Developing and Implementing a Computer-Adaptive Test for English: The SLUPE Experience

Jack Burston (jack.burston@cut.ac.cy)
Cyprus University of Technology, Cyprus
Maro Neophytou (maro.neophytou@cut.ac.cy)
Cyprus University of Technology, Cyprus
Iasonas Lamprianou (lamprianou.iasonas@ucy.ac.cy)
University of Cyprus, Cyprus

Abstract

When the granting of a degree depends upon the successful completion of an EFL course requirement, it is critically important to identify weak students as soon as possible in order to provide them with counselling and self-study guidance. To improve diagnostic efficiency, we adopted the freely available SLUPE (Saint Louis University Placement Exam) authoring system to create a computer-adaptive test for English (E-CAT). Following initial pilot testing with nearly 200 students during the Spring of 2013, the E-CAT was administered to a full cohort of some 450 first-year students during the Fall semester of 2013 and again to 350 students in the Spring of 2014. The placement results of the E-CAT were compared to student self-evaluations and instructor ratings. Of the three, E-CAT placements correlated most closely with instructor ratings. Given these satisfactory results, we now base our diagnostic testing entirely upon the E-CAT. We can, thus, confirm that the SLUPE platform will provide anyone wanting to create a computer-adaptive test with the means of doing so. Notwithstanding, the actual preparation of questions is time consuming and requires several iterations based on the statistical analysis of student responses in order to objectively determine question difficulty levels.

Keywords: CAT, Computer-Adapted Testing, English, EFL, placement

Introduction

Regardless of the procedures used to gain entry to a college or university (e.g., high school examinations, transcripts, etc.), incoming students inevitably display a wide range of foreign language proficiency. Determining the linguistic competence of incoming first-year students is thus a major challenge for any institution of higher learning with foreign language requirements. Most frequently, such information is needed in order to place students at the right level or exempt them from certain courses. It also happens, as it does in our University, that students have to take compulsory language courses for which there is no option of differential placement or exemption. In this case, it is critically important to identify the weakest students, who risk not only failing the course but, since such courses are a graduation requirement, also failing to obtain their degree.

High school records and general entrance exams rarely provide reliable measures of foreign language competence, obliging language teaching units to undertake their own assessments. This was the situation in our Language Centre where diagnostic evaluation was carried out using a commercial exam (MacMillan *Straightforward* Quick Placement & Diagnostic Test), 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 the academic year.

Computer-Adaptive Testing

As an alternative to the assessment methods we had been using, we turned to computer-adaptive testing. A computer-adaptive test (CAT) is based on Item Response Theory (IRT). IRT presupposes that proficiency in a given domain can be measured by the difficulty level of questions that can be correctly answered. Simply put, a test taker who can consistently answer questions at difficulty level X is deemed to demonstrate X level proficiency. Computer technology makes it possible to design algorithms that dynamically adapt the difficulty level of test questions to the real-time performance of the test taker (Embretson & Reise 2000; Hambleton, Swaminathan & Rogers 1991; van der Linden & Glas 2010; Wainer 1990). A CAT automatically adjusts to the proficiency level of students by presenting easier questions following incorrect responses and more difficult ones after correct answers. Ultimately, the algorithm zeroes in on the difficulty level at which correct responses are most consistent.

A CAT offers a number of advantages over traditional testing methods (Chang 2015; Chang, & Ying 2007; Chen, Lee & Chen 2005; Dandonoli 1989; Larson 1998; Wang, et al. 2012; Zheng & Chang 2015). Firstly, since students are only presented with items adapted to their level, fewer questions have to be asked and testing can be completed more quickly. Compared to a traditional non-adaptive test, which typically might contain 75-100 questions, a CAT can usually determine a student's proficiency level in 30 questions or less. Because assessments are individually adjusted for each student, test security is better since virtually no one sees the same question set. Computer-adaptive tests are automatically corrected and the results are known as soon as the test is completed. So, too, owing to the statistical precision with which question difficulty levels can be established, results can be determined with much greater accuracy than is possible with conventional tests.

Despite their considerable advantages, the use of CATs has given rise to some concerns (Burston & Monville-Burston 1995; Chao, Kuo & Tsai 2015; Canale 1986; Kaya-Carton, Carton & Dandonoli 1991; Meunier 1994; Tung 1986). From a theoretical perspective, the validity of CAT results has been called into question because, by definition, tests based on IRT assume that what is being measured is an underlying unidimensional latent trait. It goes without saying that language proficiency encompasses reading, writing, listening and speaking skills which draw upon multiple competencies: grammar, vocabulary, phonology, cultural knowledge, pragmatic and communicative abilities. In order to keep these variables to a minimum, language CATS frequently target a single

competency, such as grammar (Burston & Monville-Burston 1995), reading (Chalhoub-Deville, Alcaya & Lozier 1997; Dandonoli 1989; Kaya-Carton, Carton, & Dandonoli 1991; Keng, Kam & Wong 2010) or listening comprehension (Dunkel 1999; Olea et al 2011). Even within such restricted testing domains, however, it is not really possible to limit what is assessed to a single competency. Reading and listening comprehension, for example, necessarily entail vocabulary and grammar knowledge. Moreover, according to theorists such as McNamara (1991), objections raised against the construct validity of CATs on the basis of unidimensionality are ill-founded and stem from a failure to distinguish two types of model: a measurement model and a model of the various skills and abilities potentially underlying test performance. Language tests based on IRT represent a measurement model which posits that, owing to an underlying general linguistic ability, it is possible to meaningfully sum scores on different parts of a test. IRT makes no direct pronouncements about the different language skills that test results may entail. It simply assumes a single dimension relating ability and difficulty.

