AN EVALUATION OF A COMPUTER ADAPTIVE TEST IN A UK UNIVERSITY CONTEXT

Mariana Lilley and Trevor Barker

An Evaluation of a Computer Adaptive Test in a UK University Context

Mariana Lilley and Trevor Barker

University of Hertfordshire, F`aculty of Engineering and Information Sciences,
College Lane, Hatfield, Hertfordshire AL10 9 AB, United Kingdom
M.Lilley@herts.ac.uk and T.1.Barker@herts.ac.uk and

Abstract

The work described in this paper relates to the design, development and evaluation of software to perform a computer-adaptive test for a Visual Basic programming module at the University of Hertfordshire. The application we developed uses an adaptive algorithm based on the Three-Parameter Logistic Model from Item Response Theory. The application was designed to select the questions presented to each individual student based upon their ability, as measured by performance in the test.

The first part of the paper describes the model used in the adaptive algorithm and our design and implementation approach. In addition, it outlines earlier work by the authors in which the efficacy of this approach in the domain of English as a foreign language was shown. In this study, we were able to show that the application was of pedagogical interest to teachers and the interface did not impose any barriers to assessment. To this end, academic staff and a group of international students evaluated the prototype and compared the software with a traditional computer-based test, using evaluation techniques that ranged from heuristic evaluation to focus group.

In the second part of the paper, the results of a study of performance with 133 participants in two large-scale computer-delivered assessments are reported. We were able to show, using statistical analysis of the data obtained in the tests, that the participants were not disadvantaged by computer-adaptive testing, when compared to traditional computer-based tests and other forms of assessment, such as coursework and practical exams. In order to evaluate the students' attitude to testing, two debriefing sessions were held following the tests and the main findings from these sessions are also reported here.

The benefits and potential limitations of this method of assessment are presented in the final section of this paper.

Keywords: Computer-Assisted Assessment, Computer-Adaptive Test, Item Response Theory, Evaluation methodologies

Introduction

The ability to configure a computer interface according to the individual user is an important goal for many software developers. This is especially true in complex educational software systems. However, it is often difficult to tell whether or not a user has achieved his or her objectives when using such complex systems. User objectives are often individual, complex and difficult to measure and their satisfaction may well lie outside of the immediate software domain.

In order to overcome some of these difficulties, student models are sometimes employed to individualise the presentation of the interface (Brusilovsky, 1996). Barker and colleagues described an attempt to configure a multimedia educational system based on a cooperative psychological student model (Barker *et al.*, 2002). In this work, information and the level of tasks and questions presented to users were adapted cooperatively based upon their ability. A major limitation of their approach was the difficulty in establishing the appropriate level of the user model based on performance. This task required input cooperation from both tutor and student in order to establish the level of difficulty. Despite its pedagogical value, this method was inefficient in time and resources. Feedback to the adaptive student model was therefore slow and often for this reason it was difficult to be certain that the model had been configured optimally.

In the work described here, a mathematical model was used to adapt the presentation of questions to users automatically. The application described in this research was a computer-assisted assessment that was used to test 133 second year Computer Science students at the University of Hertfordshire. It is hoped that the model developed and evaluated here will be used in the future as part of a fuller psychological student model in order to adapt the presentation of information to users, based upon an understanding of their ability and requirements as they learn. In the following sections, the development of our computer-adaptive testing prototype is described.

Computer Adaptive Tests

Computer-assisted assessments are increasingly being viewed as a feasible alternative to traditional paper-based assessment methods (O'Reilly and Morgan, 1999; Conole and Bull, 2002). In this paper we concentrate on two subsets of computer-assisted assessments: traditional computer-based tests (CBTs) and computer-adaptive tests (CATs).

A traditional CBT mimics a paper-and-pencil test in that a predefined set of questions is presented to all participants. The main benefits of the use of such an electronic tool over a traditional paper-based test range from the potential for a more precise and faster marking process to the possibility of timely feedback. In a CAT, the qualities of speed and accuracy of marking plus the opportunity to provide the students with immediate feedback remain unchanged. CATs differ mainly from traditional CBTs in the way that the questions are selected (Wainer, 1990).

