuring academic vocabulary knowledge in the context of independent EFL users with high exposure to English.

An alternative instrument to the VLT-Ac was recently published: the Academic Vocabulary Test (AVT; Pecorari et al., 2019). The AVT includes 57 items that measure knowledge of words sampled from the Academic Vocabulary List (AVL; Gardner & Davies, 2014), and can serve "as a tool for working pedagogically [with. . .and] highlight more things we need to learn and think about with respect to academic vocabulary" (Pecorari et al., 2019, p. 69). It uses the same meaning-recognition matching format as the VLT. No ceiling effects have been observed when the test has been administered to high-exposure-to-English Nordic EFL learners (e.g., Pecorari et al., 2019; Warnby, 2022), which suggests it could be a suitable test of receptive academic vocabulary knowledge in this context. However, no guidance for interpreting AVT scores or relating them to other variables was provided by the developers (Pecorari et al., 2019). It is, therefore, essential to place the AVT in the context of other more established tests, such as the VLT-Ac, to facilitate our understanding of the measurement property of the test.

Currently, however, no existing criteria support the comparison between AVT-scores and VLT-Ac scores on a group level. The main purpose of the present study is, therefore, to present a comparison model by linking the AVT and the VLT-Ac using concurrent calibration procedures in Item Response Theory (IRT) (Feuer et al., 1999; Kolen & Brennan, 2014).

The following research question guides the study:

Research Question: How does a score on the AVT compare to a score on the VLT-Ac?

This research opens up possibilities for scholars and teachers to use AVT scores or VLT-Ac scores—depending on the context and purpose—and make approximate comparisons, for example, within and between populations in similar contexts over time, in order to discern trends or patterns in written receptive academic vocabulary knowledge. Moreover, as new vocabulary tests are introduced and then used, the methods adopted in this study may be of value as an example of linking scores between old and new tests.

Background

This section starts by briefly setting the present study within the appropriate linguistic and educational context; this is done so that readers appreciate the need in this context to engage in the assessment of upper secondary school students' academic vocabulary knowledge. Subsequently, academic vocabulary is operationalized and the two tests of academic vocabulary knowledge—the VLT-Ac and the AVT—are set against each other, and we problematize the kinds of knowledge claims that can be made when using such tests. Finally, a brief foundation is provided for the process used to compare scores on the AVT and the VLT-Ac, so-called "linking."

Context of the study

Sweden, along with all the other Nordic countries, occupies a place among the top ten countries on EF's global ranking of English proficiency (Education First, 2020). The ability to read and understand spoken English among the Swedish population is on average very high according to EF. Swedish adolescents, too, are recognized for their general English skills (Bonnet, 2004; European Commission, 2012). There are many possible explanations for this situation. The fact that Swedish and English are both Germanic languages is often referenced as is the prominent status of English in Swedish society and the increasingly vast amount of extramural English exposure that young Swedish people experience (National Agency for Education [NAE], 2012; Sundqvist, 2009; Swedish Media Council, 2019). Even if the Nordic countries are considered part of the expanding circle, where English is usually awarded a "foreign language" status, the omnipresence of English in Sweden, almost regardless of social setting, has led scholars to argue that English in Sweden holds an L2 (second language) position (e.g., Hyltenstam, 2004; Sundqvist, 2009).

English is also used to a high degree in Swedish higher education. Figures from a recent report (Malmström & Pecorari, 2022) indicated that, at an advanced level of study, 64% of all programs use English as the medium of instruction. While English is adopted as a teaching language to a lesser degree at the undergraduate level, as much as 50% to 80% of the obligatory reading that students do in Swedish-medium courses at the undergraduate level constitutes reading in English (interestingly, in 25% of the Swedish-medium university courses *all* the required reading is in English).

English has a formal presence also at lower levels of education in an evolving educational landscape (a new national curriculum and grading system was implemented in 2011, with direct implications for how English is taught in schools). When students in Sweden complete their upper secondary education, the majority of students have taken English courses for 10 years. According to the National Agency for Education (NAE, 2022), two thirds of the upper secondary school student population are enrolled in one of the university-preparatory study programs (Arts, Economy, Humanities, Natural sciences, Social sciences, or Technology). Most of these students complete their upper secondary education with a passing grade from the last compulsory English course considered equivalent to a B2-level in the *Common European Framework of Reference* (CEFR) (Council of Europe, 2022; NAE, 2021). With a B2 level, they can be seen as "independent" users of English who meet the English basic requirement for university education eligibility not only in Sweden but also in many other countries.

Despite such expectations regarding the independent use of English among upper secondary school graduates, concerns have been raised that Swedish students are ill-prepared for university study. Specifically in view of the significant amount of reading in English expected by Swedish students the moment they start university, their ability to engage with English reading has been called into question (Pecorari et al., 2011). The causes of the English reading difficulties experienced by Swedish students have not been subject to much research, but it is unlikely to be due to a lack of general English vocabulary; adolescents in Sweden do well when tested on their general English vocabulary

knowledge at the high-frequency level (2000 frequency band) (e.g., Gyllstad, 2007; Lemmouh & Snoder, 2019; Sundqvist, 2009).

It is possible, therefore, that part of the English reading challenge can be explained by students having lexical gaps beyond high-frequency vocabulary, for example, in terms of their English academic or disciplinary vocabulary (e.g., Coxhead, 2016). For this reason, it is worthwhile testing such vocabulary knowledge among upper secondary school students, not least to enable possible preventive pedagogical measures to be taken by upper secondary schools before students enroll at the tertiary level. Students studying in one of the university-preparatory programs have typically not decided on an academic discipline (the programs offer broad eligibility to university education and students can choose whatever discipline they want, regardless of the upper secondary school program). Consequently, it makes sense to test their written receptive *academic* vocabulary knowledge rather than their disciplinary vocabulary knowledge.

Operationalizing academic vocabulary

The academic vocabulary of the kind we want to assess in the Swedish upper secondary education context is usually operationalized with reference to lists of words identified as "academic" (see Therova, 2020, for an overview of such lists and the identifying principles, and pitfalls, when creating the lists). To date, two such lists of academic words have received particularly widespread attention: Coxhead's (2000) AWL and Gardner and Davies' (2014) AVL. The criteria adopted for creating the more recent AVL—frequency, ratio; range; dispersion; discipline—recall the fundamental principles used in the design of the AWL, although there are differences in how the criteria were applied (see Gardner & Davies, 2014, for details).

The AVL differs from the AWL in three main respects. First, the words for the AVL were drawn from a significantly larger corpus. Second, as a result of the extraction criteria, the AVL has a larger frequency level range than the AWL. For example, Coxhead (2000) excluded all high-frequency words that also occurred among the 2000 most common words in the General Service List (GSL; West, 1953), whereas Gardner and Davies (2014) did not use such a stop criterion. Instead, they allowed high-frequency words to be included, provided they were significantly more frequent in an academic corpus than in a nonacademic corpus. The AWL and the AVL have a substantial and comparable share of words in the 3000- to 9000-frequency range; depending on where one sets the limit for high-frequency words, this category of words may be labeled "mid-frequency" (Nation, 2013; Schmitt & Schmitt, 2014). Moreover, the AVL, in relation to the AWL, contains more words that could be placed in the lower frequency bands (see Appendix 1). Thus, it can be argued that both the AWL and the AVL present academic core words that are also found in general English at high-, mid-, and low-frequency levels (this does not change the fact that they are more frequent in academic texts compared to non-academic texts). Third, while Coxhead used the more inclusive conception of a word family (base form of the word plus inflected and derived forms) as a basis for the AWL, Gardner and Davies designed the AVL based on word lemmas (base form of the word plus inflected forms of the same part of speech). In the context of vocabulary testing, several scholars question the appropriateness of word families as a word counting unit and, instead,

advocate the use of /f/lemmas (e.g., Brown et al., 2020; McLean, 2018; Schmitt, 2010) since it gives a "clearer idea of what a correct answer on an item does and does not mean" (Kremmel, 2016, p. 979). The lemmatized AVL could be considered a list of flemmas, that is, identical forms of different parts of speech, for example, the AVL noun lemma *project*, also a verb lemma, could better be grouped as a flemma. Such arguments point toward a change of measurement instruments from word-family-based tests to lemma-based tests. However, these much-debated issues are beyond the scope of the current study which aims to calibrate scores on the AVT and the VLT-Ac. Despite these differences, it can be argued that both the AWL and the AVL largely represent a vocabulary common across academic disciplines (words from both lists occur frequently in academic texts, Coxhead, 2000; Gardner & Davies, 2014) and can, thus, be regarded as constituting one domain—written English academic vocabulary.

Testing written receptive meaning-recognition English academic vocabulary

The complexity involved in "knowing" a word (and therefore in delineating the domain "vocabulary knowledge") is widely acknowledged. The present study makes use of two existing tests—the VLT-Ac and the AVT—targeting knowledge of one main domain, written receptive English academic lexis at the level of meaning-recognition. This means that the kind of academic word knowledge tested involves a connection of form and meaning ("*What meaning does this word form signal?*") and associations of words ("*What other words does this make us think of?*") (Nation, 2019, p. 16).