Theoretical considerations aside, CATs are also subject to important practical constraints which stem from the need to operate with questions of known difficulty level. This is only possible with fixed answer response types. As a consequence, all CATs are restricted to some variant of a multiple-choice format (e.g., true-false, completion, cloze item selection), whence the focus on vocabulary, grammar, reading and listening comprehension. By its design, a CAT cannot assess productive speaking or writing ability. Likewise, the closest it can come to evaluating communicative competence is by presenting students with linguistic prompts to which they have to select contextually appropriate responses.

While they do not give a complete profile of student language competency, CATs can provide an efficient and reliable means of streaming students into broad proficiency levels, whence their attractiveness for placement testing and the reason for our particular interest in them.

SLUPE

Having neither the resources to fund the recurrent administration of a commercial CAT, nor the ability to pass the costs on to students, we were obliged to create one ourselves. The construction of any CAT requires two essential components. The first is a database of questions of known difficulty levels. The second is a computer-based programme to handle the logistics of question presentation and calculation of student ability level. Our Language Centre had the experienced teachers needed to generate the content of a CAT, but lacked the programming expertise to create the presentation software. Fortunately, we were able to use the freely available web-based SLUPE (Saint Louis University Placement Exam) CAT authoring system. SLUPE requires test makers to create their own question database, which then becomes available to other instructors who use the system. Questions and answers are simply entered into online text boxes. SLUPE employs 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. Audio and video prompts can either be uploaded to the SLUPE website or linked to an external source (e.g., YouTube).

Once the student placement level is determined by the text-based multiple-choice questions, the audio/video-based questions are presented at that level to provide a point of comparison with listening comprehension proficiency.

Test makers assign a difficulty level of 1-4 (easy-hard) to each question and SLUPE takes care of everything else. By default, the four difficulty levels within SLUPE correspond to semester divisions. However, these can be associated with whatever proficiency scale test authors choose.

The Study

Our principle aim in testing students at the beginning of the academic year is to identify those with the weakest English language proficiency in order to provide them with counselling and self-study guidance in courses at the B1 level (CEFR). From previous experience, we knew that about 30% of our first-year students would be below this level, with half of these below A2. Our practical concern, then, was to determine whether experienced language teachers without computer programming expertise could construct an effective CAT using the free SLUPE authoring system. More specifically, we sought to determine whether we could reliably stream incoming students into English proficiency levels that clearly identify those with a competency level below B1.

The E-CAT

Multiple-choice text-based questions in the E-CAT, as our test is called, assess grammar and vocabulary. As well they also test appropriate communicative responses to written dialogue prompts. For example,

Text Prompt
- When did she leave?
- _____ .
(Possible responses)

A. In a minute
B. For half an hour
C. Just before lunch
D. Until noon

Audio/video-based items consist of sets of true-false listening comprehension questions as well as audio dialogue prompts which elicit communicatively appropriate responses. For example:

Audio Prompt
- Hello, I'd like to speak to Mr Jones, please.
- _____
(Possible responses)
[T/F] Sorry, can you say that again?
[T/F] I'm sorry, I'll call again later.
[T/F] I'm afraid I don't know.
[T/F] Who is calling?
[T/F] One moment, please.

For our purposes, the (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.

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.

Methodology

Students taking the E-CAT were all in their first-year at the university and came from all of its six faculties (Geotechnical Sciences and Environmental Management, Management and Economics, Communication and Media Studies, Health Sciences, Fine and Applied Arts, Engineering and Technology). About 60% of the cohort was female and 40% male. The average age of the females was between 18-19 years old. Owing to compulsory military service, the average age of the males was two years older. Greek was the native language of virtually all students and all had previously studied English as a second language for at least nine years.

Pilot Testing

Initial Implementation Difficulties

The E-CAT was the first pilot tested in April 2013 with approximately 200 students during the second semester in their compulsory first-year course. Two problems which resulted in data loss were encountered during initial pilot testing, though neither of these were caused by any malfunction of the E-CAT itself. The first problem was caused by incorrectly configured computers which could not play audio-video files. This caused the test to terminate before a placement score could be determined. In response, steps were taken to insure that all computers in all labs were identically configured. The SLUPE

developer also changed its operation to save student placement data before the presentation of audio/video-based questions. The second difficulty was brought about by the failure of students to follow instructions properly. By design, in a CAT it is not possible to back up to review (much less change) responses to previous questions. Despite warnings not to use the web browser backspace key during the test, some students persisted in doing so. This resulted in the test aborting and loss of placement information. To alleviate the problem, the SLUPE developer altered the programme to intercept the back-up command and display a warning message that the test would abort unless the key press were cancelled.

