Article:

Lessons Learned in Designing and Implementing a Computer-Adaptive Test for English

Jack Burston* and Maro Neophytou**
Language Centre
Cyprus University of Technology

*jack.burston @ cut.ac.cy | **maro.neophytou @ cut.ac.cy

Abstract

This paper describes the lessons learned in designing and implementing a computer-adaptive test (CAT) for English. The early identification of students with weak L2 English proficiency is of critical importance in university settings that have compulsory English language course graduation requirements. The most efficient means of diagnosing the L2 English ability of incoming students is by means of a computer-based test since such evaluation can be administered quickly, automatically corrected, and the outcome known as soon as the test is completed. While the option of using a commercial CAT is available to institutions with the ability to pay substantial annual fees, or the means of passing these expenses on to their students, language instructors without these resources can only avail themselves of the advantages of CAT evaluation by creating their own tests. As is demonstrated by the E-CAT project described in this paper, this is a viable alternative even for those lacking any computer programing expertise. However, language teaching experience and testing expertise are critical to such an undertaking, which requires considerable effort and, above all, collaborative teamwork to succeed. A number of practical skills are also required. Firstly, the operation of a CAT authoring programme must be learned. Once this is done, test makers must master the art of creating a question database and assigning difficulty levels to test items. Lastly, if multimedia resources are to be exploited in a CAT, test creators need to be able to locate suitable copyright-free resources and re-edit them as needed.

Keywords: Computer-Assisted Testing, CAT, English, placement, test authoring.

1. Background

In our Language Centre, as in many European universities with an EFL course requirement, the linguistic level of incoming students can vary across the entire range of the Common European Framework of Reference for Languages (CEFRL) scale. Since all first-year students at our university have to complete a two-semester B1 level Academic English course as a graduation requirement, those who enter the university with English language proficiency below this level risk not only failing the course but also failing to obtain their degree. As there is neither time in the schedule nor funding for remedial classes, at the start of every academic year an urgent need arises to identify weak students in order to provide them with counseling and self-study guidance. To meet this need, our Centre previously carried out diagnostic evaluation

using a commercial paper and pencil test (MacMillan), in-class oral interviews and a writing assignment. Although this procedure gave satisfactory results, it was time consuming to administer and evaluate, with results not being known for at least two weeks after the start of classes. In order to improve diagnostic efficiency, we turned to computer-based testing since such evaluation can be administered more quickly, automatically corrected, and the outcome known as soon as the test is completed.

2. Computer-based test options

2.1. Non-adaptive tests

In seeking an alternative to our previous diagnostic testing procedures, one non-adaptive online option was considered: DIALANG. DIALANG attracted our attention because it evaluates a wide range of skills (reading, writing, listening, grammar and vocabulary) in English as well as more than a dozen other European languages. So, too, it is freely accessible and aligned with the CEFRL. However, since it is non-adaptive, students have to answer all questions at whatever level they self-select for testing. In a class environment this can be problematic since the test can take longer to administer than the time available in a single session. So, too, DIALANG is based on a relatively small question inventory and, being the product of a long completed EU project, lacks funding for ongoing maintenance and development. Moreover, since DIALANG does not run over the Internet (or even a local area network server), it must be individually installed on all computers. Aside from the initial complications this can entail when several labs have to be used, it also restricts flexibility should access to suitably configured labs change at the last moment. Added to these constraints, DIALANG provides no record keeping at all. At the end of a test, students are given their result, but can only write it down or, provided a printer link is available, hand in a screen print of it. For these reasons we were obliged to look elsewhere for a computer-adaptive alternative for our diagnostic testing.

2.2. Computer-adaptive test design

Computer-adaptive tests are based on Item Response Theory (Hambleton, Swaminathan & Rogers 1991). The simplest, and most frequently implemented, are constructed according to a single parameter Rasch model (Rasch, 1980), which is governed only by the difficulty of the item and the ability of the person located on the same continuum. In such a test, responses are sought to questions of pre-established difficulty level. Students who can consistently answer questions at difficulty level X are deemed to demonstrate X level proficiency. A computer-adaptive test (CAT) automatically adjusts to the proficiency level of students by presenting easier questions following incorrect responses and more difficult ones after correct answers.