The VLT—first developed by Nation (1983)—may well be the most utilized measure of English second/foreign language learners' written receptive meaning-recognition vocabulary knowledge (Read, 2000; Schmitt et al., 2020). The VLT contains four frequency-based general vocabulary sections and one academic section. In 2001, the original VLT was revised, updated, and validated by Schmitt et al. (2001; refer to their paper for details). The academic section of the revised VLT (VLT-Ac) made use of the, then, recently presented AWL (Coxhead, 2000). The VLT-Ac targets 30 words from the AWL grouped in 10 clusters. The test taker is asked to match a definition with a single-word unit from a list of six alternatives (each including three target words and three distractors).[1] While the VLT is "still a well-used standard vocabulary measurement, the authors have not revised it since it was launched over 17 years ago," confirming that "most tests, once launched, are not revised in any systematic way" (Schmitt et al., 2020, p. 110, but see Webb et al., 2017 for an "updated" VLT, but notably a levels test without an academic section).

Two recent studies in our context of Nordic upper secondary education have used the VLT-Ac: Edgarsson (2018) and Skjelde and Coxhead (2020). Both studies observed ceiling effects in the VLT-Ac scores. In his Icelandic sample, Edgarsson (2018) correlated the VLT-Ac scores with scores on an academic reading task, and the ceiling effects meant that information was lost in the correlation at the higher level of VLT-Ac. Similar problems were experienced by Skjelde and Coxhead (2020) who used Norwegian upper secondary students' English grades and their negatively skewed VLT-Ac scores in

correlational and regression analyses. Testing relatively similar participants, Busby (2020) correlated Norwegian (undergraduate 1st year) students' VLT-Ac scores with extramural English factors, but did not find any significant relationship and, due to ceiling effects in the VLT-Ac scores, concluded that the effect would possibly have been seen "with a test based on the Academic Vocabulary List (Gardner & Davies, 2014) which uses lemmas instead of word families, and provides greater coverage of academic texts" (Busby, 2020, p. 76). There is ample evidence, thus, that using the VLT-Ac in this Nordic context of later upper secondary/early university education is problematic.

The AVT (Pecorari et al., 2019) tests knowledge of academic words sampled from the AVL (Gardner & Davies, 2014) and the test design and validation recall the principles and procedures used by Schmitt et al. (2001) (see Pecorari et al., 2019 for details). The resulting matching test format is virtually the same as that used in the VLT-Ac. A test taker's knowledge of 57 target words—at the level of meaning-recognition—is tested across 19 clusters.[2] In their validation of the AVT, Pecorari et al. (2019) did not provide any means of interpreting a test score or relating it to other measures. In that respect, the AVT suffers from the same shortcomings as many other tests with few indications of how to use the scores (Schmitt et al., 2020). In our Swedish context, AVT scores have been normally distributed among university students (Pecorari et al., 2019) and upper secondary school students (Warnby, 2022). Furthermore, as opposed to Busby (2020), the participants' scores in Warnby (2022) correlated positively and significantly with, for example, extramural English factors. This suggests that the AVT may be a more appropriate instrument in this context.

Scores from written receptive meaning-recognition tests—such as the VLT-Ac and the AVT—are often considered to bear on reading ability and, therefore, often correlated with reading scores (e.g., Edgarsson, 2018; Laufer & Aviad–Levitzky, 2017; Shaw & McMillion, 2011). However, the meaning-recognition format suffers from two important factors affecting its similarity to reading. First, the matching format in meaning-recognition tests like the VLT-Ac and the AVT may suffer from local item dependence (e.g., Ha, 2021; Kamimoto, 2014) and is prone to guessing or construct-irrelevant test strategies that may lead to an overestimation of word knowledge for reading (e.g., Gyllstad et al., 2015; Kamimoto, 2008; Stewart & White, 2011). Second, despite the significant predictability meaning-recognition has for reading ability, its construct validity has been debated lately in comparison to meaning-recall formats (Laufer & Aviad–Levitzky, 2017; McLean, 2021) mainly for two reasons. First, empirical evidence from English L2 research indicates that meaning-recall is better at predicting reading comprehension than meaning-recognition (McLean et al., 2020; Stewart et al., 2021; Zhang & Zhang, 2020). Second, meaning-recall formats may be more similar to real-life reading demands since the meaning of words has to be evoked quickly in the mental lexicon of the test taker/reader during fluent reading, compared to meaning-recognition formats that demand a lower level of word knowledge indicating partial knowledge of the words tested (e.g., Aviad-Levitzky et al., 2019; Kremmel & Schmitt, 2016; McLean et al., 2020; Nation & Webb, 2011; Stoeckel et al., 2021). Since a test answer may be attributable to partial lexical knowledge of the tested words (Nagy et al., 1985), distractors in a meaning-recognition test should be

written to provide possibilities for learners to gain credit for partial knowledge (Nation, 2012). In their revision of the VLT, Schmitt et al. (2001) adopted the principle of partial lexical knowledge, which means that the alternative words for each cluster were chosen so that their meaning differed clearly. The argument was that learners with even a minimal understanding of a word's meaning should be able to choose the correct word. This partial knowledge argument is in line with principles in the development of the AVT: "Two words in a cluster having similar meaning could lead to ambiguity. It was therefore necessary to keep words with closely related meaning or similar definitions [. . .] from occurring in the same cluster" (Pecorari et al., 2019, p. 62).

Taken together, the two test designs (the VLT-Ac and the AVT) share several similarities: the matching format is the same, the target words constitute single-word units, the vocabulary knowledge tested is at the level of meaning-recognition (showing partial lexical knowledge), and the underlying domain is in both cases written receptive academic vocabulary. Out of the 57 target words in the AVT (Version 2), 24 words can be found in the AWL word families, and, conversely, 22 of the 30 VLT-Ac (Form 1) target words are also found among the AVL /f/lemmas. The two tests are designed to be representative of the vocabulary lists underlying each test and reflect the frequency distribution within the lists. Critically, this paper argues that the two tests have the same framework, that is, the domain assessed is in both cases written receptive English academic vocabulary at the level of meaning-recognition, even if this vocabulary domain is operationalized in slightly different ways due to the two different underlying lists of academic vocabulary. Hence, a linking procedure ought to be both meaningful and possible in order to compare the estimated scores based on each of the two tests. The current study proposes a comparison model and demonstrates how this can be done using responses on the VLT-Ac and the AVT.

Linking the tests with IRT

This study aims at comparing the estimated scores by linking the AVT and the VLT-Ac as two tests of written receptive meaning-recognition English academic vocabulary knowledge. Different tests can be used to measure the same underlying domain when the framework definition (a description of the skills/areas) is shared, and the test specifications (a description of item formats, number of items, scoring rules, etc.) are similar or different (Feuer et al., 1999). A linking procedure places the parameter estimates from different tests onto a common scale and "the most direct method for establishing and evaluating a linkage is the single-group design, in which two tests are administered to a common set of examinees" (Feuer et al., 1999, p. 45). When two tests measure the same individuals on the same content domain but are built on different test specifications, for example, different test lengths with unique test items, the common person test equating (e.g., Boone & Staver, 2020) or test calibration (e.g., Feuer et al., 1999; Kolen & Brennan, 2014) can be applied to link the test scores for comparability. To link the VLT-Ac and the AVT scores, the current study relies on a concurrent calibration procedure (e.g., Hanson & Béguin, 2002) within an IRT framework using a single-group design.

Method

Data collection

To link the VLT-Ac and the AVT, this study adopted a single-group counterbalance design (Kolen & Brennan, 2014). The participants (M_{age} = 18.09, SD_{age} = .31) were 385 Swedish upper secondary EFL learners. They had just completed the final English subject course necessary for university admittance with a passing grade equivalent to CEFR-B2. The participants were enrolled in six nationally regulated study programs (Arts, Economy, Humanities, Natural sciences, Social sciences, and Technology) preparing them for university studies. All participants were administered a single booklet with the 57-item AVT (Form 2) and the 30-item VLT-Ac (Version 1). To control for possible test order effects, half the sample received Booklet A with the AVT items followed by the VLT-Ac items, and the other half received Booklet B where the order was reversed.[3]

Scoring

Binary scoring was applied for each item in both tests. All examinees were encouraged to guess the answer provided they had the slightest intuition of the word meaning; blind guessing was discouraged. A *lenient* scoring approach giving credit for partial knowledge was adopted; two items on the AVT, identified as outliers in the initial exploratory IRT analyses, were examined lexically and were rescored using a *lenient* as opposed to a *severe* scoring approach (cf., Pecorari et al., 2019) since alternative words were judged to indicate partial knowledge.[4]

Data analyses

A series of analyses were conducted to check the viability of linking the two tests. First, initial CTT results of the two tests (57-set and 30-set) including test reliability, total score, standard deviation, percentile classification, distribution, and correlation were estimated and compared using IBM SPSS 27. Second, IRT analyses of the 57-, the 30-, and a combined 87- set were performed in R studio (RStudio Team, 2020) using the mirt package (Chalmers, 2012).