Pilot Testing Results

The results that we obtained from the pilot testing of the E-CAT fall into two categories: those relating to student placement and those relating to question difficulty levels.

The results of the E-CAT placement for the Spring 2013 semester are shown in Figure 1. Roughly a quarter of the students were placed at level 1 and a quarter at level 2, 41% at level 3 and the remaining 11% above that level.

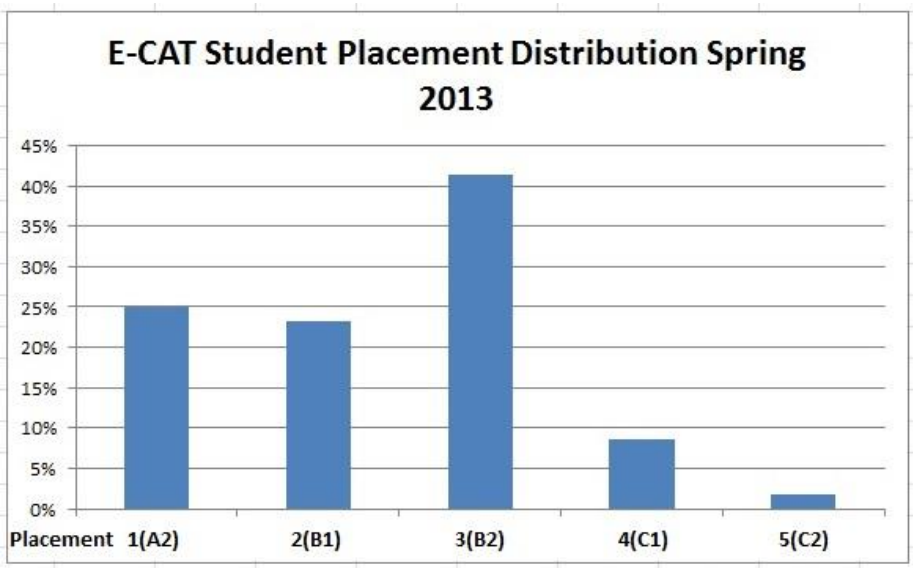

Figure 1. Student Placement Spring 2013

In order to obtain an external reference point for the E-CAT placement results, we asked instructors to rate their students on the same 1-4 scale (A2-C1) as used in the test. Compared to our instructors' evaluation of their students, the E-CAT agreed in 48% of the cases (= 0 discrepancy), underrated them by one or two CEFR levels a quarter of the time and overrated them by between one to four CEFR levels a quarter of the time (Figure 2).

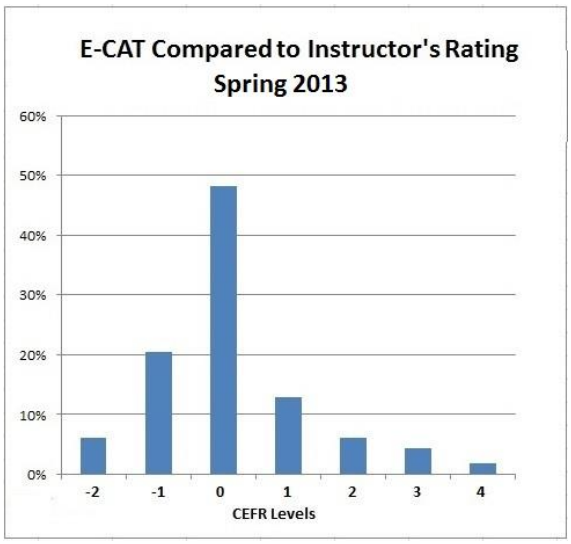

Figure 2. Spring 2013 CAT-Instructor Ratings

Based on the statistical analysis of student responses, our initial estimations of question difficulty level in the E-CAT were only correct about half of the time. While this might appear to be disappointingly low, it in fact was not unexpected since the initial determination of question difficulty could only be made through intuitive subjective estimations. Also, with only 200 students taking the test the first time, the number of responses upon which the statistical analysis was based allowed for some reservations about the accuracy of the results. The reassignment of question difficulty levels created gaps within certain of the question types which we filled through the creation of about two dozen new items.

Full-Scale E-CAT Results

In September-October 2013 approximately 450 first year-students sat the E-CAT. The test was taken again by 350 of the same students in March-April of 2014. Testing took place in University computer labs during normal class times and was completed on average within 20 questions in less than forty minutes.

Fall 2013

The statistical analysis of question difficulty levels resulting from the second iteration of the E-CAT confirmed that the adjustments made had considerably improved the accuracy of the test's settings. While, overall, correct question difficulty assignments of the test remained below 50%, nearly 91% of the E-CAT question difficulties were now within +/- 1 level of the statistical analysis.