By targeting questions within a range that a student can consistently answer correctly, a CAT can be administered using a relatively small number of question items. Compared to a traditional non-adaptive test, which typically might contain 75-100 questions, a CAT can usually determine a student's language proficiency level in 25 questions or less. Although any particular student may at most see only a couple of dozen test items, in order to have a sufficient number of items in reserve at various levels of difficulty, the operation of a CAT requires a question database several times this size. It also requires a computer-based algorithm to select the questions to be presented, determine the correctness of responses, and adjust the difficulty level of subsequent questions accordingly.

2.3. Computer-Adaptive Tests

2.3.1. Commercial tests

The most comprehensive, and undoubtedly best known, computer-adaptive programme for evaluating foreign language proficiency is the Brigham Young University *CAPE* (Computerized Adaptive Placement Exams). It tests grammar, vocabulary and reading comprehension and is aligned with the American Council for the Teaching of Foreign Languages (ACTFL) proficiency guidelines: novice, intermediate, advanced and superior. In its most recent iteration, known as *webCAPE*, it includes tests for six languages including L2 English. As its name implies, it is Internet-based and so can be accessed without installation on local computers. The *CAPE* series is based on a very large question database (nearly 1000 items per language) and provides statistically reliable results with detailed record keeping. However, its use comes at a cost (e.g., $1,700/year for 500 students, if paid by the University) which our Centre simply could not afford. Alternatively, the cost ($10) of taking the *CAPE* can be passed on directly to students, which in our public institution was not an option.

2.3.2. Free tests

Fortunately, two cost-free CAT creation options are available as an alternative to a commercial test: *Concerto* and *SLUPE*. Of the two, *Concerto* is by far the most flexible and powerful. Distributed by the University of Cambridge, *Concerto* is an online R-based adaptive testing platform. Being open-source, it can be fine-tuned to the evaluation of competence in virtually any domain. That being said, its implementation requires the services of a computer programmer fluent in R and someone with a solid background in statistical analysis. On the one hand, this makes it an ideal choice where such expertise is available. On the other, as in our case, it puts *Concerto* out of reach when the required technical expertise is not accessible.

Though much more limited in its capabilities than *Concerto*, *SLUPE* (Saint Louis University Placement Exam) has the great advantage of requiring no programming ability or statistical expertise of test creators. *SLUPE* is a user-friendly CAT authoring system which requires only that test makers create their own question database. It allows two types of testing format:

a) Text-based: multiple-choice questions with four options and only one correct answer.
b) Audio/video-based: a set of five 5 True/False options, 0-5 of which may be correct answers.

Questions and answers are simply entered into an online text box. Audio and video prompts can either be uploaded to the *SLUPE* website or linked to an external source (e.g., YouTube). Test makers assign a difficulty level of 1-4 (easy-hard) to each question. By default, the four difficulty levels within *SLUPE* correspond to semester divisions. However, these can be associated with whatever proficiency scale test authors choose. Once questions have been added to the database, *SLUPE* takes care of everything else. Like *Concerto*, *SLUPE* is web-based and so requires no local computer installation. Each test is associated with a specific URL which instructors give to students along with a log-in id and password. The CAT algorithm underlying *SLUPE* automatically handles question presentation based on difficulty levels and keeps detailed records of student responses: the questions they attempted, whether they were answered correctly or not, and their final placement level. It also tracks results organized by test item responses, thus allowing subsequent statistical analysis of actual question difficulty levels. For language teachers like ourselves, with minimal technical and/or financial support, *SLUPE* was an obvious choice when starting out to create a CAT.

3. The E-CAT

3.1. Test creation

While *SLUPE* enormously simplifies the technological and computational aspects of CAT creation, the quality of placement obtained with it very much depends upon the teaching experience and testing expertise of would-be test makers.

3.2. Theoretical considerations

As with any test, construct validity (Cronbach & Meehl, 1955) arguably must be the primary consideration, i.e., does the test actually assess what it claims to evaluate? In the case of our test, dubbed the E-CAT, its intended purpose was to assess the general L2 English proficiency of first-year university students. In particular, it sought to identify the weakest students, those below A2 (CEFRL), in order to provide them with appropriate counseling and self-study guidance.

