ARNE OLOFSSON

English Proficiency Testing – alla Bolognese

Introduction

For a university course, it is a commonplace that results have to be checked at the end of the teaching period. If the focus of the course is factual knowledge, testing is fairly straightforward. If, on the other hand, the focus is on practical proficiency in, say, a language, testing tends to be more complicated and time-consuming. Obvious ways of testing practical language mastery include interviews or speeches for pronunciation and fluency, and essays or translations for written proficiency. All of these are time-consuming for examiners and it is far from easy to attain objectivity in the sense that the same performance should be given the same mark regardless of who does the marking. In addition to the testing of active writing skills, it may also be of interest to test ability to judge a given text in terms of correctness or acceptability – a skill of particular value in the teaching professions. Also, the size of an examinee's vocabulary may be seen as an interesting component in an assessment of language proficiency.

Different universities and university colleges in Sweden (and elsewhere) use different methods and different test batteries for the measuring of English proficiency. It is the purpose of this article to describe some recent developments in the use of a well-established component of the testing system at the English department of the University of Gothenburg.[1]

Background

For the past four decades, one of the components in the testing of English proficiency has been a test battery known as the VOC/MCT.[2] The MCT (for Multiple Choice Test) consists of two, three or four English texts in which there are altogether 90 positions where the text branches out into alternative wordings, only one of which is unobjectionable.[3] The varieties represented are British English and General American. There is no element of translation. The following is a fairly typical MCT passage.

---

[1] Following an administrative reform in 2009, the English department is now merged with all the other languages (except the Nordic ones, including Swedish) studied at the University of Gothenburg and thus an integral part of the Department of Languages and Literatures.

[2] The test battery has its roots in a project carried out at the department in the 1960s, sponsored by the then Agency for Higher Education. The title of the project was MUP, an acronym for Mål, Undervisning och Prov (Objectives, Teaching and Testing) and it resulted in 20 written reports.

[3] For two 50-item sample tests, see Kjellmer (1967) and Wright (1979). In his article, Kjellmer also discusses the theoretical background to this kind of test.

"Some dictionaries assign a grade level to each entry and	A) atempt B) atemt C) attempt D) attemt	
to define the entry in words understood by	A) most B) the most	children. However,
the vocabulary skills of children at any grade level	A) are B) is	highly variable,
and the state of knowledge of these skills	A) are still very imperfect."[4] B) is	

The vocabulary part (VOC) consists of 120 English words, presented in six groups on the basis of their frequency. The marking is fully computerized and based on both MCT and VOC. The raw scores are transformed into scores on a 20-point scale, one for each part plus one for the combination of the two components. The level of difficulty of the VOC part is counted as stable and therefore able to function as a calibrator for variations in difficulty in the MCT part. Thus, for instance, the interval of 63–66 raw points (out of 120) for VOC is always 10 on the 20-point scale, whereas no such rule exists for the MCT.

The Gothenburg VOC reform

Up until 2008, each vocabulary item was followed by five Swedish words, only one of which was a possible translation. However, following the implementation of the "Bologna process", with an increased influx of international students, translation involving Swedish is no longer an option for testing across the board, although it is still possible and desirable in the teacher education programme. The MCT is not really affected by this, apart from the fact that the focus of contrastive problems should now in principle be less on difficulties typical of Swedish learners. The VOC, on the other hand, has had to undergo a drastic change but still within the established framework.

For stability and comparability between different generations of learners, the general format has been kept: 120 items presented in six groups according to their frequency in English text, with the underlying assumption that there is some degree of correspondence between frequency and learners' mastery of lexical

[4] It may seem over-generous to university-level testees to include such a basic feature as subject–verb concord. However, diagnostic test results show that of prospective university first-termers in English, eight out of ten get the concord right in the former of the two cases and five out of ten in the latter. Considering that five out of ten in a dual-choice task equals the guessing chance, the result is less than impressive.

items. The original VOC test was based on Thorén (1967), which was in its turn based on Thorndike & Lorge (1944).[5] The new format initially keeps Thorén as a source but the material is checked against more recent frequency counts, mainly Cobuild (1995) and, offering greater precision through lemmatization, Leech et al. (2001). Group 1 contains fairly elementary material (within the 3,000 most common words) such as *gap, iron, quiet* and *complain*, whereas Group 6 features fairly rare items such as *deluge, knack, buoyant* and *shrivel*, which are all outside the 7,000 but within the 15,000 most frequent words.[6]

What is really new in the new format is the fact that the starting-point for each vocabulary item is not a word but a definition, written in English and formulated in as simple a manner as possible, using common and general vocabulary. The definitions are inspired by but never copied verbatim from learners' dictionaries. As a couple of examples, consider the following, which have worked well in terms of expected level of difficulty and in terms of the attractiveness of the distractors.