The results of the E-CAT placement for the Fall 2013 semester are shown in Figure 3. As comparison with Figure 1 shows, the overall distribution results are very similar to those

of the Spring 2013 semester: roughly a quarter of the students at level 1 and at level 2, 36% at level 3 and the remaining 14% above that.

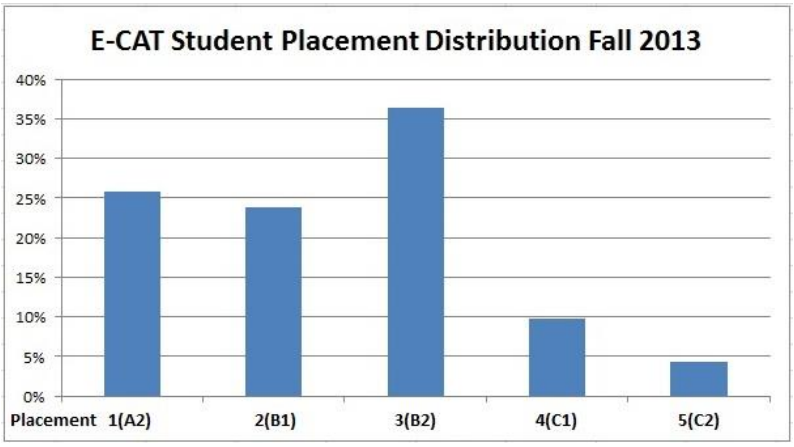

Figure 3. Student Placement Fall 2013

In preparing the statistical analysis for the second iteration of the E-CAT, we incorporated three external reference points for the test's student placement results, all based on the CEFR scale. Firstly, students took a written test (MacMillan *Straightforward* Quick Placement & Diagnostic test). When taking the E-CAT this time, students themselves were also asked to self-rate their English proficiency level. Lastly, instructors complemented test results with their evaluation of student proficiency level.

In comparing student placement ratings of the E-CAT and the three external evaluations, it is important to bear in mind the different basis of each. The E-CAT tested vocabulary, grammar, and communicative appropriateness. The MacMillan test was based on vocabulary and grammar. Instructor assessments were based on students' course performance over a couple of weeks. Student self-assessments were purely subjective. It thus cannot be expected that there would be a 100% correlation between the four evaluations. The results of the comparison of E-CAT placement results are summarized below.

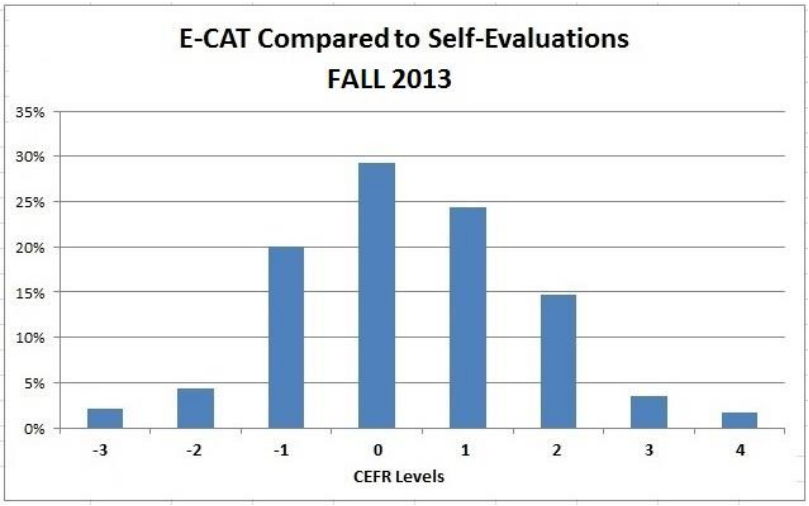

Figure 4. E-CAT / Student Self-Evaluation Fall 2013

As can be seen in Figure 4, there is great variation between the placement level of the E-CAT and students' own assessment of their language proficiency. The two agree less than 30% of the time (= 0 divergence) and the discrepancies range from -3 to +4 CEFR levels. This of course raises the question of which of the two evaluations is the most reliable. Given the much higher correspondence of the E-CAT placements with the MacMillan test and instructors' evaluations (Figure 5-6), the most reasonable conclusion is that students are not very good judges of their own English language proficiency.

Compared to the MacMillan test (Figure 5), the E-CAT results correspond in nearly 40% of the cases, with the range of discrepancy exceeding one CEFR level 13% of the time.

