Chapter 11
The Vienna English Language Test

Susanne Sweeney-Novak

Keywords Testing grammar and vocabulary · Entrance test · Proficiency test · Multiple-choice test · Test development and validation

11.1 Contextualisation and Test Purpose

The Vienna English Language Test (VELT) was developed at the Department of English and American Studies at the University of Vienna and first implemented in 2011. It is a multiple-choice test used to determine undergraduate students' language knowledge in the areas of vocabulary and grammar with a view to ensuring level B2+ according to the *Common European Framework of Reference for Languages* (CEFR, Council of Europe, 2001). The VELT is not an entrance requirement for the undergraduate programme: failing the test does not exclude a student from attending lectures. However, students have to pass the VELT to be able to register for the courses in the department's English Language Competence (ELC) programme. There is no limit to the times this test can be taken.

Since the inception of a new curriculum in 2002 at the department, a commercially available standardised test, referred to below as the "old test," had been administered to first-semester students at the beginning of their first ELC course to establish their proficiency level in accordance with the CEFR. Monitoring the results over time made it clear that about 20% of students did not meet B2 level, which is supposed to be the school-exit level in the Austrian context. It made sense that these students should first improve their language competence before being offered a place in the ELC courses at the department.

S. Sweeney-Novak (✉)
Department of English and American Studies, University of Vienna, Vienna, Austria
e-mail: susanne.sweeney-novak@univie.ac.at

A. Berger et al. (eds.), *Developing Advanced English Language Competence*,
English Language Education 22, https://doi.org/10.1007/978-3-030-79241-1_11

Test security and the wish to create a specific departmental test led to the development of the VELT. Unlike the old test, which targeted all the proficiency levels from A1 to C2 of the CEFR, the purpose of the VELT is to establish whether a test taker is proficient in grammar and vocabulary at level B2+ and above. Therefore, the VELT only marginally includes items at B1 level and below.

11.2 Test Construct

The theoretical framework adopted for the design of the VELT was taken from Purpura's theoretical definition of grammar, which covers phonological/graphological, lexical, morphosyntactic, cohesive, information management, and interactional forms and meanings at the subsentential, sentential, and suprasentential/discourse levels (2004, p. 78). The VELT has a strong focus on lexical, morphosyntactic, cohesive, and information management forms and meanings at the sentential and suprasentential levels. In his description of corpus linguistics, Purpura (2004, p. 15) shows that there are features of language use which could be "taken as both lexical and grammatical": for example, the word *since* has both a "lexical dimension," its meaning, and a "grammatical dimension" as a clause marker or a preposition.

Indeed, when analysing and labelling individual items first in the old test and then in the VELT, it was not always clear whether an item was testing vocabulary knowledge or grammatical knowledge, which confirmed Purpura's position that grammar and vocabulary are not separate traits but are interrelated in language use. One example may illustrate the point: the teaching of vocabulary should also focus on colligations, which are the syntactic environment of a specific word and are an important aspect of knowing a word. Most English as foreign language learners taking the VELT would probably be familiar with the word *upset* when used as a verb with an animate object with the meaning of making a person sad or anxious. However, in our data only a very small number of test takers were able to recognise the use of *upset* with an inanimate object in the sentence "The airline's insolvency upset our holiday plans."

In the VELT, there are no separate grammar and vocabulary sections, and vocabulary and grammar are tested context-dependently. In the example above, the use of the word *upset* is embedded in a sentence whose syntactic features should elicit the correct response over incorrect distractors. A context-independent response would be, for example, word-definition matching, as in Nation's Vocabulary Levels Test (see Read, 2000, pp. 9–13 on the dichotomy of context-dependent and context-independent and pp. 118–120 on Nation's Vocabulary Levels Test).

The test construct of the VELT includes English morphology, lexis, and syntax. In contrast to many vocabulary tests, which ask for definitions of words or elimination of non-words, the VELT focuses on the meanings of words, semantic fields, and collocations in context. In the more current versions of the test, items were added related to word order (e.g., inversion) to express emphasis.

11.3 Test Method and Administration

In format and length, the VELT mirrors the old test. Each version consists of 60 items; time given for completion is 30 minutes. For test administration and security purposes, candidates are randomly divided into two groups, necessitating the development of two equivalent versions of the VELT. Due to the large number of students tested, this paper-based test must be administered speedily and results produced quickly. Since the answer sheets can be machine read, it is possible to have the results within a matter of hours. Success or failure is reported to students as percentage points away (positively or negatively) from the cut score.

The consistent format of the VELT is four-options multiple choice (MC) with one correct answer. The decision to adopt a MC format only, rather than develop a test using a variety of formats, was governed by the assumption that this is a format most likely known to all test takers. Furthermore, according to Purpura (2004), despite the criticism they receive, MC items are well suited for testing discrete features of grammatical knowledge. This claim would also apply to knowledge of vocabulary. In addition, MC items can be scored objectively, thus avoiding any subjective interpretation of student answers.