Group 1: High-frequency words

to leave permanently	**abandon 90%**	demean 2%	divest 3%	extol 4%	grieve 1%

Group 6: Low-frequency words

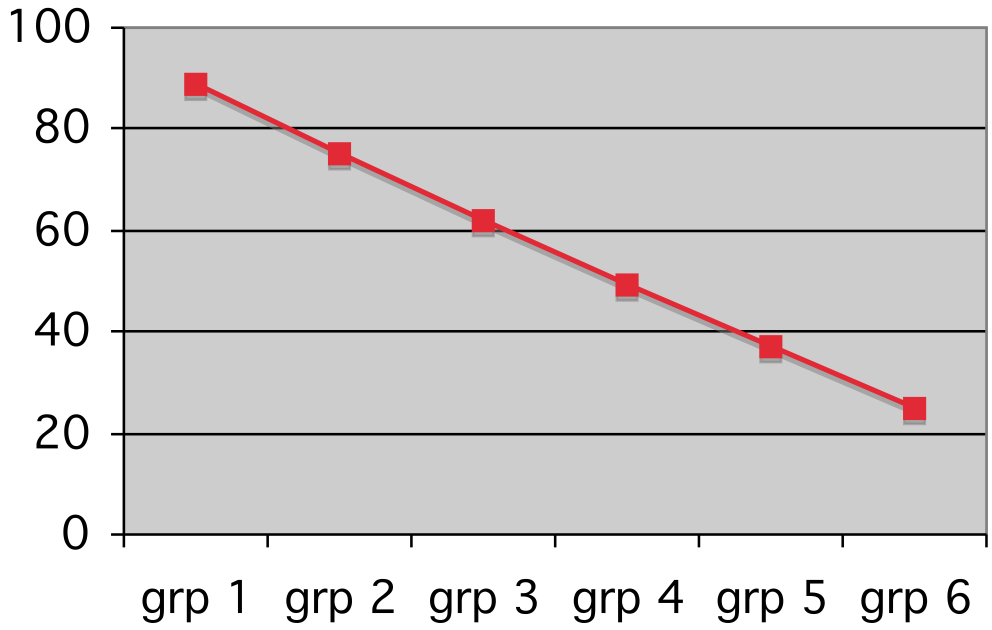
narrow strip of land that joins the coasts of two larger areas	accession 22%	**isthmus 27%**	jetty 19%	nestling 13%	tycoon 19%

The results presented above are based on an end-of-term exam with first-, second- and third-termers participating, altogether 144 students.

For the most frequent (≈ easiest) words, the aim is to get correct answers from about 90% of the participants. The corresponding aim for the least frequent (≈ most difficult) words is about 25%; it should be remembered that the guessing chance in a test item with five choices is 20%. There should be a gradual drop in the success rate through the six groups of test words, as shown in the following diagram.

[5] By "original" I mean the multiple-choice format used from about 1970 up till 2008. Even a decade before the MUP project, translation-type VOC experiments based on the first edition of Thorén's book had been carried out at the department, as described in Ellegård (1960).

[6] The latter span is based on COBUILD (1995) and refers to the frequency band with one filled diamond. (If we follow COBUILD exactly, the span is defined as between the 6,600 and the 14,700 most frequent words in the "Bank of English" corpus.)



As long as this gradient profile can be retained from one test to the next, we seem to come sufficiently close to stability in the level of difficulty and thus to reliability. The combination of (and compromise between) frequency and attested difficulty requires constant monitoring, with analyses carried out after each test round. If there is substantial discrepancy between frequency and difficulty, a test item can be moved to or towards the "right" group, but, for stability, only one step.[7]

For the first few years of development, the 120 words have been presented as four 30-item batches, each of which covers the full range of frequencies. The main reason for this arrangement is a wish to avoid fatigue; now students get a fresh start on elementary items after each 30-item batch.

The statistics provided by the computer for each test round include average student success per item and attractiveness of the distractors offered, average success per frequency group and average success for each 30-item batch, which facilitates harmonizing their levels of difficulty.