There are many different IRT models. For example, a one-parameter logistic (1PL) model/Rasch model (see, for example, Aryadoust et al., 2021; Baker, 2001; Wilson, 2004, for treating Rasch as 1PL), where the item discrimination parameter (a) is constant, estimates the item difficulty (b) which is located on the latent person ability scale (theta). Adding the item discrimination as a second parameter, the two-parameter (2PL) model estimates both a and b. In a three-parameter model (3PL), a third added parameter attempts to account for guessing (Hambleton et al., 1991). IRT models with a varying number of parameters have different levels of complexity. In general, the more complex the model, the larger the sample is needed to ensure the precision of the parameter estimates. Different requirements for the number of test takers and test items have been suggested. The range of used sample sizes varies considerably for the 1PL model (Aryadoust et al., 2021), and a minimum sample size of at least 300 in 1PL has been recommended (e.g., Guyer & Thompson, 2011). For 2PL,

recommendations differ, for example, 250 participants with 25 items (Harwell & Janosky, 1991) or with 30 items (Şahin, & Anıl, 2017), 300 with 75 items (Yoes, 1995). For a 3PL model, a sample of at least 1,000 participants would offer adequate estimates (Lord, 1968) also with respect to the number of items (e.g., Stoeckel et al., 2021). Two 1PL models (Rasch and $a = 1.7$), a 2PL model and a 3PL model, were tested with the current data in the exploratory phase. The exploratory analyses revealed that the 2PL model was the best-fitting model and was, therefore, chosen for the main analysis.

In the main analysis, the estimated item and person parameters in the three different item sets (VLT-Ac, AVT, and VLT-Ac & AVT combined) were compared to check for stability. The Test Characteristic Curves (TCC), the Test Information Functions (TIF), and the reliability curves were examined. Furthermore, two important assumptions about the data, that is, unidimensionality and local independence (LI), were assessed and verified (see, for example, Aryadoust et al., 2021; Hambleton et al., 1991).

One threat to test validity is that the items within a test battery measure several different abilities. It is, therefore, essential to test whether the assumption of unidimensionality holds, that is, if the data display one dominant factor influencing the test performance (Hambleton et al., 1991). This can be examined through, for example, exploratory factor analysis (EFA). If the first factor in EFA is distinctively larger in eigenvalue, then one can assume unidimensionality in the test (Wiberg, 2004). A principal component analysis (PCA) was used to extract the underlying factors of the test items (Aryadoust et al., 2021). Correlations, EFA, and reliability analysis were conducted in the current study to help to decide whether the data can be judged unidimensional.

Another threat to test validity is the interdependency of the responses among items in a test battery. The LI assumption of the test items can be examined with residual correlations among test items. A correlation (labeled Q3) above .3 warrants further consideration on whether the LI assumption holds (Aryadoust et al., 2021). Another way to assess LI is to examine whether the words in the booklets are lexically common. For example, if one item assesses knowledge of the word "procedure" and another item assesses the same root, for example, "proceed," the LI assumption is not met. However, since the matching test format in the tests can be contested regarding item interdependence for reasons of, for example, item exclusion strategy, the LI assumption of conducting IRT is violated and should be considered a limitation (Ha, 2021; Kamimoto, 2014; Stewart, 2012).

The final part of the main analysis adopted a concurrent calibration procedure. From the combined 87 set of item estimates placed on the mutual ability scale, the 57 AVT items and the 30 VLT-Ac items were extracted, and their respective TCCs were plotted in the same graph. This provided a means to identify the latent ability level (Θ) estimated for a VLT-Ac score that corresponds to an expected AVT score at the same ability level, and vice versa. Additional TCCs of the two tests were plotted with the 95% confidence intervals from the bootstrapped parameter estimates to check the robustness of the results.

Thereafter, a *score comparison table* was made with VLT-Ac scores in the left column and linked AVT scores in the right column. Since the estimation of person thetas has a standard error and the expected true score estimated with IRT will provide intervals

Table 1. Descriptive statistics of test scores on the VLT-Ac and the AVT, *N* = 385.

		VLT-Ac	AVT
Mean		24.52	30.4
95% CI for Mean	Lower bound	23.99	29.25
	Upper bound	25.06	31.55
Std. Deviation		5.3	11.46
Standardized mean		0.82	0.53
Standardized *SD*		0.18	0.20
Mode		29	31
Minimum		3	2
Maximum		30	54
Quartiles	25th	22	22
	50th	26	31
	75th	28.5	39
Skewness		–1.43	–0.11
Std. Error of Skewness		0.12	0.12
Kurtosis		1.76	–0.63
Std. Error of Kurtosis		0.25	0.25
Booklet A	*N* = 192	24.36	29.91
Booklet B	*N* = 193	24.69	30.9

Note: CI: confidence interval; VLT-Ac: academic section of the Vocabulary Levels Test; AVT: Academic Vocabulary Test.

between and across integer scores, a decision was made to include ranges of scores at certain levels.

Results

CTT results

Descriptive statistics of the scores on both tests are presented in Table 1. The standardized mean score (percentage correct) differs largely between the two tests. The average percentage correct for the VLT-Ac is higher than for the AVT ($M_{VLT\text{-}Ac}$ = .83 and M_{AVT} = .53), indicating that the VLT-Ac is generally easier in this test taker context. However, the standard deviation for the percentage correct is similar in both tests. The AVT-scores are normally distributed, whereas the VLT-Ac-sores are negatively skewed. No statistically significant differences were found between scores from the two booklets, VLT-Ac, $t(383) = .61$, $p = .54$; AVT, $t(383) = .85$, $p = .40$.

Reliability analyses for all three sets (87-set = the VLT-Ac plus the AVT; 57-set = the AVT; 30-set = the VLT-Ac) displayed high Cronbach's alphas (87-set $a = .95$; 57-set $a = .93$; 30-set $a = .89$). Moreover, the correlation between observed AVT and VLT-Ac test scores was strong ($r = .80$, $p < .001$; $\rho = .84$, $p < .001$). It should be noted that measurement errors in the observed test scores may attenuate the correlation coefficient. Thus, the tested latent trait by the two tests may correlate even higher.

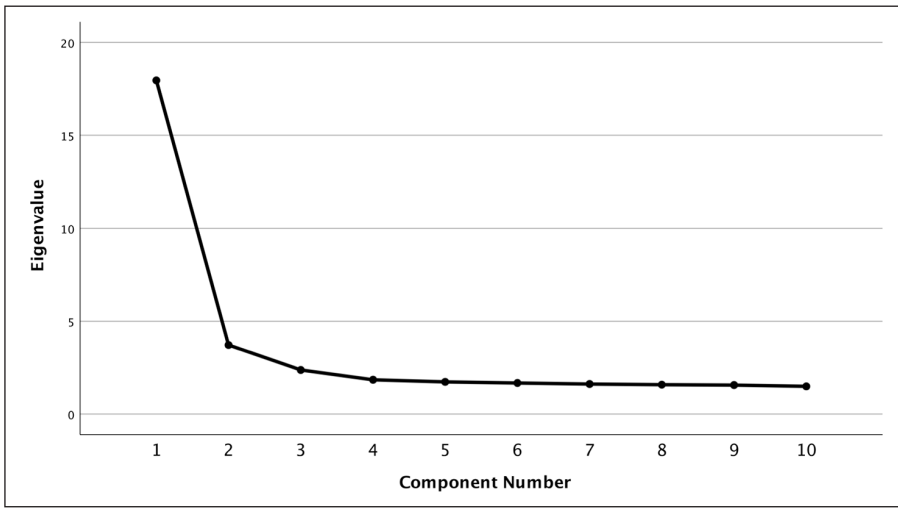

Figure 1. Scree plot showing one major factor.

Checking unidimensionality

A PCA was conducted on the normally distributed data in the combined 87 set. As displayed in the scree plot (Figure 1) one major factor explains most of the variance. The eigenvalue of the main factor (F1 = 17.95) was five times larger than the eigenvalue of the second factor (F2 = 3.72). Bartlett's test of sphericity was significant ($p < .001$), and the Kaiser–Meyer–Olkin measure (KMO; Kaiser & Rice, 1974) showed excellent sampling adequacy (KMO = .90). The average loading from each factor is .44 (*SD* = .12) (Appendix 3).

The inferences made from the reliability analysis, the correlation of observed scores and the factor analyses support the unidimensionality argument, that is, that both tests measure one main domain, namely written receptive English academic vocabulary knowledge at the level of meaning-recognition. Therefore, the items from the two tests can be merged into one 87-item bank in further IRT analyses.

Checking local independence

The correlation of item residuals was performed to check LI in the 87 set. Out of a total of 3741 correlated pairs, only three Q3 coefficients yielded a value above .3. The three correlated item pairs (Q3 > .3) were investigated qualitatively in the booklet and no lexical affinity between the targeted words could be claimed, *divergence—omission* (Q3 = .48); *exigence—proclivity* (Q3 = .64); *vexing—parsimonious* (Q3 = .37). All target words taken together, no common items exist. Notwithstanding the limitation of the matching format, the assumption of local independence was deemed met.