Whilst in a CBT the set of questions presented is usually predefined regardless of students' performance, in a CAT the questions administered are dynamically selected and depend on student performance during the test. This is due to the fact that, in a CAT, a correct response is typically followed by a more difficult question. Conversely, an incorrect response results in an easier question being presented next. A mathematical model is often used in order to select the most appropriate level of difficulty of the next question for each individual student.

CATs therefore, have the potential to offer higher levels of individualisation when compared to conventional CBTs. Given that in a CAT the responses provided by each individual student dictate the level of difficulty of the following questions, the set of questions presented during a given session of assessment is tailored for each individual student.

The main benefit of this approach is that students are presented with questions that are challenging and motivating, rather than questions that are either too difficult and therefore frustrating, or too easy and thus uninteresting (Wainer, 1990). Moreover, questions that are too difficult or too easy tell us little or nothing about a student. Only those questions exactly at the boundary of the student's knowledge tell us anything about the level of a student's ability. In a CAT, it is intended that all questions presented are at this level of difficulty.

The mathematical model used in the prototype developed was based on the Three-Parameter Logistic Model (3-PL) from Item Response Theory (IRT), which is explained in the next section.

Item Response Theory

The adaptive algorithms used in CATs are based on Item Response Theory (IRT). The central element of IRT is a family of mathematical functions that aims to calculate the probability of a specific student answering a particular question correctly. At present there are various IRT mathematical models for dichotomously scored questions (i.e. items), such as the One-Parameter Logistic Model (1-PL), the Two-Parameter Logistic Model (2-PL) and the Three-Parameter Logistic Model (3-PL) (van der Linden, 1997).

The 1-PL Model makes use of the mathematical function shown in Equation 1 to evaluate the probability *P* of a student with an unknown ability θ answering a question of difficulty b correctly (Wainer, 1990).

$$P(\theta) = \frac{1}{1 + e^{-(\theta - b)}}$$

Equation 1: Probability P(θ) of a student answering a given question correctly (1-PL)

The mathematical function used within the 2-PL Model to evaluate *P(θ)* is shown in Equation 2 (Wainer, 1990).

$$P(\theta) = \frac{1}{1 + e^{-1.7a(\theta - b)}}$$

Equation 2: Probability P(θ) of a student answering a given question correctly (2-PL)

In the 2-PL Model, the meaning of the parameter b remains the same as in the 1-PL Model. The additional parameter a represents the question's discrimination and its usefulness when distinguishing among students near an ability level θ (Hambleton, 1991).

The mathematical function provided by the 3-PL Model to evaluate *P(θ)* is shown in Equation 3 (Lord, 1980).

$$P(\theta) = c + \frac{1 - c}{1 + e^{-1.7a(\theta - b)}}$$

Equation 3: Probability P(θ) of a student answering a given question correctly (3-PL)

The meaning of the parameters b and a shown in Equation 3 is identical to the 2-PL Model. The parameter c, which is known as pseudo-chance or guessing parameter, represents the chance of a student answering a question correctly by guessing.

Ward (1980) suggests that a potential limitation of objective questions is that a student with no knowledge could possibly answer a question correctly by guessing. As the CAT prototype was based on the use of objective questions, it was important to ensure that the probability of a student answering a question correctly by chance was taken into account. It was also deemed necessary to consider the discrimination level of all questions responded. Hence the 3-PL model, rather than the 1-PL or 2-PL, was selected for the prototype introduced here.

Both CATs and CBTs are typically based on the idea of using a questions database. Questions are expected to cover topics from a given subject widely, and at a range of difficulty. Databases used for CBTs are the same as those used for CATs, except that in the latter each question stored in the database has the parameters b, a and c associated with it.