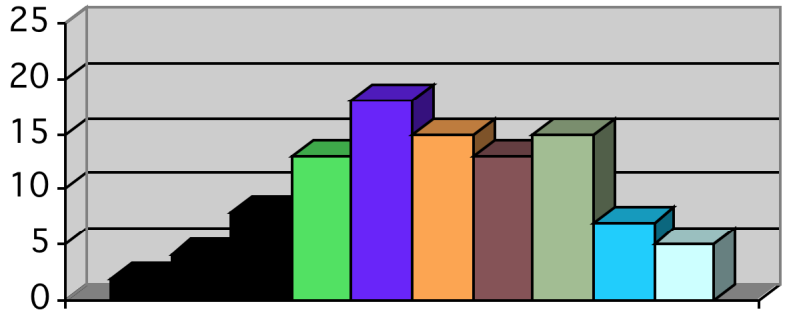
[7] A striking example of such discrepancy is the word *parrot*, which is outside the 6,000 most frequent words in English in general but is mastered by 100% of those students who have been exposed to the textbooks used in Swedish schools. (The Monty Python parrot sketch may of course have helped too.)

Results

Both the MCT and the VOC (and the combination of them) have proved to be very powerful ranking instruments, which in its turn means that, when given as beginning-of-term diagnostic tests, they provide valuable information to individual students as to their positions in their study groups in terms of practical proficiency and as to their chances of passing the course within reasonable time.[8] For illustrations, see the following diagrams, which are based on the extreme situation at the beginning of the autumn term of 2010, when not only 145 prospective students of English for general purposes or for teacher education but also 464 prospective economics students (Swedish and international) took the same test. Particularly noteworthy is the spread of results, in view of the fact that the Swedish students involved are generally assumed to have a fairly uniform educational background.

In all of the diagrams, each bar represents 2 points (1–2, 3–4, etc., from left to right) on the 20-point scale. In order to facilitate comparisons between categories, the bars show percentages of each population, not numbers of individuals. Black is used for the three leftmost bars (1–6 points) to indicate results that signal probable difficulties to meet proficiency requirements within the given time frame of the education.

Fig. 1. Diagnostic VOC results.
609 students (mixed).



[8] The ranking function was successfully used (with the old version of the test battery) at the University of Gothenburg in the mid-1990s, when proficiency ranking for admission was temporarily allowed.

Fig. 2. Diagnostic VOC results. 326 students, economics programme

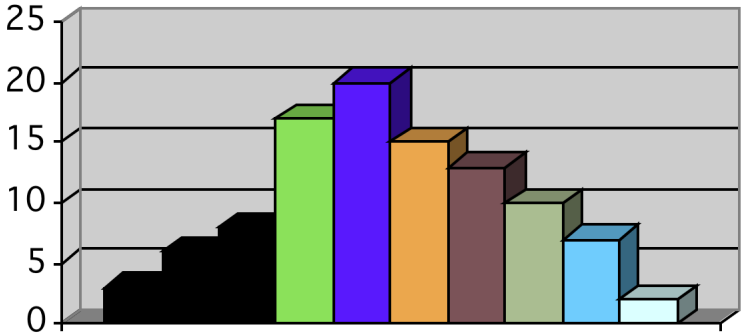


Fig. 3. Diagnostic VOC results. 138 students, Graduate School Master Programme

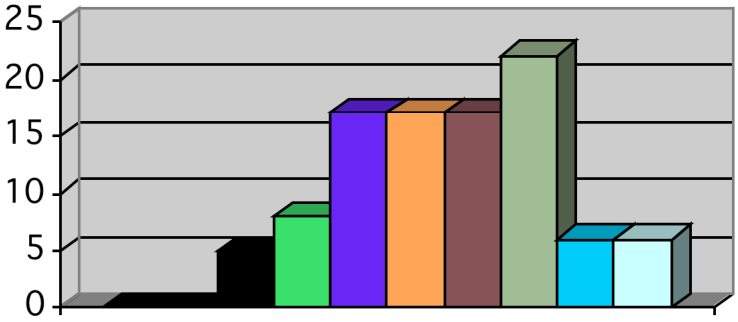
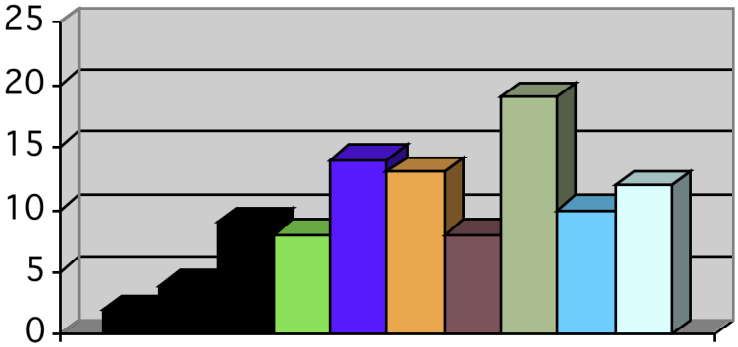
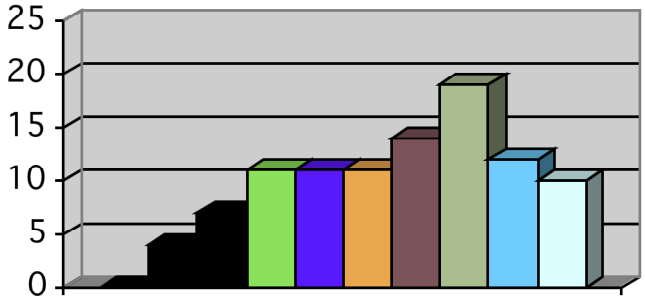