Comparisons of 2PL estimates in the different item sets

To ensure the stability of estimates, three models were performed, one for each set of items. The item parameter estimation is fully stable when the parameters estimated in the two separate tests are compared with the corresponding estimates from the 87 set. However, the 57 set (AVT) has more harder items than the 30-set (VLT-Ac), indicated by more items in the AVT with *b*s located above theta = .5 for this population. The 30-set (VLT-Ac) offers, on the other hand, more items located below theta = –2.5., that is, easier items for this test taker population, which is in line with the lists' word frequency distribution previously referenced in the background (Appendix 1).

When comparing the estimated person parameters in the three sets, a similar pattern occurs, namely that the person thetas estimated in the 30-set are not located at higher levels of the latent trait scale, whereas the AVT-estimated person thetas are located on the full scale. The correlation between the estimated person thetas in the 57-set and those in the 30-set is .84 ($p < .001$), indicating a large amount of shared variance between the two sets, and this supports the possibility to do a concurrent calibration. However, the imperfect correlation also revealed the lack of common information at the higher end of the scale in the two test sets.

Finally, the TIF and the reliability curves were examined for the two tests using the estimates from the combined set (Appendix 4). For the 30-set, the TIF (≈ 20) is above the threshold (> 10) (Wiberg, 2004), and the reliability is best in the range of approximately $\Theta = -3$ to 0.5. For the 57-set, the TIF is good (≈ 25) and the reliability is best in the range of approximately $\Theta = -2.5$ to 2.5.

Calibration and score comparison table

The results presented earlier indicated that a concurrent calibration of the AVT and the VLT-Ac was possible. Therefore, from the combined set of 87 items the 30-item VLT-Ac and the 57-item AVT were extracted. Their TCCs were then plotted in the same graph in Figure 2 with a shaded area illustrating the area where the TIF and the reliability are best shared.

The linking of scores between the tests was made by identifying an ability score on the VLT-Ac scale corresponding to the level of ability measured by AVT test. A first example, as is shown in Figure 2, is that the TCCs of the two tests intersect and this intersection is located on the ability scale of theta $\approx -.8$ (on the x-axis) at which level a score of approximately 22-point (on the y-axis) is expected on either test. Another example is the dotted vertical line in the center of the ability scale, i.e., located at theta = 0. At this ability level, an AVT score of 32 is expected, which corresponds approximately to a VLT-Ac score just above 26.

However, when the estimated person thetas and their standard errors are placed together with the expected true scores, it is evident that the estimation includes scores between possible integer test scores. For example, a VLT-Ac score of 26, located at theta ≈ 0, may range from a rounded AVT score of 31 to 33 as shown in Table 2.

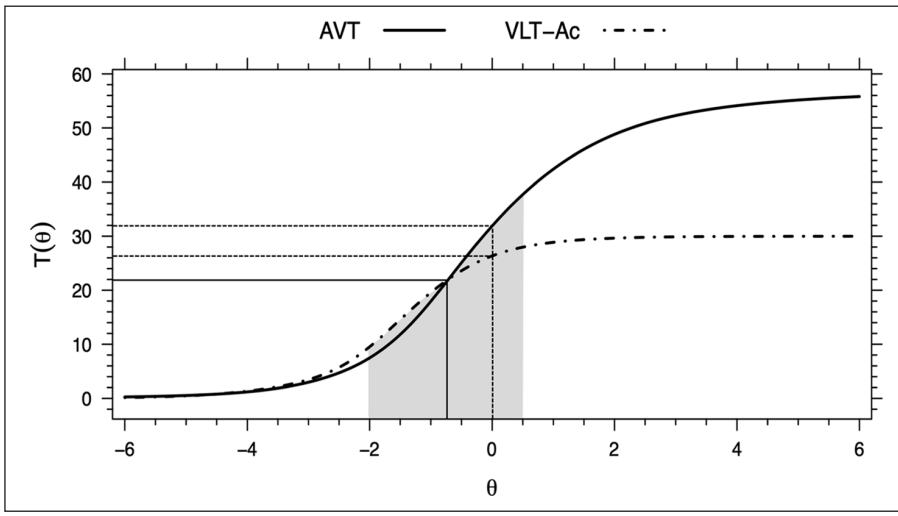

Figure 2. TCCs of the AVT and the VLT-Ac from the concurrent calibration.
Note: VLT-Ac: Vocabulary Levels Test Academic Section; AVT: Academic Vocabulary Test; TCC: Test Characteristic Curves.

Table 2. Example of variation of estimated thetas and standard error together with the expected test scores.

Participant	Estimated theta	*SE*	Expected VLT-Ac score	Expected AVT-score
198	–0.1	0.2	25.9	30.7
194	–0.1	0.2	26.1	31.1
214	0.0	0.2	26.3	31.6
200	0.0	0.2	26.4	32
248	0.0	0.2	26.6	32.5
351	0.1	0.2	26.7	33

Note: VLT-Ac: academic section of the Vocabulary Levels Test; AVT: Academic Vocabulary Test; *SE*: standard error.

Furthermore, the confidence intervals of the TCCs using 500 bootstrapped estimations were plotted (Appendix 5). In general, the variation in the estimations is very small, and this is particularly true for the AVT test.

Table 3 presents a score comparison between integer VLT-Ac scores in the left-hand column and the corresponding integer AVT scores in the right-hand column. By looking at the extracted TCCs, the shaded area, the plotted confidence intervals, and the estimations (see Table 2), it was decided, at certain levels, to include score ranges, which sometimes overlap in the lower or upper bounds. The reliability is lower at the ends of the scale and, therefore, the ranges in these levels are larger. The table is, therefore, appropriate for approximate comparisons of scores.

Table 3. Score comparison table.

VLT-Ac Version 1	AVT Form 2
1–10	0–8
11–13	9–10
14	11
15	12
16	13
17	14–15
18	15–16
19	17–18
20	18–19
21	20–21
22	22–23
23	24–26
24	26–28
25	28–31
26	31–33
27	33–37
28	37–41
29	42–49
30	50–57

Note: VLT-Ac: academic section of the Vocabulary Levels Test; AVT: Academic Vocabulary Test.

Discussion

The current study set out to find a model for linking scores from the AVT (Form 2) with scores from the VLT-Ac (Version 1) using a single-group counterbalance design with an IRT-based concurrent calibration procedure. The underlying purpose was to enable comparisons of measurements of academic vocabulary knowledge among CEFR-B2 level English users (students) enrolled in study programs preparing for university study across a wide range of academic disciplines.

The score comparison table (Table 3) constitutes the key finding of the present study since it enables the comparison of scores achieved on the VLT-Ac and the AVT. For instance, at the latent ability level of theta ≈ –.8, a VLT-Ac score of 22 compares approximately to an expected AVT score of 22, which may be considered a weak result possibly indicating very limited knowledge for the purpose of reading academic texts (Edgarsson, 2018).

A relevant application of the score comparison table is that it offers a means for comparing scores in situations where either the VLT-Ac or the AVT was administered or for predicting a test taker's score on the VLT-Ac based on his/her AVT score (or vice versa). A first example serves to illustrate the utility of score comparison between populations in a similar educational setting (upper secondary students in Norway and Sweden): For the Norwegian sample, Skjelde and Coxhead (2020) administered the VLT-Ac

(M_2=24.27), whereas Warnby (2022) used the AVT (M=29.63) with Swedish students. Based on the score comparison table, the two group-level mean scores can be approximately compared: a VLT-Ac score of 24.27 is approximately equivalent to an AVT score in the 26- to 28-point range; similarly, an AVT score of 29.63 is approximately equivalent to a VLT-Ac score of 25. A second example illustrates a comparison within the same population (upper secondary students in Sweden) over time: Gyllstad (2007) gave his students the VLT-Ac (M_2=18.20), whereas Warnby (2022) administered the AVT (M=29.63). Based on the score comparison table, the group-level mean scores from the two studies can be approximately compared: Gyllstad's observed VLT-Ac score of 18.20 corresponds more or less to an AVT score in the 15- to 16-point range; similarly (and as noted in the previous example), Warnby's observed AVT score of 29.63 is approximately equivalent to a VLT-Ac score of 25.

Score linking of the kind we are proposing here, and use of the score comparison table, are not without detractors, however, and caution should be exercised when interpreting the results. The AVT measures broadly on the full latent ability scale, whereas the VLT-Ac does not provide any information at the higher end of the scale. This makes it hard to compare the maximum VLT-Ac score with an AVT score; a test taker's VLT-Ac score of 30 may be predicted to range from approximately 50 to 57 points on the AVT. We can be more certain, however, that an AVT-score above 41 compares to a high score on the VLT-Ac (approximately in the 29- to 30-point range).