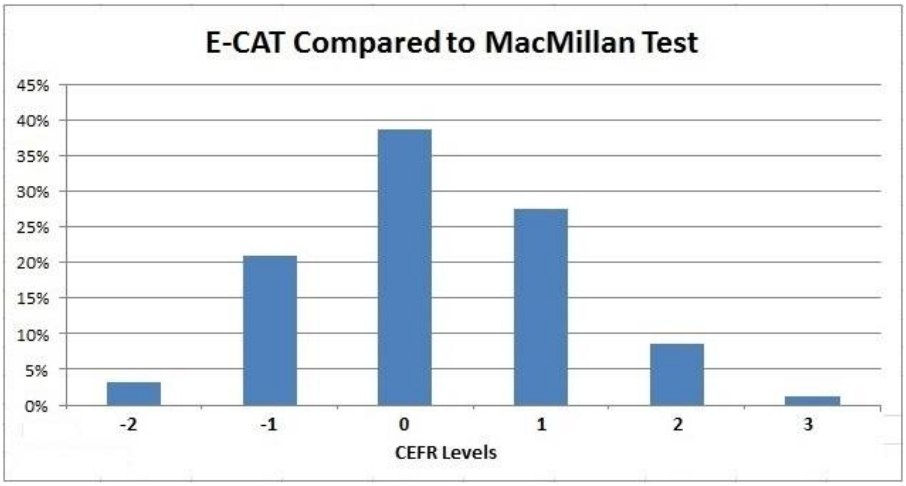

Figure 5. E-CAT / MacMillan Test

As regards the performance of the E-CAT relative to instructor ratings (Figure 6), it is to be noted that the degree of correspondence with the revised version of the test compared to the original version (Figure 2) improved from 48% to 53%, which is considerably greater than with the MacMillan test. So, too, the degree of divergence was markedly reduced and did not exceed 2 CEFR levels. Within this range, only 9% of the cases exceeded one CEFR level.

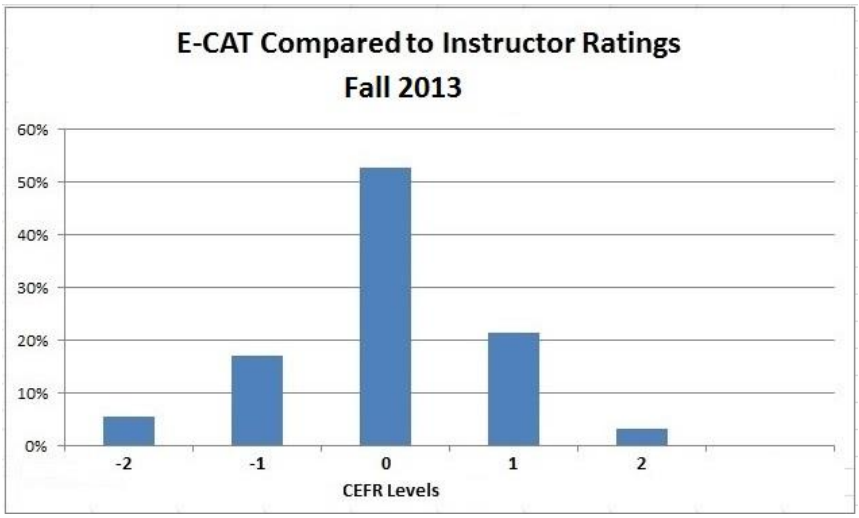

Figure 6. E-CAT / Instructor Ratings Fall 2013

Lastly as can be seen in Figure 7, when the overall placement results of all four evaluations are correlated with each other, the E-CAT, MacMillan test and instructor ratings all show very similar correlations, as indicated by the colour and size of the circles in the graph. Student self-evaluation, on the other hand, demonstrates a very weak correlation.