In a typical CAT, as each student answers a question, his or her response is evaluated as being either correct or incorrect and then a value for the probability *P(θ)* is estimated. If the response to a question is evaluated as incorrect, the function *Q(θ) = 1-P(θ)* is used instead. Increasingly larger sets of responses permit the estimate of θ to be refined further. This is achieved by multiplying the probability functions P1(θ), P2(θ), Q3(θ) and so on to get the composite probability (or likelihood) function for the set of all responses. Given that the peak of the likelihood function corresponds to the most likely value of θ for a given set of responses, the most recent value of ability estimate θ is taken to be the peak value of this composite function. If a given student's ability estimate θ is the same as the level of difficulty of a question, that student has a 50% chance of answering that question correctly (Wainer, 1990). Thus the question to be administered next is determined by the estimated θ for each individual stage, as the student is supplied with a question from the question bank for which the difficulty *b* is the closest value to the provisional ability θ.

The process of displaying questions, evaluating the responses and selecting the next question to be administered based on the student's previous responses is repeated until a stop condition has been met. The most common stop conditions are (*i*) a certain number of questions have been administered, (*ii*) a time limit has been reached and (*iii*) a certain standard error for the ability θ has been met.

Prototype development

In this section we describe the development and initial evaluation of the CAT prototype. In previous work, it was important to ensure that the prototype was usable and academically useful. To this end, an expert evaluation was performed on the prototype, which comprised a Graphical User Interface and a 250 objective questions database in the domain of English as a foreign language. A panel of eleven experts, comprising university lecturers in Computer Science and English evaluated both usability and effectiveness aspects of the prototype. This work is described in full by Lilley and Barker (2002).

The experts performed a heuristic evaluation (Molich and Nielsen, 1990) to assess the prototype's usability. In this evaluation, each of the eleven experts independently assessed elements of the interface design. These interface elements were then rated according to ten usability principles, using a Likert scale (1-5, with 5 being best). All interface elements evaluated obtained mean scores ranging from 3.9 to 4.5, and these results suggested that the interface did not present any major usability problems.

The experts also evaluated the prototype's effectiveness as a computer-assisted assessment tool in an academic context. The findings of this evaluation supported the view that the prototype could lead to a reduction in time spent marking assessments. Furthermore, it suggested that the prototype would be beneficial in terms of efficiency of marking and error reduction. The prototype was evaluated as being easy to use and to learn in addition to being potentially capable of offering higher levels of interaction than offered by a traditional CBT.

Experts also expressed the opinion that the prototype's ability to help students to detect their own potential educational needs in both summative and formative assessment environments was poor. This was due to the fact that the students were unaware of the adaptive process and therefore possibly unable to learn from their mistakes. The evaluators considered that the students would be more receptive to using a CAT in a formative rather than in a summative assessment environment. These results suggest that lecturers foresee problems regarding the scoring method used within a CAT. Candidates answering the same number of questions correctly would almost certainly have different final scores, and this could bring uncertainties about the "fairness" of the assessment.

To further assess the interface's usability and evaluate student attitude to the prototype, a modified version of the prototype was subjected to a user evaluation study (Lilley *et al.*, 2002). This evaluation involved 27 students whose first language was not English. In this study, each student – without any prior training on how to operate the application - was given 30 minutes to answer 20 questions on the use of English language and grammar. A trained observer in Human-Computer Interaction (HCI) monitored the student use of the software during the test, and the results from this observation corroborate the findings from the heuristic evaluation in that the prototype was easy to use and learn and thus unlikely to have a negative effect on student performance.

Fifteen out of the 27 students who participated in this study answered 10 adaptive questions (i.e. CAT) followed by 10 non-adaptive ones (i.e. CBT). The remaining 12 students answered 10 CBT questions followed by 10 CAT ones. In both scenarios, students were unaware of the question order. Following the usability test, 12 students from the original group were randomly selected to take part in a focus group session.

Findings from the focus group suggest that, despite students in general preferring the CAT set of questions to the CBT one, some students were concerned about the fact that they were unable to return to previous questions once they had submitted their responses. Furthermore, some students expressed concern that stopping conditions based on consistent standard error scores might mean that some students have tests of different lengths. The participants in the focus group feared that this characteristic might disadvantage students who initially perform badly and provide no opportunity for them to catch up.