In contexts such as ours (involving CEFR-B2 EFL users with high exposure to English), the AVT is arguably a more appropriate test of written receptive meaning-recognition English academic vocabulary knowledge than is the VLT-Ac. This is because the AVT *(i)* gives test takers better opportunities to display a broader scope of their vocabulary knowledge and (*ii*) provides vocabulary researchers in this context with an instrument that is better at discriminating variation in the test takers' academic vocabulary knowledge. We noted earlier ceiling-effect issues experienced by some scholars using the VLT-Ac in contexts similar to ours (Edgarsson, 2018; Skjelde & Coxhead, 2020), resulting in potential underestimation of the true variation and increasing the risk of introducing type I errors (Austin & Brunner, 2003). In other contexts, however, depending on the specific context and purpose of the testing, the VLT-Ac could be entirely appropriate and offer high reliability with, for example, less proficient test takers (e.g., younger learners of English, as in Sundqvist (2009), or in expanding circle (Kachru, 1985) education settings where the overall exposure to English is more limited than in our context, as in McLean (2021). The findings of this study thus lend support to the claim that the use of a test must always be related to the purpose of the testing and to the context (e.g., Read, 2000; Schmitt et al., 2020). In this regard, Lin and Morrison's (2010) notion of a "standard" test of academic vocabulary knowledge becomes problematic ("standard" very much being a relative concept]. Instead, this study adopts a principle of "standardized" comparison, that is, a systematic way of comparing scores on two different vocabulary tests.

Vocabulary tests of the kind in this study, that is, measuring word knowledge at the meaning-recognition level using single-word units out of context, offer limited information on authentic language usage. First, the meaning-recognition format is not ideally suited to represent the lexical knowledge employable when reading; meaning-recall may be a better indicator (e.g., Kremmel & Schmitt, 2016; McLean et al., 2020; Nation

& Webb, 2011). Second, in addition to vocabulary knowledge, reading proficiency depends on a variety of skills, for instance, reading strategies and content knowledge. However, the single-word meaning-recognition format in the two linked tests in this study offers a quick way of collecting low-stakes information for diagnostic purposes of learners' academic vocabulary size. Relating such scores to other variables, for example, reading, it is evident that vocabulary knowledge, even at the low level of meaning-recognition, correlates significantly and positively with reading ability (e.g., Zhang & Zhang, 2020). This study was not designed to provide such additional variable comparison and cannot, therefore, say whether a VLT-Ac score of 26 (or a comparable AVT score of 32) may indicate "the extent to which learners and teachers need to focus on goals for academic vocabulary learning" (Skjelde & Coxhead, 2020, p. 6). However, the already identified relationships that academic vocabulary scores have with academic reading scores (e.g., Edgarsson, 2018; Shaw & McMillion, 2011) and with school grades (e.g., Skjelde & Coxhead, 2020) show how scores on tests such as the VLT-Ac and the AVT may explain variance in academic reading ability and academic achievement. This study adds a possibility to compare scores on such tests and may be used for comparisons in future correlational second language research.

Limitations and further work

We want to draw attention to some obvious limitations of this work and issue a caution concerning the utility of score comparison using the model presented in this paper.

First, readers are reminded that the comparison model arrived at is based on certain scoring principles adopted in relation to one form of the AVT and one version of the VLT-Ac; any comparison beyond these constraints may not be externally valid and may inflict upon, for example, the assumption of unidimensionality and/or the stability of IRT-estimations.

Second, this study offers a possibility to compare AVT-scores in relation to scores on the VLT-Ac and vice versa. However, errors of measurement should always be considered, and when aggregate scores are compared, the score distribution and sample sizes may vary. Therefore, we recommend that the score comparison table only be used for approximate comparisons in low-stake testing situations. Furthermore, this study is limited insofar as it cannot indicate what test scores mean in terms of, for example, mastery of the AVL/AWL or its relationship with reading or academic achievement.

Third, the matching format used in both the VLT-Ac and the AVT can be criticized for violating the assumption of local independence and introducing guessing possibilities (e.g., Gyllstad et al., 2015; Ha, 2021; Kamimoto, 2008, 2014; Stewart & White, 2011). The selection of IRT model is not really a concern in relation to this study since no substantial differences in linked scores were observed in the exploratory phase. However, with larger sample sizes, future studies may use the 3PL model to explore the intercepts for each item of a guessing parameter (McLean et al., 2015; Stewart, 2012; Stewart et al., 2017; Stewart & White, 2011).

Finally, this study uses a Swedish sample of CEFR B2 students preparing for university, and it is therefore impossible to speak to the generalizability of the estimations in populations with very different L1s (non-Germanic) or another English proficiency level.

Conclusion

The objective of the current study was to present a model for linking scores on two existing meaning-recognition tests of written receptive English academic vocabulary knowledge (the recently developed Academic Vocabulary Test, AVT, and the academic section of the Vocabulary Levels Test, VLT-Ac), in order to understand what a score on one test means in relation to the other.

The key contribution of this study is a score comparison table enabling approximate comparisons of scores achieved on the VLT-Ac and the AVT within the same context or between similar contexts to explore differences and trends over time. By using IRT modeling of the scores from both tests in a concurrent calibration, this study concludes that, in our context, (*i*) both tests measure test takers' knowledge of one broad domain (i.e., written receptive English academic vocabulary at the meaning-recognition level), (*ii*) the AVT measures on the whole latent scale and, therefore, seemingly functions better than the VLT-Ac as a test of written receptive meaning-recognition English academic vocabulary knowledge for test takers in this proficiency range, and especially, (*iii*) AVT and VLT-Ac scores can now be directly, albeit approximately, compared. Such comparisons can be highly relevant when vocabulary researchers or other test users want to establish links between academic vocabulary knowledge over time and/or with other closely related variables.

Acknowledgements

We thank the editors and the reviewers for their detailed and constructive comments during the revisions of the paper. Furthermore, we want to thank Dr. Eugenio Gonzalez at ETS Princeton for early advice on the linking methods and Senior Professor Gudrun Erickson at the University of Gothenburg for general comments regarding language assessment. Finally, we thank all participating schools and students.

Author contributions

Marcus Warnby (Conceptualization; Data curation; Formal analysis; Investigation; Project administration; Visualization; Writing—original draft; Writing—review & editing).
Hans Malmström (Conceptualization; Writing—review & editing).
Kajsa Yang Hansen (Conceptualization; Formal analysis; Methodology; Writing—review & editing).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Marcus Warnby https://orcid.org/0000-0001-9317-0233
Hans Malmström https://orcid.org/0000-0003-2450-7650
Kajsa Yang Hansen https://orcid.org/0000-0001-7071-2482

Notes

1. The VLT is free to download here: https://www.norbertschmitt.co.uk/vocabulary-resources.
2. Both forms and keys of the AVT can be downloaded here: https://www.en.cityu.edu.hk/Vocabulary-Tests.
3. In Appendix 2, the test formats used in the VLT-Ac, in the AVT and in the present study are presented.
4. The two leniently rescored items: the alternative *interconnect* can, according to Merriam-Webster dictionary, be treated as similar to the given definition *attach* matched originally with the target word *append*. The same relationship goes for the alternative *defensible* which, justifiably, may be matched with the definition *providing support* (original target word: *ancillary*). Future revisions of the AVT are suggested to rewrite these items.

References

Aryadoust, V., Ng, L., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing*, *38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Austin, P. C., & Brunner, L. J. (2003). Type I error inflation in the presence of a ceiling effect. *The American Statistician*, *57*(2), 97–104. https://www.jstor.org/stable/30037242

Aviad-Levitzky, T., Laufer, B., & Goldstein, Z. (2019). The new computer adaptive test of size and strength (CATSS): Development and validation. *Language Assessment Quarterly*, *16*(3), 345–368. https://doi.org/10.1080/15434303.2019.1649409

Baker, F. B. (2001). *The basics of item response theory* (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Bonnet, G. (Ed.). (2004). *The assessment of pupil's skills in English in eight European countries 2002*. The European Network of Policy Makers for the Evaluation of Educational Systems.

Boone, W. J., & Staver, J. R. (2020). Common person test equating. In J. W. Boone & J. R. Staver (Eds.), *Advances in Rasch analyses in the human sciences* (pp. 147–158). Springer. https://doi.org/10.1007/978-3-030-43420-5_11

Brown, D., Stoeckel, T., Mclean, S., & Stewart, J. (2020). The most appropriate lexical unit for L2 vocabulary research and pedagogy: A brief review of the evidence. *Applied Linguistics*, *43*(3), 596–602. https://doi.org/10.1093/applin/amaa061

Busby, N. L. (2020). Words from where? Predictors of L2 English vocabulary among Norwegian university students. *ITL–International Journal of Applied Linguistics*, *172*(1), 58–84. https://doi.org/10.1075/itl.19018.bus

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Charles, M., & Pecorari, D. (2015). *Introducing English for academic purposes*. Taylor & Francis. https://doi.org/10.4324/9781315682129

Cobb, T. (2022). *Compleat Web VP v.2.6* [Vocabulary Profiler computer program]. Lextutor. https://www.lextutor.ca/vp/comp/

Council of Europe. (2022). *The CEFR levels*. Council of Europe. https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Coxhead, A. (2016). Acquiring academic and disciplinary vocabulary. In K. Hyland & P. Shaw (Eds.), *The Routledge Handbook of English for Academic Purposes*, (pp. 177–190). Routledge. https://doi.org/10.4324/9781315657455

Edgarsson, G. (2018). Academic vocabulary proficiency and reading comprehension among Icelandic secondary school students. In B. Arnbjörnsdóttir & H. Ingvarsdóttir (Eds.), *Language*

development across the life span (pp. 95–112). Springer. https://doi.org/10.1007/978-3-319-67804-7_6

Education First. (2020). *The EF English proficiency index 2020*. https://www.ef.com/wwen/epi/

European Commission. (2012). *First European survey on language competences: Final report*. Publications office of the European Union.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Uncommon measures: Equivalence and linkage among educational tests*. National Academy Press. https://doi.org/10.17226/6332

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

Guyer, R., & Thompson, N. A. (2011). *User's manual for Xcalibre 4.1*. Assessment Systems Corporation.