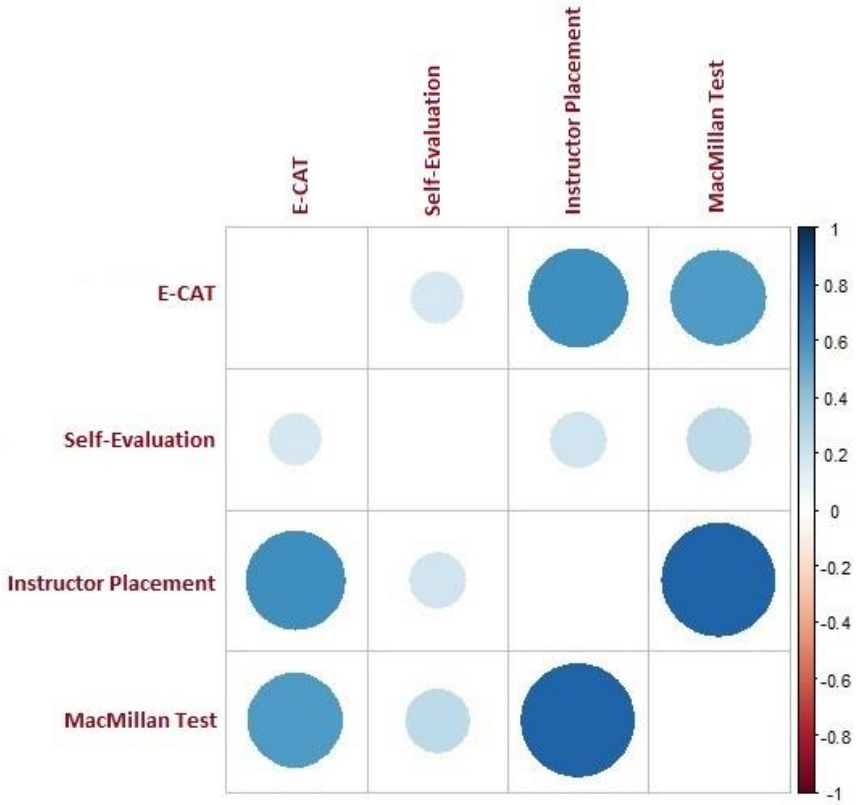

Figure 7. Overall Test Correlations Fall 2013

Spring 2014

The E-CAT was administered a third time during the Spring 2014 semester. Aside from actual placement, this iteration of the test served two purposes. As before, it provided additional results to further improve the statistical accuracy of question difficulty levels. Secondly, since it involved the same cohort of first-year students as in the Fall and was administered at the end of the second semester, it allowed measurement of proficiency gains over the academic year.

The analysis of question difficulty levels resulting from the third administration of the E-CAT demonstrated continuing improvement of the accuracy of the test's settings. Most notably, 72% of the question difficulty assignments of the test now accorded with the statistical estimates. At the easiest level of difficulty, over 95% of the settings corresponded precisely with the statistical analysis. So, too, overall nearly 95% of the E-CAT question difficulties were within +/- 1 level of the statistical findings.

As can be seen in Figure 8, compared to the Fall 2013 results (Figure 3), there was a very substantial increase in the measured proficiency level of students. At 13%, the proportion of those at the lowest level was just half what it had been at the beginning of the year. Those at the second level increased from 24% to 36%. Likewise, the proportion of students at level 4 jumped from 10% to 18%. This upward shift at levels 2 and 4 resulted in the overall percentage of students at level 3 to drop from 36% to 29%.

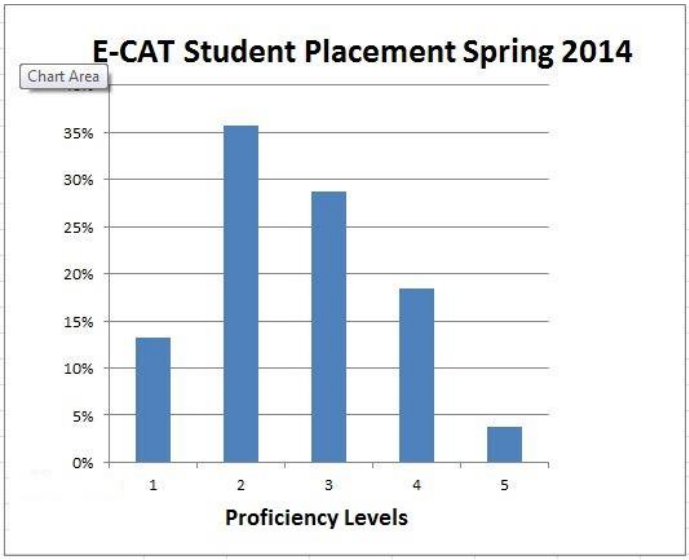

Figure 8. Student Placement Spring 2014

In preparing the statistical analysis for the third iteration of the E-CAT, it being impractical to re-administer the MacMillan test, we restricted comparison of its placement results to two external reference points: student-self evaluations and instructor ratings. Given the weak correlation that had been demonstrated between student self-evaluations and the CEFR-based ratings of the other assessment measures in the Fall 2013 results, the question arose as to whether this could have been caused by the inability

of students to accurately rate themselves in terms of the presumed CEFR descriptors. To simplify matters, this time students were asked to rate their English proficiency on the following four point scale: 1=fair, 2= good, 3= strong, 4= very strong. Instructors again rated their students, this time also using the same simplified 4 point scale.

Compared to the Fall 2013 results (Figure 4), placements based on a simplified scale resulted in a higher rate of agreement (= 0 divergence) between the E-CAT and student self-evaluations (Figure 9), which increased from 29% to 35% in Spring 2014. Again, however, about 45% of students either under-estimated or over-estimated their proficiency by one level compared to the results of the E-CAT. The percentage of students under-estimating or over-estimating their proficiency by two or more levels declined from about 26% to 18%. While the use of a simpler rating scale very likely contributed to the increased accuracy of student self-evaluations, so, too, it is reasonable to assume that a year's experience in the course also gave students a more realistic assessment of their proficiency level. That being said, it is clear that student self-evaluations are no substitute for formal placement testing.