The second part of the study aimed to observe how students used the software over longer periods of time. In this study, 7 volunteers were randomly selected from the original group. The number of questions presented was increased to 40 and the test duration to 60 minutes (Lilley *et al.*, 2002). In this part of the study, we were able to investigate how standard error varied over the duration of the interaction with the software. It was found that standard error on student performance did achieve constant levels within a reasonable time, and could indeed form the basis for a stopping condition for the test. No major usability problems were uncovered in this study.

In the next section we describe a study of performance with 133 participants in two computer-assisted assessments in the domain of Computer Science, namely assessment 1 and assessment 2. In this study, a modified version of the prototype introduced here was used. The results obtained by these students in two other assessments are also included.

Method

A group of 133 students enrolled in a second year Visual Basic programming module of the Higher National Diploma programme in Computer Science at the University of Hertfordshire participated in two sessions of assessments using the CAT application we developed. The application used for these assessments was an enhanced version of the high-fidelity prototype described earlier. To facilitate better formatting of questions, the prototype's interface was modified to support additional graphical elements. A new database containing 119 questions covering the module syllabus was created and independently calibrated by subject experts. The database size was deemed satisfactory by the experts. This view is supported by Carlson (1996), who suggests that a database containing approximately 100 questions is usually appropriate, provided that the questions cover the entire difficulty range. Questions for both adaptive and non-adaptive components of assessments 1 and 2 were drawn from this database.

The first assessment took place in week 7, and comprised 10 predefined questions (i.e. CBT) followed by 10 questions dynamically selected (i.e. CAT). The second assessment took place during week 10 and comprised 10 predefined questions followed by 20 dynamically selected ones. Prior to the first session of assessment using the application, students were given a brief introduction to the use of the software. However, students were unaware of the full purpose of the study and of the CAT component of the tests until after both assessments had been completed. In both sessions of assessment, the order in which the CBT questions were presented was randomly selected, as an attempt to minimise unauthorised collaboration amongst students. Participants took the tests as part of their regular assessment and the fact that each assessment comprised a non-adaptive element followed by an adaptive one, was not only a useful addition for comparative purposes, but it also helped to ensure that the test was fair and that no student would be disadvantaged by taking part in the study. Both assessments 1 and 2 were conducted under supervision in computer laboratories.

Assessment 3 took place on week 12. In this assessment, each individual student was expected to spend 25 hours over a period of 4 weeks to write a program that would offer similar functionality to a word editor. The practical exam took place on week 18. In this assessment, each individual student had to create a Visual Basic application based on a set of specifications provided on the day, within a time limit of two hours. Like assessments 1 and 2, assessment 4 was conducted under supervised conditions.

Results

Table 1 shows the scores obtained by participants in the four assessments undertaken. In Tables 1 and 2, "Level CAT1" and "Level CAT2" respectively represent the mean value of the estimated ability (θ) for assessments 1 and 2. The estimated ability ranged from –2 to +2, with intervals of 0.01. The "%CBT1" and "%CBT2" scores correspond to the amount of correct responses in CBT1 and CBT2. Similarly, the "% CAT1" and "% CAT2" scores denote the amount of correct responses in the adaptive element of the first and second assessments. As for "%A3" and "%A4" columns, these represent the scores obtained in assessments 3 and 4 respectively.

Level CAT 1 Mean	Level CAT 2 Mean	Score %CBT 1 Mean	Score %CAT 1 Mean	Score %CBT 2 Mean	Score %CAT 2 Mean	Score %A3 Mean	Score %A4 Mean
-0.83	**-0.91**	**51.5%**	**59.8%**	**42.3%**	**53.0%**	**71.7%**	**49.7%**

Table 1: Mean scores obtained in the four assessments undertaken (N=133)

A Pearson's Product Moment correlation was also performed on the data summarised in Table 1. The results of this correlation are shown in Table 2.