Gyllstad, H. (2007). *Testing English collocations: Developing receptive tests for use with advanced English learners* [Doctoral dissertation, Lund University]. Lund University Publications. https://lup.lub.lu.se/search/ws/files/5893676/2172422.pdf

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

Ha, H. T. (2021). A Rasch-based validation of the Vietnamese version of the listening Vocabulary Levels Test. *Language Testing in Asia*, *11*, Article 16. https://doi.org/10.1186/s40468-021-00132-7

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (*Vol. 2*). SAGE.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*(1), 3–24. https://doi.org/10.1177/0146621602026001001

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied psychological measurement*, *15*(3), 279–291. https://doi.org/10.1177/0146621691015003

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, *41*(2), 235–253. https://doi.org/10.1002/j.1545-7249.2007.tb00058.x

Hyltenstam, K. (2004). Engelskan, skolans språkundervisning och svensk språkpolitik [English, language teaching and Swedish language policy]. In O. Josephson (Ed.), *Engelskan i Sverige. Språkval i utbildning, arbete och kulturliv [English in Sweden. Language choice in education, working life and culture]* (pp. 36–110). Norstedts.

Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In Q. Randolph & H. G. Widdowson (Eds.), *English in the world* (pp. 11–30). Cambridge university press.

Kaiser, H. F., & Rice, J. (1974). Little Jiffy, Mark IV. *Educational and Psychological Measurement*, *34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Kamimoto, T. (2008). *Nation's Vocabulary Levels Test and its successors: A re-appraisal* [Doctoral dissertation, Swansea University]. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.564017

Kamimoto, T. (2014). Local item dependence on the Vocabulary Levels Test revisited. *Vocabulary Learning and Instruction*, *3*(2), 56–68. https://doi.org/10.7820/vli.v03.2.kamimoto

Kolen, M. J., & Brennan, R. L. (2014). *Statistics for social and behavioral sciences. Test equating, scaling, and linking: Methods and practices* (3rd ed.). Springer.

Kremmel, B. (2016). Word families and frequency bands in vocabulary tests: Challenging conventions. *TESOL Quarterly*, *50*(4), 976–987. https://doi.org/10.1002/tesq.329

Kremmel, B. & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, *13*(4), 377–392. https://doi.org/10.1080/15434303.2016.1237516

Laufer, B., & Aviad–Levitzky, T. A. (2017). What type of vocabulary knowledge predicts reading comprehension: Word meaning recall or word meaning recognition? *The Modern Language Journal*, *101*(4), 729–741.

Lemmouh, Z., & Snoder, P. (2019, August 29). Vocabulary learning milestones: A study of the receptive vocabulary size of Swedish adolescent EFL learners [Conference presentation]. *EuroSLA 29*, Lund, Sweden.

Lin, L. H. F., & Morrison, B. (2010). The impact of the medium of instruction in Hong Kong secondary schools on tertiary students' vocabulary. *Journal of English for Academic Purposes*, *9*(4), 255–266. https://doi.org/10.1016/j.jeap.2010.09.002

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, *28*(4), 989–1020. https://doi.org/10.1177/001316446802800401

Malmström, H., & Pecorari, D. (2022). *Language choice and internationalisation: The roles of Swedish and English in research and higher education*. Institutet för språk och folkminnen.

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*(6), 823–845. https://doi.org/10.1093/applin/amx003

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. http://hdl.handle.net/10125/67396

McLean, S., Kramer, B., & Stewart, J. (2015). An empirical examination of the effect of guessing on vocabulary size test scores. *Vocabulary Learning and Instruction*, *4*(1), 26–35. https://doi.org/10.7820/vli.v04.1.mclean.et.al

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *37*(3), 389–411. https://doi.org/10.1177/0265532219898380

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly*, *20*(2), 233–253. https://doi.org/10.2307/747758

Nation, I. S. P. (1983). Testing and teaching vocabulary. *Guidelines*, *5*(1), 12–25.

Nation, I. S. P. (2012). *The Vocabulary Size Test: Information and specifications*. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-tests/the-vocabulary-size-test/Vocabulary-Size-Test-information-and-specifications.pdf

Nation, I. S. P. (2013). *Learning vocabulary in another language. Second edition*. Cambridge university press.

Nation, I. S. P. (2019). The different aspects of vocabulary knowledge. In P. Nation (Ed.), *The Routledge handbook of vocabulary studies* (pp. 13–29). Routledge. https://www.routledgehandbooks.com/doi/10.4324/9780429291586-2

Nation, I. S. P., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Cengage learning.

National Agency for Education. (2012). *Internationella språkstudien [International language study 2011]*. Skolverket. https://www.skolverket.se/publikationer?id=2832

National Agency for Education. (2021). *Kommentarmaterial till ämnesplanerna i moderna språk och engelska [Commentary material for the syllabi in modern languages and English]. Skolverket*. https://www.skolverket.se/publikationer?id=7842

National Agency for Education. (2022). *Gymnasieskolan—Elevstatistik [The upper secondary school – Statistics about students]*. Läsåret 2021/22. Skolverket. https://siris.skolverket.se/reports/rwservlet?cmdkey=common&geo=1&report=gy_elever&p_ar=2021&p_lankod=&p_

kommunkod=&p_skolkod=&p_sub=1&p_inriktning=0&p_hmantyp=&p_visahmt=0&p_flik=G&p_programkod=

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

Pecorari, D., Shaw, P., Malmström, H., & Irvine, A. (2011). English textbooks in parallel-language tertiary education. *TESOL Quarterly*, *45*(2), 313–333. https://doi.org/10.5054/tq.2011.247709

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

RStudio Team. (2020). *RStudio: Integrated development for R*. RStudio. http://www.rstudio.com/

Sahin, A., & Anil, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory and Practice*, *17*(1), 321–335. https://doi.org/10.12738/estp.2017.1.0270

Schmitt, N. (2010). *Researching vocabulary*. Palgrave Macmillan.

Schmitt, N., Nation, I. S. P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *52*(4), 1–12. https://doi.org/10.1017/S0261444819000326

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503. https://doi.org/10.1017/s0261444812000018

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

Shaw, P., & McMillion, A. (2011). Components of success in academic reading tasks for Swedish students. *Ibérica*, *22*(22), 141–162. https://www.diva-portal.org/smash/get/diva2:466682/FULLTEXT01.pdf

Skjelde, K., & Coxhead, A. (2020). Mind the gap: Academic vocabulary knowledge as a predictor of English grades. *Acta Didactica Norden*, *14*(3), Article 6. https://doi.org/10.5617/adno.7975

Stewart, J. (2012). Does IRT provide more sensitive measures of latent traits in statistical tests? An empirical examination. *Shiken Research Bulletin*, *16*(1), 15–22.

Stewart, J., McLean, S., & Kramer, B. (2017). A response to Holster and Lake regarding guessing and the Rasch model. *Language Assessment Quarterly*, *14*(1), 69–74. https://doi.org/10.1080/15434303.2016.1262377

Stewart, J., Stoeckel, T., McLean, S., Nation, P., & Pinchbeck, G. (2021). What the research shows about written receptive vocabulary testing: A reply to Webb. *Studies in Second Language Acquisition*, *43*(2), 462–471. https://doi.org/10.1017/s0272263121000437

Stewart, J., & White, D. (2011). Estimating guessing effects on the Vocabulary Levels Test for differing degrees of word knowledge. *TESOL Quarterly*, *45*(2), 370–380. https://www.jstor.org/stable/41307638

Stoeckel, T., McLean, S., & Nation, P. (2021). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, *43*(2021), 181–203. https://doi.org/10.1017/s027226312000025x

Sundqvist, P. (2009). *Extramural English matters: Out-of-school English and its impact on Swedish ninth graders' oral proficiency and vocabulary* [Doctoral dissertation, Karlstad University]. Karlstad University Studies.