The VELT consists of individual sentences and five short text passages of 70–90 words with 7–9 gaps each. The passages include a range of text types with selected gapped items which require the test taker to supply a missing word or phrase chosen from four options. The purpose of the reading passages is to test beyond the sentence and to focus on text-specific features, for example past tense, participle clauses, or logical connectors. All short reading passages are authentic texts taken from different sources. Sometimes it is necessary to make minor adaptations to the text, for example to ensure that the content is not biased, that world knowledge is not required to understand the content, or to be in keeping with the required length of these short passages.

11.4 Test Development and Piloting

11.4.1 Collecting Response Data

Initially, selected items from various published test papers were given to students at the beginning of their first semester. These items were also administered to final-year pupils in schools to compare the results with first-semester students at the department. The purpose of using papers from published tests was to gain an understanding of the level of proficiency that pupils and university students of English were at and which items denoting lexical or grammatical features were typical of a specific level of proficiency. At the same time, independently constructed items were piloted to see whether these correlated with standardised items at specific levels of proficiency. Lexical items were included which were taken specifically from the Academic Word List developed by Coxhead (2000). A considerable part of the

discrete sentences was taken from corpora, such as the British National Corpus, or dictionaries whose examples are based on a corpus, such as the COBUILD dictionary, which is based on the UK's Birmingham University Language Database and is a pioneering work in dictionary compilation of modern English usage. Once we had accumulated a bank of at least 200 items, we investigated item difficulty, item discrimination, and distractor quality. Those items which proved statistically unsatisfactory as regards their level of difficulty and/or their ability to discriminate between high and low scorers were discarded. For the major trial, all items were classified regarding their level and purpose, and two versions of a trial test were developed. Up to the present time, the trialling and piloting procedures of two versions described above remain the same.

11.4.2 Including and Excluding Items

Items which look appropriate from the point of view of item difficulty and discrimination are piloted. Distractor analyses are conducted, and adjustments of weak distractors are made. Analytical software is used for analysis, namely SPSS and Winsteps (Linacre, 2019). The Rasch dichotomous model (Winsteps) can compare *person ability* and *item difficulty*, informing test developers of the probability of test takers answering specific items correctly. In addition, the Rasch model calculates *fit data* (person and item fit or misfit). This means that items or persons that produce surprising answers are indicated as not fitting the model. Items that the model specifies as misfitting are discarded.

In addition to omitting items which are found statistically wanting, some items are discarded or amended because of the feedback received from students and colleagues. These could include biased items or items which do not take into consideration language change. For example, on one occasion a student pointed out that they could only complete an item by elimination, because the "correct" answer was given in 'British English' rather than 'American English,' with which they were familiar. Besides, with an increasing number of students whose first language is not German or who come from a non-Austrian background, cultural bias in test development should be borne in mind. For test security reasons, student feedback is limited to some points they would like to make after the trial tests have been collected. By contrast, feedback from colleagues, especially those who revise the first drafts, is extensive.

11.4.3 Trialling

Originally, two versions, referred to as Version 1 and Version 2 below, of the future test were trialled and correlation studies with the old test were conducted. Today we trial the test with students at the beginning of their ELC programme. The trial

population has, in fact, passed a VELT and should at least be at level B2+. All items and more for two versions are trialled, and correlation studies are conducted between each student's test and trial result.

11.4.4 Standard Setting

Another step in the development of the VELT was the setting of cut scores between CEFR levels and, most specifically, between a pass and a fail. Various parameters were taken into account to tackle the question. First of all, the difficulty (facility value) of an item gave some indication as to whether an item was easy or difficult. The results of the old test enabled us to match test takers' CEFR level based on the old test with their scores on the trial versions. We also used the Rasch-based person/item map, together with the facility values, to determine the cut-off points between levels.

A second parameter was the judgement of experienced colleagues who had extensive teaching experience at the school-exit level. They were asked to scrutinise the items and decide whether a student at this level, which is supposed to be B2, would be able to answer an item correctly, whether they would regard an item as below the school-exit level, or whether only more advanced students would be able to answer an item correctly.

Thirdly, classifications according to CEFR levels were taken into account. To this end, the English Vocabulary Profile and the English Grammar Profile (English Profile, 2015), as well as Lextutor (Cobb, n.d.), an online platform for the analysis of texts and words, were consulted.

Finally, we drew on the results from the previous 16 semesters about the distribution of proficiency levels of beginning students. This gave us a good idea as to which percentage of test takers would be below B2 and which would be in the B2, C1, and C2 ranges.