Variable	**Level CAT 2**	**%CBT 1**	**%CAT 1**	**%CBT 2**	**%CAT 2**	**%A3**	**%A4**
Level CAT 1 Pearson's R Sig	0.633 p<0.001	0.832 p<0.001	0.498 p<0.001	0.537 p<0.001	0.566 p<0.001	0.369 p<0.001	0.541 p<0.001
Level CAT 2 Pearson's R Sig	*	0.541 p<0.001	0.329 p<0.001	0.800 p<0.001	0.696 p<0.001	0.399 p<0.001	0.559 p<0.001
%CBT 1 Pearson's R Sig	*	*	0.329 p<0.001	0.467 p<0.001	0.499 p<0.001	0.300 p<0.001	0.445 p<0.001
%CAT 1 Pearson's R Sig	*	*	*	0.379 p<0.001	0.428 p<0.001	0.339 p<0.001	0.467 p<0.001
%CBT 2 Pearson's R Sig	*	*	*	*	0.595 p<0.001	0.398 p<0.001	0.527 p<0.001
%CAT 2 Pearson's R Sig	*	*	*	*	*	0.309 p<0.001	0.537 p<0.001
%A3 Pearson's R Sig	*	*	*	*	*	*	0.528 p<0.001

Table 2: Pearson's Product Moment correlations between the scores and levels for participants in the four assessments undertaken (N=133)

It can be seen that there is a high correlation between scores in the CBT and CAT sections of both assessments. Participants who performed well in CBT sections also performed well in CAT sections and vice versa ($p<0.001$). It was also found that the CAT levels achieved by students in assessment 1 were highly correlated to the CAT levels in assessment 2. This was taken to indicate that the test was a fair reflection of user ability in the assessment and the CAT assessment was at least as good and probably a better indicator of student ability as the CBT section of the prototype. The high correlation between the levels achieved in assessments 1 and 2 and the scores obtained in assessments 3 and 4 were also taken to support this view.

No major usability or other problems were identified during the study. The attitudes of students to the prototype expressed during the debriefing session were generally positive. No student reported that they had felt disadvantaged by the CAT approach used in the study. Some students reported that would have liked the opportunity to return to questions after they had entered their responses into the application.

Discussion and future work

This paper presents the latest stage in the development of our research at the University of Hertfordshire (Lilley and Barker, 2002; Lilley et al., 2002), in which the increased use of computer-assisted assessments in Higher Education and the potential benefits and limitations of the CAT approach were investigated. Large amounts of resources are invested in the development and use of educational software. There is also a great investment on the part of the learners to undertake computer-based courses and assessments, and thus full evaluation of such software is a vital issue. It is essential that tutors and learners have confidence in computer-based learning and assessment tools, and thorough evaluation is one way of achieving this. The evaluation approach used here was based on the work of Barker and Barker (2002) and Laurillard (1993), in which it is suggested that the evaluation of educational software is complex and to be relevant should be undertaken in a real educational context and involve all stakeholders, including students, academic staff and the academic institution. In addition, a variety of quantitative and qualitative evaluation methods, ranging from online questionnaires to focus groups may be applied to help ensure a complete view of learner attitude and performance. In the research reported in this paper and earlier work by the authors (Lilley and Barker, 2002; Lilley et al., 2002), a fairly large range of evaluation methods have been employed, involving focus groups, interviews, expert evaluation, online data logging, questionnaires and test score comparisons. A problem with such complex evaluation strategies relates to the interpretation and generalisation of the results. Barker and colleagues (2002) describes the development of an evaluation framework and how this was used to provide a context for their evaluation of educational multimedia software. We are planning to extend these ideas in future work to describe a framework intended to be useful in the context of the research reported here. It is hoped that this will allow us to gain deeper insight into the effectiveness of CATs in Higher Education and to make more comprehensive generalisations from our findings.