Swedish Media Council. (2019). *Ungar & medier 2019*. Statens medieråd. https://www.statensmedierad.se/download/18.1ecdaa0017633a0d6666107/1607510806657/Ungar%20och%20medier%202019%20tillganglighetsanpassad.pdf

Therova, D. (2020). Review of academic word lists. *TESL-EJ–Teaching English as a Second or Foreign Language*, *24*(1), 1–15. https://doi.org/10.7820/vli.v09.1.therova

Vilkaitė-Lozdienė, L., & Schmitt, N. (2019). Frequency as a guide for vocabulary usefulness. High-, mid- and low-frequency words. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 81–96). Routledge. https://doi.org/10.4324/9780429291586-6

Warnby, M. (2022). Receptive academic vocabulary knowledge and extramural English involvement—Is there a correlation? *ITL–International journal of applied linguistics*, *173*(1), 120–152. https://doi.org/10.1075/itl.21021.war

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test. *ITL–International Journal of Applied Linguistics*, *168*(1), 33–69. https://doi.org/10.1075/itl.168.1.02web

West, M. (1953). *A general service list of English words*. Longman, Green and Co.

Wiberg, M. (2004). *Classical test theory vs. Item response theory: An evaluation of the theory test in the Swedish driving-license test*. Umeå university.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Routledge. https://doi.org/10.4324/9781410611697

Yoes, M. (1995). *An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model*. Saint Paul, MN: Assessment Systems Corporation.

Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, *26*(4), 696–724. https://doi.org/10.1177/1362168820913998

Appendix 1

Frequency distribution of the AWL, the AVL, and words included in the two tests

Although the purpose of this study is not to perform lexical analyses of the words contained in the lists (and, thus, in the tests), some sort of visualization of the frequency levels within the lists may give the reader a better understanding of how the academic words in the lists and the tests would be placed in terms of frequency levels in a corpus of general English. The presentation below broadly illustrates the frequency levels. More in-depth and elaborated analyses are beyond the scope of this study.

In Figure 3, the words in the AWL, the AVL, the VLT-Ac and the AVT are grouped together in four frequency categories: 1 K to 2 K, followed by 3 K to 9 K, then 10 K to 12 K, and, finally, 13 K to 25 K. This categorization roughly follows Nation (2013) who labels "high-frequency words (the most frequent 2,000 word families [. . .], the mid-frequency words (7,000 word families from the 3rd to the 9th 1,000-word lists inclusive) [. . .], and the low-frequency words (10th 1,000- word list onward)" (p.16). However, the low-frequency category has been divided into two categories, the first category ending at 12,000 (12 K). The reason was that such a division better reflected the tested frequency level (see Figure 3).

Separately, we entered the AWL words, the AVL words, the VLT-Ac, and the AVT test item options into the vocabulary research tool Compleat Web Vocabulary Profiler v.2.6 (VP-Compleat) (Cobb, 2022) which gives the frequency placement of the words from 1 K to 25 K in the *British National Corpus* (BNC) and the *Corpus of Contemporary American English* (COCA). For the AWL, all the words were entered as given in the

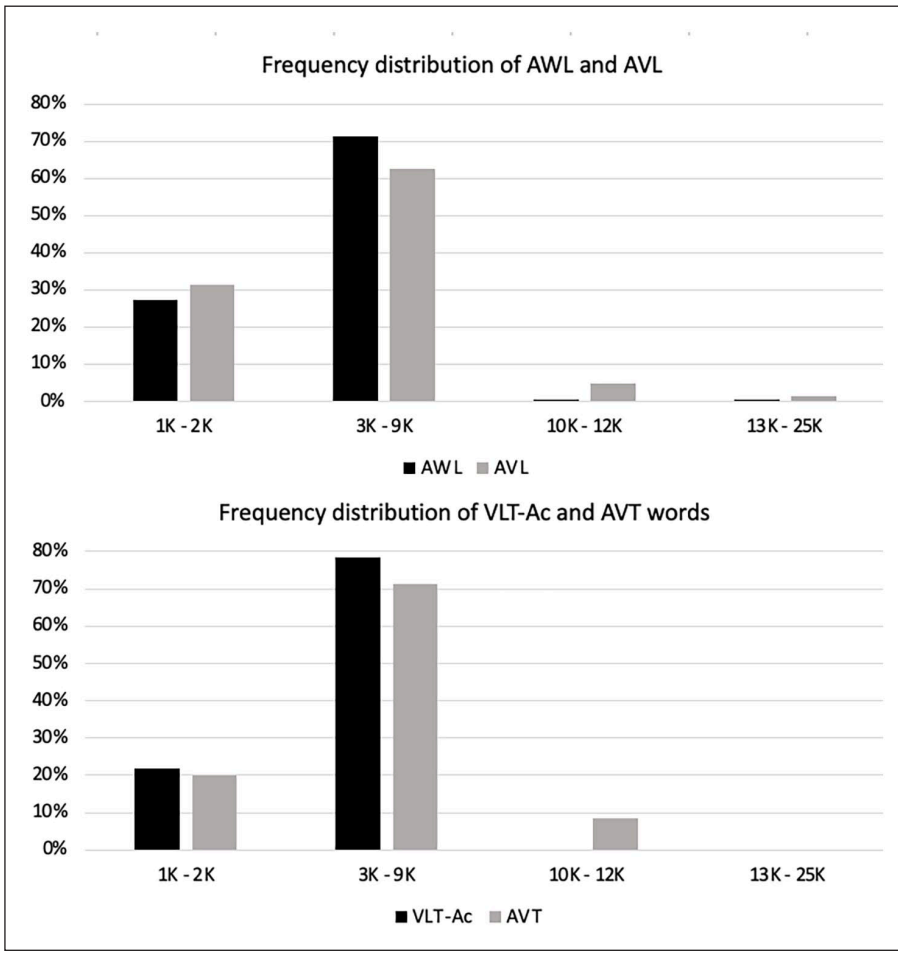

Figure 3. Frequency distribution of the AWL word family types, the AVL lemmas, the VLT-Ac and the AVT test item options.

sublist family version on Coxhead's homepage, https://www.wgtn.ac.nz/lals/resources/academicwordlist/sublist. For the AVL, we took all the lemmas as they are provided on https://www.academicwords.info/.

Appendix 2

Matching formats in the VLT-Ac, the AVT, and the present study

In Figure 4, the matching formats used in the revised VLT-Ac (Schmitt et al., 2001), the AVT (Pecorari et al., 2019), and the present study are exemplified.

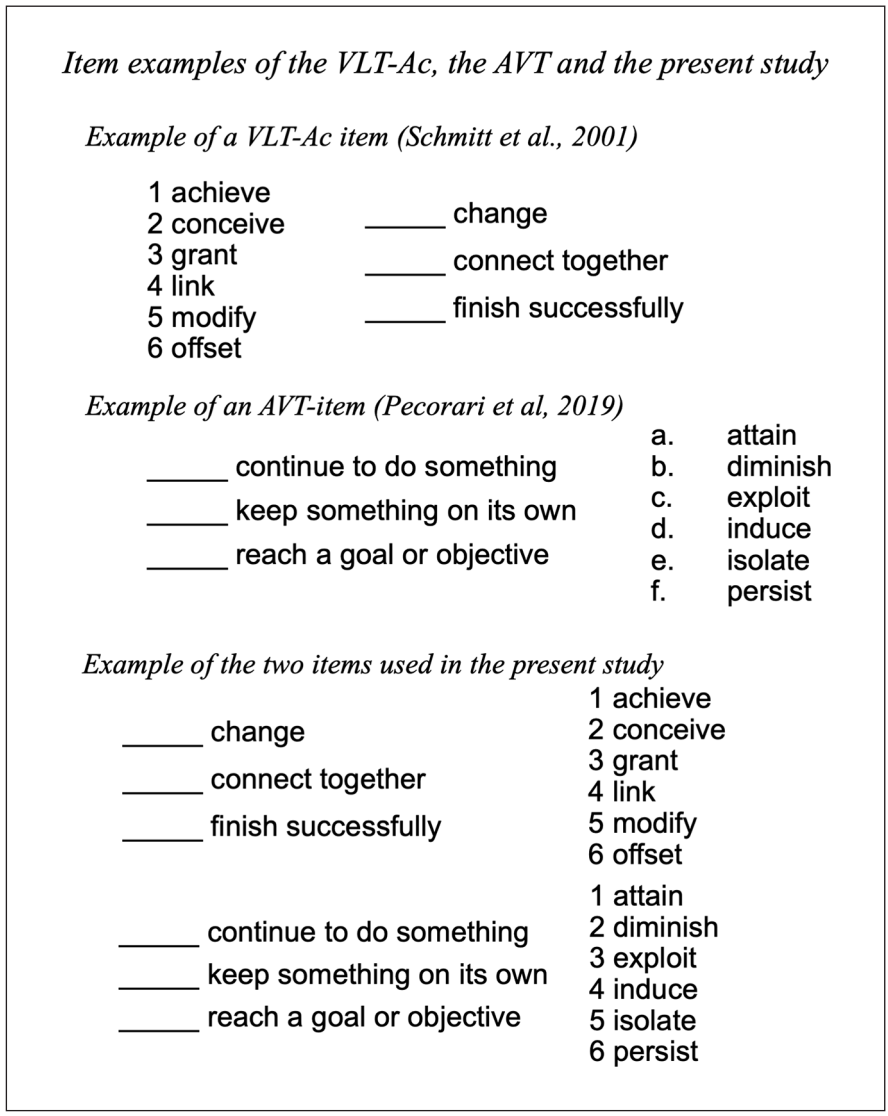

Figure 4. Examples of items in the VLT-Ac (Schmitt et al., 2001), the AVT (Pecorari et al., 2019), and the present study.