Fig. 4. Diagnostic VOC results. 145 general-purpose and teacher-education students



When given as a test, whether as a diagnosis or as an examination, the VOC is always combined with the MCT and a computer program calibrates them. As an example of results of the full test, see the following diagram, which can be seen as supplementing or superseding the one above based on VOC only for the 145 prospective general-purpose and teacher-education students.

Fig. 5. Diagnostic full VOC/MCT. 145 general-purpose and teacher-education students



Further comments on the results

For the students who had signed up for English for one term ("general-purpose" in the heading) and for the teacher-education students, the VOC test may seem to have been too easy, since so many (17 individuals, i.e. nearly 12 per cent) reached the top scores (19–20) on the 20-point scale (see Fig. 4). However, in terms of raw scores, none of those who reached 20 (11 individuals) "hit the ceiling": The top scorer reached 119 (out of the 120 items) but was alone in that position, leaving five fellow students in the 110–114 range and five from 103 (the qualification raw score for 20) to106. The spread was similar in the Graduate School group, but the top scorer stopped at 115. For the Graduate School students, it is noteworthy (but natural, in view of their academic level) that not one of them scored less than 5 (see Fig. 3).

Some methodological points

At least the trial rounds and the subsequent monitoring have to be carried out on large groups of testees; at the University of Gothenburg a first-term intake for full-time studies of English usually amounts to between 150 and 200 students belonging to different categories (programme and non-programme courses). Additional students and thus statistics are made available through cooperation between the University of Gothenburg and some of the regional university colleges. However, once the levels of difficulty have been established based on the large-scale testing, the tests can be successfully used by smaller groups as well.[9]

[9] Strictly speaking, this is relevant only for assessment in examination format. For ranking purposes, absolute levels of difficulty are not necessary.

A computer program is being developed for random selection of items within the different frequency groups. The program includes a "quarantine" function which blocks items from repeated use within a specified period of time. With this powerful selection machinery, a student who studies for all the three terms during which the VOC is used as an examination tool will never meet the same test items, even if he or she should have to re-sit the examination every term; all it takes to achieve this is a "bank" of as few as 720 items (= definitions plus choices). With this system fully in place, a teacher (or administrator) can have a unique test generated by the machinery by simply pressing a button instead of spending hours dreaming up a new test on more than one occasion per term. The underlying careful construction and monitoring of these tests are time-consuming and thus expensive activities, but in the long run the money will probably be found to have been well spent.

References

Collins COBUILD English Dictionary 1995. London: HarperCollins Publishers.

Ellegård, A. 1960. Det rätta ordförrådet? *Moderna Språk* 54: 117–126.

Kjellmer, G. 1967. Measuring Language Proficiency is Easy / Difficult / Impossible. *Moderna Språk* 61: 10–19.

Leech, G., Rayson, P., & Wilson, A. 2001. *Word Frequencies in Written and Spoken English. Based on the British National Corpus*. Harlow: Longman.

Thorén, B. 1967. *10 000 ord för tio års engelska*. Lund: Gleerups.

Thorndike, E.L. & Lorge, I. 1944. *The Teacher's Word Book of 30,000 Words*. New York: Teachers College Press.

Wright, D. 1979. Life as a Multiple Choice Test – or Multiple Choice Testing as a Way of Life. *Moderna Språk* 73: 321–327.