11.5 Test Validation

Having used the old test for a number of years and having found the results of the test to be consistent and sound in determining which students are at the required ability level, it was clear that this test should be used to establish the new test's concurrent validity by way of correlation studies. For a detailed study of the research parameters, see Sweeney-Novak (2012).

Originally, there were three sets of data to work with. The 189 students in the trial had taken the old test before the start of the semester and had subsequently taken both versions of the new test in class. It was therefore possible to correlate the two new test versions with an external measurement instrument, namely the old test. Scatterplots showed a positive relationship between the three variables.

As a further step to see to which extent the two sets of data correlated, *Pearson product-moment correlation coefficient* (*r*) was calculated. Data with significant discrepancies were taken out of the data set: for example, isolated cases with a considerable difference between the trial and old test and cases with high numbers of items missing. In these instances, it was not clear whether students had arrived late to do the trial test, whether they had not taken the trialling process seriously enough, or whether test security of the old test had been compromised. The correlation indices showed a strong positive correlation between the old test and trial Versions 1 and 2 (.794 and .761 respectively), as well as between trial Version 1 and trial Version 2 (.809). A correlation index of .809 shows a clear relationship between the two trial versions, although an *r* in the high .80s or .90s, according to Hatch and Lazaraton (1991, pp. 440–444), would be desirable.

Initially, statistical information was acquired by means of Classical Test Theory. Subsequently, data was analysed using Item Response Theory (IRT) for additional information about test items (difficulty) and test takers (ability), which can be placed on one common linear scale, in order to establish the item difficulty hierarchy for each version and to ensure equivalence of the two versions. IRT is a powerful statistical tool which is used to make informed claims about a test's overall quality, about item and person characteristics, and about their relationship. IRT models are based on formalised expectations about person and item behaviour which is not directly observable; hence, IRT models are also referred to as "latent trait" models. The IRT model used in all VELT analyses is the one-parameter Rasch model based on Winsteps (Linacre, 2019).

At the present time, there are a considerable number of test versions, which begs the question whether the results of these versions are comparable and whether the VELT has continuously been a stable measurement instrument. In fact, data for all versions compares well, and the degree of reliability is very high. By way of example, Table 11.1 presents the reliability results of five versions. The data shows the results of the two original versions (Version 1 and Version 2), of one version given to one group in the winter semester 2017 (Version 5), and of the two versions from the winter semester 2018 (Version 7 and Version 8). The data is taken from the

Table 11.1 Rasch reliability and separation

Semester	Version	Number of candidates	Person reliability	Person separation	Item reliability	Item separation
WS 2011	1	333	.89	2.86	.99	8.43
WS 2011	2	288	.89	2.79	.98	7.90
WS 2017	5	189	.90	3.03	.97	5.81
WS 2018	7	239	.90	2.97	.97	5.68
WS 2018	8	210	.89	2.85	.97	6.01

Winsteps analysis, which reports two types of reliability: person and item reliability. Person reliability is the equivalent of Cronbach's alpha. Item reliability is a specific Rasch entity indicating the "reproducibility" of item difficulties. Person separation classifies candidates: low person separation (below 2 and reliability below .8) could mean that the test does not distinguish well enough between high and low achievers. Item separation verifies item hierarchy: item separation indices should be above 3 and item reliability above .9 to confirm item hierarchy (or construct validity) of a test.

The data shows invariance between the test dates and two versions given at one test date. This kind of stability is consistent through all test administrations to date.

11.6 Conclusion

In describing the development of the VELT and the validation processes, this chapter explains the work that has produced an effective instrument to measure vocabulary and grammar knowledge of students wanting to major in English at the Department of English and American Studies at the University of Vienna. It is hoped that the work so far has provided some validity evidence to support the claim that the VELT measures grammatical and lexical knowledge at a proficiency level which would form a sound basis for foreign-language students in an academic context in which English is the medium of instruction.

References

Cobb, T. (n.d.). *Compleat lexical tutor* (Version 8.3) [Computer software]. University of Quebec. Retrieved from http://lextutor.ca/

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly, 34*(2), 213–238. https://doi.org/10.2307/3587951

English Profile. (2015). English Profile: The CEFR for English. Retrieved from www.english-profile.org

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics*. Heinle & Heinle.

Linacre, J. M. (2019). *Winsteps* (Version 4.4.7) [Computer software]. Retrieved from https://www.winsteps.com

Purpura, J. E. (2004). *Assessing grammar*. Cambridge University Press.

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

Sweeney-Novak, S. (2012). The Vienna English language test (VELT). *Views, 21*, 51–71. Retrieved from https://anglistik.univie.ac.at/fileadmin/user_upload/i_anglistik/Department/Views/Uploads/VIEWS_21_2012_.pdf