Attaining construct validity is challenging for any CAT used for language proficiency assessment, all the more so when aligned with the CEFRL. By definition, CEFRL criteria are all performance-based, i.e., they describe what students are able to do with the language in given situations. On the other hand, by design, all computer-adaptive tests are based on fixed answer responses (e.g., multiple-choice questions), which most easily targets grammar and vocabulary knowledge. Typically, listening and reading comprehension are the only performance-related language skills tested in a CAT. As a consequence, the construct validity of any CAT-based assessment of language proficiency depends critically upon the content validity of the grammar and vocabulary that is tested, i.e., the degree to which their mastery is representative of a given proficiency level. In the case of the CEFRL, content validity equates to the mastery of those elements of grammar and vocabulary that allow defined language functions to be successfully performed. While listening and reading comprehension tasks allow receptive language skills to be tested, it is also possible to assess more active skills by using prompts (text as well as audio) to solicit communicatively appropriate responses. For example:

> Audio Prompt - *They live on a shoe string nowadays.*
> (Possible text-based responses, 0-5 of which may be correct)
>
> - Yes, they have it pretty easy.
> - Yes, they have little money.
> - They should buy sandals.
> - They are just stringing you along.
> - They are frugal, they'll get by.

3.3. Practical considerations

Owing to their fixed nature, *SLUPE* questions are subject to two notable constraints. Firstly, while audio-video-based listening comprehension testing is easily accommodated through the use of multiple true-false questions, reading comprehension tasks cannot be effectively exploited. Text-based prompts can only be associated with a single multiple-choice question, i.e., one text passage cannot serve as the basis for multiple comprehension questions. It could easily take a student a couple of minutes to read a passage of any substance, which is far too long to devote to a single question. Secondly, while question prompts may be in written, oral or video form, only text-based responses are supported. As a consequence, *SLUPE* cannot be used to present audio-based communicatively appropriate responses (see 3.2 above).

Although the creation of text-based questions is very straightforward, the exploitation of audio and video resources as question prompts is considerably more demanding. Finding appropriate materials can be very time consuming and, once located, copyright

permission must be obtained for their use. Because of the complications involved in obtaining copyright permission, would-be test creators are well advised to limit their search for audio-video materials to copyright-free or creative commons sources.

Aside from general copyright permission, the exploitation of audio-video resources makes two other demands on test makers. Firstly, copyright usage must allow the material to be modified in order to extract just that portion of the audio-video file needed as a test prompt. Typically, this would be no more than 60-90 seconds from a passage that might run for five minutes or more. Secondly, the test creator must either possess the editing skills needed to modify audio-video resources or have access to technical assistance to get the job done.

In principle, *SLUPE* can operate with as few as 52 test questions:

- 10 text-based at four levels (= 40 items)
- 3 audio/video-based at 4 levels (= 12)

However, statistical reliability requires at least twice this number of test items in the database. The E-CAT was first created with 112 testing items. Subsequent to initial testing, this was increased to 144. The E-CAT test was pilot tested in April 2013 with approximately 200 students during the second semester in their compulsory first-year course. In September-October 2013 approximately 450 first year-students sat the test. Another 350 students sat the test in March-April of 2014.

3.4. Difficulty level calibration

For our purposes, in assigning question item difficulty, the *SLUPE* semester levels 1-4 were equated with CEFR A2, B1, B2 and C1. Since SLUPE places students who score above the top level in semester 5, we equated this with C2.

By definition, the proficiency level of a student taking an IRT-based CAT is equated with the difficulty level of test-items that are correctly answered. Consequently, the reliability of such placement is critically dependent upon the accuracy of the difficulty level assigned to each question. Although *SLUPE* itself allows question difficulty levels to be determined freely by whatever means test makers choose, until a question database has been administered to a reasonably large number of students, i.e., several hundred at least, there is no way of knowing with any certainty the actual difficulty of any question. This can only be determined by an *ex post facto* analysis of the relative frequency with which questions were answered correctly or incorrectly.