Appendix 3

PCA—factor loadings on one major factor

In Table 4, the factor loadings from the PCA on the one major factor within the AVT and the VLT-Ac.

Table 4. PCA—factor loadings on one major factor.

Item	F1	Item	F1	Item	F1
AVT1	.55	AVT30	.50	VLT2	.47
AVT2	.52	AVT31	.39	VLT3	.52
AVT3	.54	AVT32	.60	VLT4	.53
AVT4	.51	AVT33	.23	VLT5	.52
AVT5	.24	AVT34	.47	VLT6	.39
AVT6	.32	AVT35	.46	VLT7	.44
AVT7	.56	AVT36	.45	VLT8	.65
AVT8	.52	AVT37	.38	VLT9	.44
AVT9	.50	AVT38	.35	VLT10	.35
AVT10	.58	AVT39	.47	VLT11	.30
AVT11	.36	AVT40	.31	VLT12	.49
AVT12	.51	AVT41	.20	VLT13	.45
AVT13	.50	AVT42	.48	VLT14	.47
AVT14	.56	AVT43	.16	VLT15	.33
AVT15	.59	AVT44	.19	VLT16	.64
AVT16	.58	AVT45	.18	VLT17	.40
AVT17	.39	AVT46	.24	VLT18	.47
AVT18	.48	AVT47	.35	VLT19	.44
AVT19	.55	AVT48	.43	VLT20	.14
AVT20	.53	AVT49	.52	VLT21	.46
AVT21	.55	AVT50	.50	VLT22	.40
AVT22	.42	AVT51	.36	VLT23	.66
AVT23	.52	AVT52	.30	VLT24	.28
AVT24	.50	AVT53	.45	VLT25	.46
AVT25	.34	AVT54	.22	VLT26	.49
AVT26	.68	AVT55	.40	VLT27	.48
AVT27	.57	AVT56	.45	VLT28	.47
AVT28	.56	AVT57	.34	VLT29	.47
AVT29	.36	VLT1	.50	VLT30	.33

Note: PCA: principal component analysis; AVT: Academic Vocabulary Test; VLT: Vocabulary Levels Test.

Appendix 4

TIF and Reliability curves for the three sets as estimated with the combined 87set

Figure 5 presents two plots. The first plot presents the TIFs of the combined set of items, the AVT and the VLT-Ac. The second plot shows the reliability curves for the three sets.

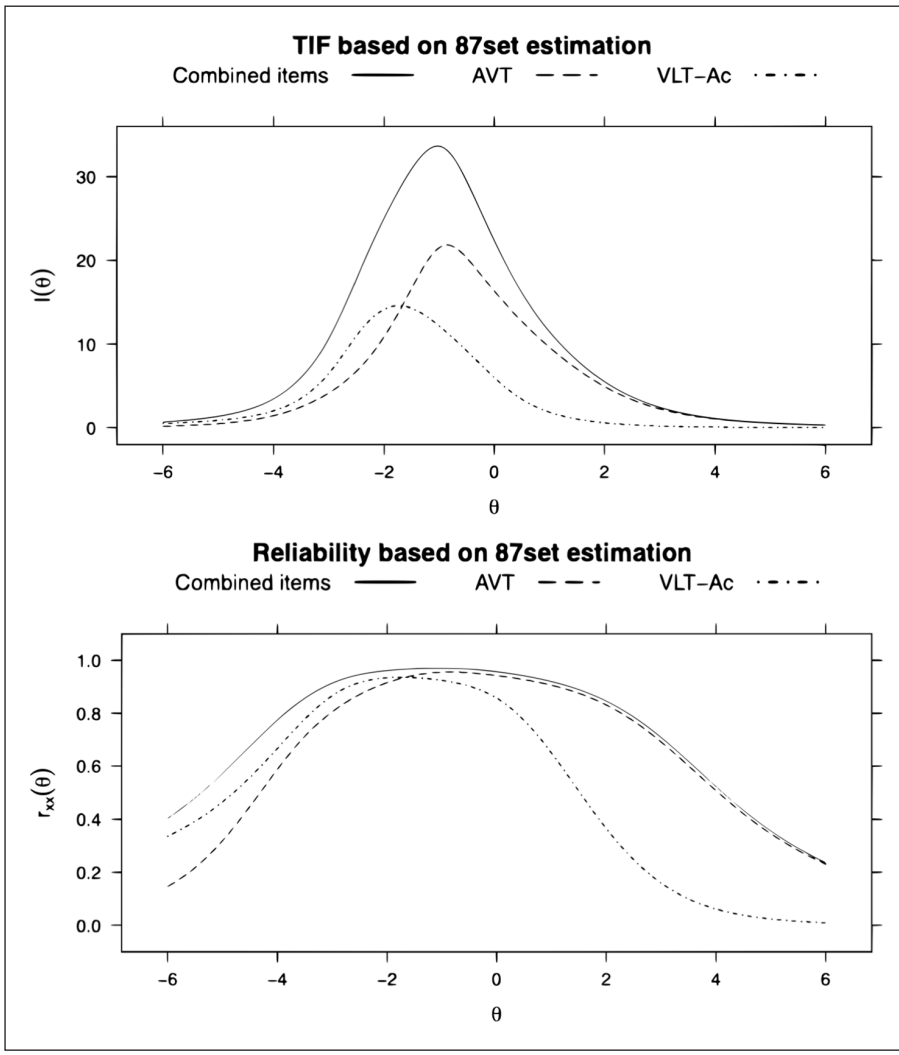

Figure 5. TIF and Reliability curves for the three sets as estimated with the combined 87set.

Appendix 5

TCCs with confidence interval from 500 bootstraps

Figure 6 presents the TCCs for each test estimated with the 87-set when the information matrix is computed with 500 iterations. The gray shadow around the solid TCC curve illustrates the confidence interval based on these 500 bootstraps.

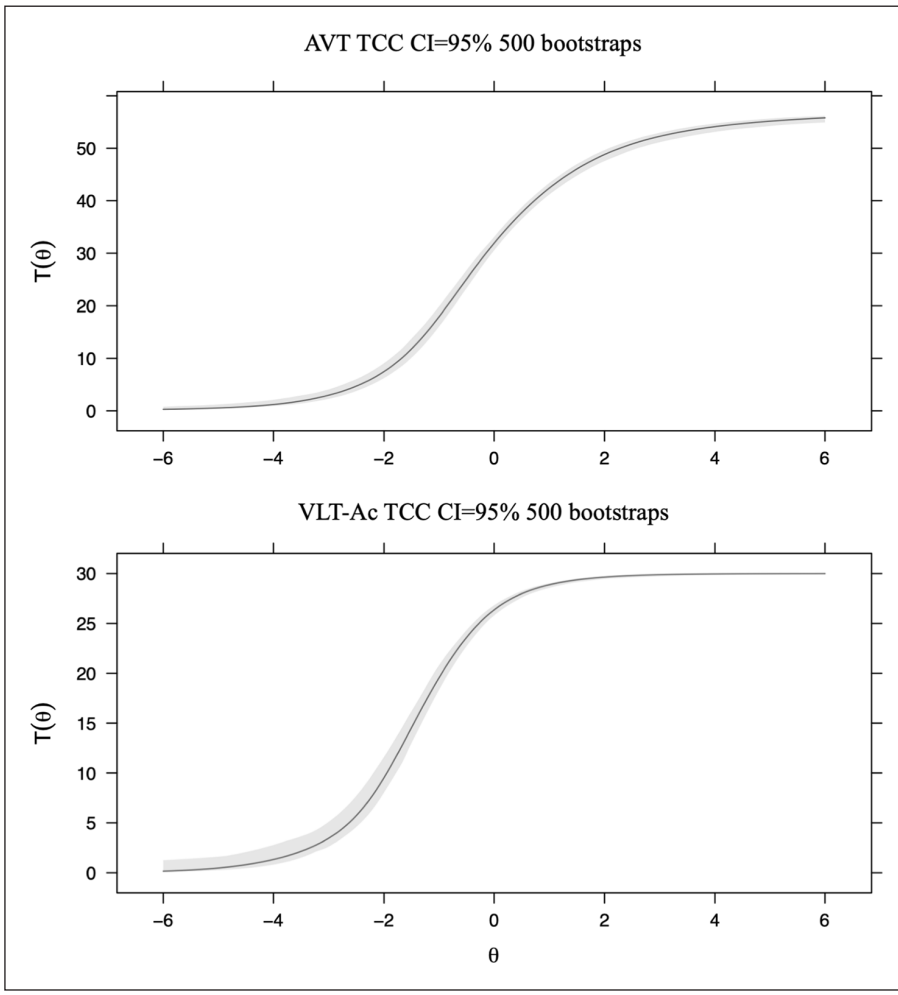

Figure 6. TCCs with a shaded 95% confidence interval from 500 bootstraps.