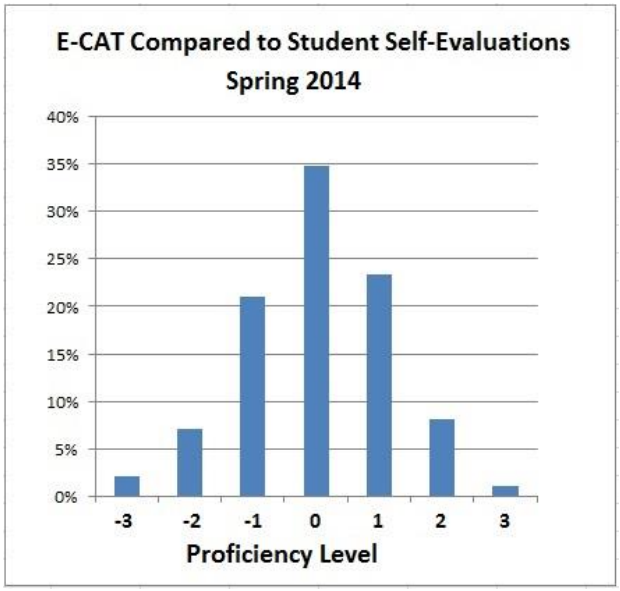

Figure 9. E-CAT / Student Self-Evaluation Spring 2014

In comparison with instructor evaluations, E-CAT placements at the end of the academic year (Figure 10) varied considerably more than at the beginning of the year (Figure 6).

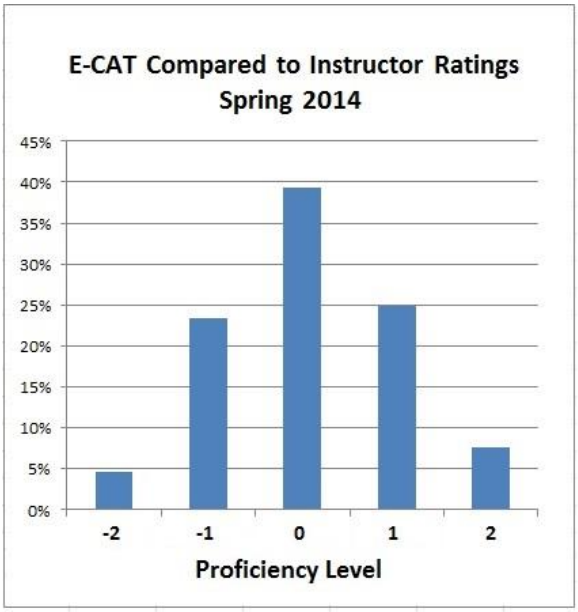

Figure 10. E-CAT / Instructor Ratings Spring 2014

Two factors very likely contributed to these results. Firstly, during the Fall 2013 semester instructors had access to E-CAT placements when rating their students, so may have been inclined to align their ratings with the test results. Such influence was not possible during the Spring 2014 semester since all instructor ratings were collected prior to the re-administration of the E-CAT. Secondly, whereas instructors had only had contact with their students for a couple of weeks when rating them at the beginning of the year, during the Spring 2014 semester they had had them in class for at least 10 weeks, and in some cases since Fall 2013. Which is to say, they knew their students much better when evaluating their proficiency level at the end of the year.

Discussion

On the basis of our experience over nearly two years, we can confirm firstly that the SLUPE platform will provide anyone wanting to create a computer-adaptive test with the means of doing so. It does not presuppose of its creators any formal background in testing, it is easy to learn to use, requires little time or effort to produce an operational test, and, of course, is free. Notwithstanding, the actual preparation of questions is time consuming and requires several iterations based on the statistical analysis of student responses in order to objectively determine question difficulty levels. While SLUPE does not make excessive technological demands, care does need to be taken to insure that operational configurations in a lab are the same for all computers. Likewise administering the test is very straightforward – provided students follow instructions and in particular do not attempt to return to previous questions.

Our initial results with student placement were satisfactory and, as a result of the adjustments we have made to question difficulty levels, improved with the second and

third versions of the E-CAT. In comparing E-CAT placement levels with student self-evaluation, the *MacMillan* test and instructor ratings, we have discovered that by far the weakest correlation is with self-assessments.

Conclusion

The E-CAT has provided us with an acceptable, cost and time effective, replacement for the commercial paper and pencil test we were using. The E-CAT is no match for instructor ratings based on a semester or more of contact with students. However, at the beginning of the academic year when such rating expertise is simply not available, it does a good job identifying the weakest students, which is our priority. Moreover, because the E-CAT is easy and quick to administer, it allows us to measure end-of-year progress as well. The E-CAT is an ongoing project and we have good reason to expect further improvements in student placement accuracy as the statistical analysis of question difficulty levels becomes more precise thanks to an ever growing database of student responses. Finally, we are confident enough with the current results to now base our diagnostic testing entirely upon the E-CAT.

Limitation of the Study

Firstly, it must be remembered that the accuracy of student placement in an IRT-based CAT is directly determined by the accuracy of question difficulty levels, a factor which is very much affected by the native language of test takers. In order to be used with speakers of a different L1, the entire question database of the E-CAT would have to be recalibrated. As we have demonstrated, it is possible to evaluate a student's ability to identify contextually appropriate responses to textual and oral prompts. However, because of the need to operate with fixed responses to questions of known difficulty, the E-CAT cannot evaluate productive language usage. For the foreseeable future at least, writing and speaking proficiency can only be evaluated through manually corrected tests. Within the context of what a CAT can evaluate, one area for future improvement would be for the test to augment purely numerical level placements with qualitative information about specific areas which are easiest or hardest for students. This is precisely where the most recent work in CD (Cognitive Diagnosis) CAT studies is focused. The SLUPE CAT authoring platform itself has recently taken a step in this direction by allowing test designers to tag the content domain of questions, from which student area proficiency profiles can be generated. This is an enhancement which the next iteration of the E-CAT will definitely be investigating.