In principle, it is possible to create a CAT on the basis of a question database previously analyzed for difficulty level, for example one derived from an earlier paper and pencil version of a test. However, doing so assumes that differences in testing conditions (e.g., with or without the use of a computer) and student populations will not significantly affect question difficulty levels. In the absence of an existing question database of known difficulty level, as was our case, the initial assignment of item difficulty of necessity can only be done intuitively. In any event, however difficulty levels are initially determined, a CAT question database needs to be recalibrated several times based on actual responses from a representative student population before reliable placement can be assumed. Very often, especially at the early stages of CAT development, the recalibration of item difficulty level results in gaps being created in the database which have to be filled by the creation of new test items at the levels that have been vacated. The difficulty level accuracy of these additions then needs to be validated through the analysis of subsequent administrations of the CAT.

While the easiest and most difficult items in a question database are relatively easy to identify, i.e., those which the most students answer correctly or incorrectly, any detailed determination of question difficulty level can only be done by proper statistical

analysis. Even the most experienced language teachers cannot intuitively assign question difficulty levels with any high degree of accuracy. Compared to the statistical analysis of student responses, our initial estimations of question difficulty level in the E-CAT were correct less than half of the time, with considerable standard error and many discrepancies of 2-3 levels. Following the first recalibration, the statistical analysis of the second administration of the test again revealed an accuracy rate of less than 50% in question difficulty assignment, but this time with a considerably lower standard error of measurement. Moreover, 91% of the level assignments resulting from the recalibration were within +/-1 level of the statistical estimates of question difficulty. Analysis of the third iteration of the test demonstrated further improvements in test accuracy, with 72% of the difficulty settings agreeing with the statistical estimates.

3.5. Placement results

As a reference point for placement accuracy, the E-CAT results from its third pilot testing were compared against our instructors' evaluation of their students' proficiency level based on a whole semester (and in some cases an entire academic year) of class performance. Across all levels, the E-CAT agreed exactly about 40% of the time, with no more than +/-1 level divergence in another 48% of the placements. Below the A2 level, which was our primary concern, exact agreement was higher at 50% with no more than +1 level divergence in another 33% of the placements. Overall, then, in well over 80% of the cases the E-CAT successfully placed students with reasonable accuracy in less than one class period compared to instructors who had the advantage of at least an entire semester to make their judgment. As the accuracy of question difficulty levels improves through continued statistical analysis of test results, it is expected that so, too, will placement accuracy.

4. Conclusion

Based on our experience with the E-CAT, we can say with confidence that it is definitely feasible for language teachers without computer programming skills to create reliable computer-adaptive tests using the freely accessible *SLUPE* authoring programme. That being said, the process is neither quick nor effortless. Above all, it requires collaborative teamwork to succeed, which in our case involved five experienced language teachers. Initial test construction, learning how to use the *SLUPE* system and even more so building an operational question database, can be expected to take a whole semester. If multimedia resources are to be effectively exploited, test creators need to be able to locate suitable copyright-free resources and re-edit them as needed. Undoubtedly, the most challenging and critical aspect of question creation is the proper assignment of difficulty level. As our experience demonstrates, on their own, even the most experienced language teachers are unlikely to get this right more than half the time. Since by definition the reliability of any CAT-based student placement is directly determined by the accuracy of question difficulty assignments, access to *ex post facto* statistical analysis of item difficulty levels is essential. At least two pilot testing sessions, typically spread over two semesters and involving several hundred students, are required to evaluate placement results and adjust the question database accordingly.

For those fortunate enough to have the financial resources to pay the recurrent fees for the use of a commercial language test such as *webCAPE*, constructing a CAT may very well appear to be too demanding a task. On the other hand, making a virtue of necessity, once a locally developed CAT is operational it has one great advantage over any commercial test. Having been calibrated against the local student population for which it is intended, the difficulty level of its test items is much more closely matched to the proficiency of its test takers, with correspondingly greater placement accuracy. In cases where the native language of students being assessed is quite different from that

typically used to calibrate a commercial CAT, e.g., L1 Greek, Chinese, or Arabic speakers learning L2 English, this can make a significant difference.

References

Cronbach, L. J.; Meehl, P.E. (1955). Construct Validity in Psychological Tests. *Psychological Bulletin* 52: 281–302.

Hambleton, R., Swaminathan, H., & Rogers, J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Lawshe, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology,* 28, 563–575.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests.* Copenhagen: Danmarks Paedagogiske Institut, 1960. Reprint, Chicago: University of Chicago Press.