In earlier work by the authors it was suggested that CBTs are typically constructed on the premise that a broad range of abilities must be considered, covering the full range of learner abilities being assessed. A limitation of this approach however is that more able students are required to answer questions that are below their level of ability, often before they are presented with questions that are challenging. In the same vein, less able students start answering questions that are likely to be appropriate for their level of ability and then, at a later stage, are presented with questions that are far above their level of ability and might contribute to student anxiety and frustration. In both cases, questions that are not appropriate for the level of ability of a particular student provide no valuable information about this student. We argue that only those questions tailored for the student provide challenge to the learner and useful information to the teacher. Findings from this study suggest that CATs would not only address this issue by administering questions that are appropriate for each student level of ability, but also provide a more individualised and interactive assessment method.

The results presented here suggest that CATs have the potential to offer a more consistent and accurate measurement of student's abilities than the one offered by traditional CBTs. In our study, the statistical analysis of our data suggested that CATs are a fair measure of ability, producing higher test-retest correlations than either CBT or assessments performed away from the computer. The greater consistency and accuracy of a CAT however must be balanced against the effort required to develop a CAT. This represents a potential limitation of this assessment method, as the underlying algorithm and database required by a CAT is more difficult to be implemented than that required by a traditional CBT. It is our belief however that CATs have the potential to overcome these limitations and play a role of increasing importance in the use of computer-assisted assessments in UK Higher Education. Thus we are currently engaged in developing the work presented here further. The user interface has already been modified to support question formats beyond text-only. At present, images are already being supported, and we aim to develop the user interface further to be able to support a wider range of multimedia contents. We are also engaged in modifying this application to allow students to change previously entered answers, a concern for some learners in this evaluation. It will be important at all stages to ensure that learners, teachers and academic institutions are involved fully in the evaluation of our work.

References

Barker, T. and Barker, J. (2002). *The evaluation of complex, intelligent, interactive, individualised human-computer interfaces: What do we mean by reliability and validity?* Proceedings of the European Learning Styles Information Network Conference, University of Ghent, June 2002.

Barker, T., Jones, S., Britton, C. and Messer, D. (2002). *The use of a co-operative student model of learner characteristics to configure a multimedia application*. User Modelling and User Adapted Interaction 12 (2/3), pp 207-241.

Brusilovsky, P. (1996). *Methods and techniques in adaptive hypermedia*. User Modelling and User-Adapted Interaction 6 (2-3), pp 87-129.

Carlson, R. D. (1994). *Computer adaptive testing: A shift in the evaluation paradigm*. Journal of Educational Technology Systems, 22(3), pp 213-224.

Conole, G. and Bull, J. (2002). *Pebbles in the Pond: Evaluation of the CAA Centre*. Proceedings of the 6th Computer-Assisted Assessment Conference, Loughborough, pp 63-73.

Hambleton, R. K. (1991). *Fundamentals of Item Response Theory*. California: Sage Publications Inc.

Laurillard, D. (1993). *Rethinking university teaching: A framework for the effective use of educational technology*. London: Routledge.

Lilley, M. and Barker, T. (2002). *The Development and Evaluation of a Computer-Adaptive Testing Application for English Language*. Proceedings of the 6th Computer-Assisted Assessment Conference, Loughborough, pp 169-184.

Lilley, M., Barker, T., Bennett, S. and Britton, C. (2002). *How computers can adapt to knowledge: A comparison of computer-based and computer-adaptive testing*. Proceedings of the International Conference on Information and Communication Technologies in Education (ICTE 2002), Badajoz, Spain, pp 704-708.

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems*. New Jersey: Lawrence Erlbaum Associates.

Molich, R. and Nielsen, J. (1990). *Improving a human-computer dialogue*. Communications of the ACM 33(3), 338-348.

O'Reilly, M. & Morgan, C. (1999). *Online Assessment: creating communities and opportunities* in Brown, S., Race, P. and Bull, J. Computer-Assisted Assessment in Higher Education. London: Kogan Page.

van der Linden, W. J. (1997). *Handbook of Modern Item Response Theory*. NewYork: Springer-Verlag.

Wainer, H. (1990). *Computerized Adaptive Testing (A Primer)*. New Jersey: Lawrence Erlbaum Associates.

Ward, C. (1980). *Preparing and Using Objective Questions.* Cheltenham: Stanley Thornes.