References

Burston, J., & Monville-Burston, M. (1995). Practical design and implementation considerations of a computer adaptive foreign language test: The Monash/

Melbourne French CAT. *CALICO Journal* 13(1), 26-46. Retrieved from http://www.equinoxpub.com/journals/index.php/CALICO/article/view/23404/19409

Canale, M. (1986). The promise and threat of computerized adaptive assessment of reading comprehension. In C. Stansfield (Ed.), *Technology and language testing* (pp. 29-45).Washington, DC: TESOL.

Chalhoub-Deville, M., Alcaya, C., & Lozier, V. (1997). Language and measurement Issues in developing computer-adaptive tests of reading ability: The Minnesota approach. In A. Huhta, V. Kohonen, L. Lurki-Suonio, & S. Luoma (Eds.), *Current developments and alternatives in language assessment* (pp. 546-585). Jyvaskyla University.

Chang, H-H. (2015). Psychometrics behind computerized adaptive testing. *Psychometrika*, 80(1), 1-20.

Chang, H-H. & Ying, Z. (2007). Computerized-adaptive testing. In N. Salkind (Ed.), *The sage encyclopedia of measurement and statistics* (pp. 170-174). Thousand Oaks, CA: Sage.

Chao, R-C., Kuo, B-C. & Tsai, Y-H. (2015). Development of Chinese adaptive test system based on higher-order item response theory. *International Journal of Innovative Computing, Information and Control*, 11(1), 57-76.

Chen, C-M., Lee, H-M., & Chen, Y-H. (2005). Personalized e-learning system using Item Response Theory. *Computers & Education*, 44(3), 237-255.

Dandonoli, P. (1989). The ACTFL Computerized Adaptive Test of Foreign Language Reading Proficiency. In W. F. Smith (Ed.), *Modern technology in foreign language education: application and projects* (pp. 291-300). Lincolnwood, IL: National Textbook.

Dunkel, P. (1999). Research and development of a computer-adaptive test of listening comprehension in the less-commonly taught language Hausa. In M. Chaloub-Deville (Ed.), *Issues in computer-adaptive testing of second language reading proficiency* (pp. 91-121). New York: Cambridge University Press.

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

Kaya-Carton, E., Carton, A., & Dandonoli, P. (1991). Developing a computer-adaptive test of French reading proficiency. In P. Dunkel (Ed.), *Computer-assisted language learning and testing: Research issues and practice* (pp. 259-284). New York: Newbury House.

Keng, L-T., Kam, H-W., & Wong, P. (2010). Computerized adaptive testing in reading comprehension. *TEFLIN Journal* 8(1), n pg.

Larson, J. (1998). An argument for computer adaptive language testing. *Multimedia-Assisted Language Learning*, 1(1), 9-24. Retrieved from http://kmjournal.bada.cc/wp-content/uploads/2013/05/1-1-1jerry.pdf

McNamara, T. (1991). The role of item response theory in language test validation. In A. Anivan (Ed.), *Current developments in language testing* (165-184). Singapore: SEAMEO Regional Language Centre.

Meunier, L. (1994). Computer Adaptive Language Tests (CALT) offer a great potential for functional testing. Yet, why don't they?. *CALICO Journal* 11(4), 23-39. Retrieved from

http://www.equinoxpub.com/journals/index.php/CALICO/article/view/23426/19431

Olea, J., Abad, F. J., Ponsoda, V., Barrada, J. R., & Aguado, D. (2011). eCAT-listening: Design and psychometric properties of a computer-adaptive test on English listening. *Psicothema* 23(4), 802-807. Retrieved from http://www.psicothema.com/pdf/3959.pdf

Tung, P. (1986). Computerized adaptive testing: Implications for language test developers. In C. Stansfield (Ed.), *Technology and language testing* (pp. 9-11). Washington, DC: Teachers of English to Speakers of Other Languages.

van der Linden, W., & C. Glas (Eds.) (2010). *Elements of adaptive testing*. New York: Springer. Retrieved from http://dx.doi.org/10.1007/978-0-387-85461-8

Wainer, H. (1990). *Computer-adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wang, H-P., Kuo, B-C., Tsai, Y-H., & Liao, C-H. (2012). A CEFR-based computerized adaptive testing system for Chinese proficiency. *TOJET*, 11(4), 1-12.

Zheng, Y., & Chang, H-H. (2015). On-the-fly assembled multistage adaptive testing. *Applied Psychological Measurement*, 39(2), 104-118.