THE DEVELOPMENT AND EVALUATION OF A COMPUTER-ADAPTIVE TESTING APPLICATION FOR ENGLISH LANGUAGE

Mariana Lilley and Trevor Barker

The Development and Evaluation of a Computer-Adaptive Testing Application for English Language

Mariana Lilley
Trevor Barker

Department of Computer Science
Faculty of Engineering and Information Sciences
University of Hertfordshire
Hatfield
Hertfordshire
AL10 9AB

t.1.barker@herts.ac.uk

Abstract

This paper reports on research undertaken at the University of Hertfordshire into the development and initial expert evaluation of a computer-adaptive testing programme based on Item Response Theory (IRT). The paper explains how the Three-Parameter Logistic model was implemented in the prototype. The underlying theory and assumptions of the model used in its development are also explained, along with the limitations and benefits of the computer-adaptive test (CAT) approach compared to traditional computer-based test (CBT) methods. In this paper use of the prototype as an alternative to the current method used by the University is evaluated by experts, and summaries of their reports and recommendations are presented. This paper also describes plans for developing this work further, including its use in computer-based student modelling where an accurate estimation of performance within a subject domain can be used to inform and adapt the choice of presentation of learning materials. Considerations for extending the CAT model to encompass other types of questions rather than multiple-choice or multiple-response questions are also presented.

Introduction

The use of computer-based tests (CBTs) has increased significantly over the last few years, and there are a number of reasons why this trend is found in Higher Education (Harvey and Mogey, 1999). These reasons vary from the possibility of marking large numbers of assessments accurately and quickly to the reduction of the time necessary to prepare these assessments by storing and reusing questions. In addition to these factors associated with higher efficiency, another benefit of CBTs would be to bring the assessment environment closer to the learning environment. Software tools and web-based sources are frequently used to support the learning process, so it seems reasonable to use similar computer-based technologies in the assessment process.

In this context, a traditional CBT is a computer-based test that mimics a traditional "paper-and-pencil" test. In a traditional "paper-and-pencil" assessment, the examiner selects a set of questions and hence the level of difficulty before the assessment session. As a result, all the examinees receive the same set of questions during a given session of an assessment. Alternatively, examiners can prepare assessments in which the examinees can select a set of questions to be answered from a larger set. For example, the examinee can choose 10 questions from a pool of 15 questions. However, in both situations, the examiner is responsible for selecting the questions that make up the pool.

Like the "paper-and-pencil" format, in a traditional CBT the questions presented to the examinees during a given session of assessment are usually the same for all examinees. Alternatively, some CBTs are designed to randomly select questions from a questions bank.

In all the cases described above, the questions presented to the examinees are not tailored according to their individual performance during the test. As a result, examinees can be presented with questions that are either too easy or too difficult.

An Overview of Computer-Adaptive Testing

Computer-adaptive tests (CATs) differ mainly from traditional CBTs in the way that the questions are selected (Wainer, 1990). In a CAT the questions presented to the examinees are dynamically selected, and depend on the examinee's individual performance during the test. If the examinee answers the question correctly, a more difficult question is presented next. Conversely, if the examinee answers the question incorrectly, an easier question is presented next.

Since the questions in a CAT are selected in an interactive way, a CAT would be able to mimic both aspects of a traditional "paper-and-pencil" test and aspects of an oral interview (Freedle, 1997). In addition, it has been suggested that CATs could positively contribute to the examinee's motivation during the assessment session (Wainer, 1990). During a given session of assessment, examinees might lose interest if the questions presented are too easy or might feel frustrated if the questions presented are too difficult. As a

result, in an ideal situation where the examinee's motivation is to be maintained, the questions presented should have an appropriate level of difficulty tailored for each individual examinee.

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, in generic terms, calculates the probability of a specific examinee answering a particular item (question) correctly. When using IRT, questions are usually referred to as items and this is the terminology that will be used in this paper from now on. At present IRT offers more than one different mathematical model to estimate the examinee's ability. The best known models for items with dichotomously scored responses are the One- Parameter Logistic Model, the Two- Parameter Logistic Model and the Three- Parameter Logistic Model (Van der Linden, 1997).

The Logistic Model used here

The prototype presented here uses principles of the Three-Parameter Logistic Model (3-PL) from IRT to rate an examinee's ability θ based on the examinee's own responses.

As its name implies, the 3-PL Model makes use of three parameters. Firstly, the parameter b, which represents the item's difficulty. Secondly, the parameter a, which represents the item's discrimination, or in other words, represents the degree to which a given item response varies according to the ability level (Lord, 1980). Finally, the parameter c, which is known as pseudo-chance or guessing parameter, represents the chance of an examinee answering an item correctly by guessing. The fact that the 3-PL Model accommodates the possibility of an examinee answering an item correctly by chance was the main reason why this model was chosen over the One and Two- Parameter Logistic Models.

The three parameters of the 3-PL Model are used in the mathematical function shown in Equation 1 to evaluate the probability P of an examinee with an unknown ability θ answering an item correctly (Lord, 1980).

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

Equation 1: The Three-Parameter Logistic Model

The values of the parameters b, a and c can be estimated using a calibration process which is beyond the scope of this paper. There should always be sufficient items in the pool in order to test examinees for all possible values of θ. For the purposes of this prototype, it was decided to restrict the range of values of θ from -2 to +2. This allowed the fundamentals of IRT to be demonstrated, and limits the number of items that must be generated to support the theory. The parameter b (item difficulty), must therefore range between the minimum and maximum values of ability θ, such that $-2 \leq b \leq 2$. The parameter a (item discrimination) is typically a positive number, such that $0 < a < 2$. When $a > 1$, it indicates that the given item has a higher discrimination. Finally, usual values for the parameter c (pseudo-chance) are positive

numbers, such that $0 \leq c \leq 1$. For example, a well-designed multiple-choice item with five options would typically have the parameter $c=0.2$, since the examinee has one in five chance of answering the item correctly by guessing.

By applying the formula shown in Equation 1, it is possible to plot an Item Characteristic Curve (ICC) for any given item. When using the 3-PL model, each item in the items bank (pool of questions) would have an Item Characteristic Curve (ICC) associated with it. A typical ICC would look similar to the curve illustrated in Figure 1. This curve indicates the likelihood $P(\theta)$ of an examinee answering this item correctly. All ICCs have this typical S-curve shape, but each will differ in detail according to the parameters associated with the item.

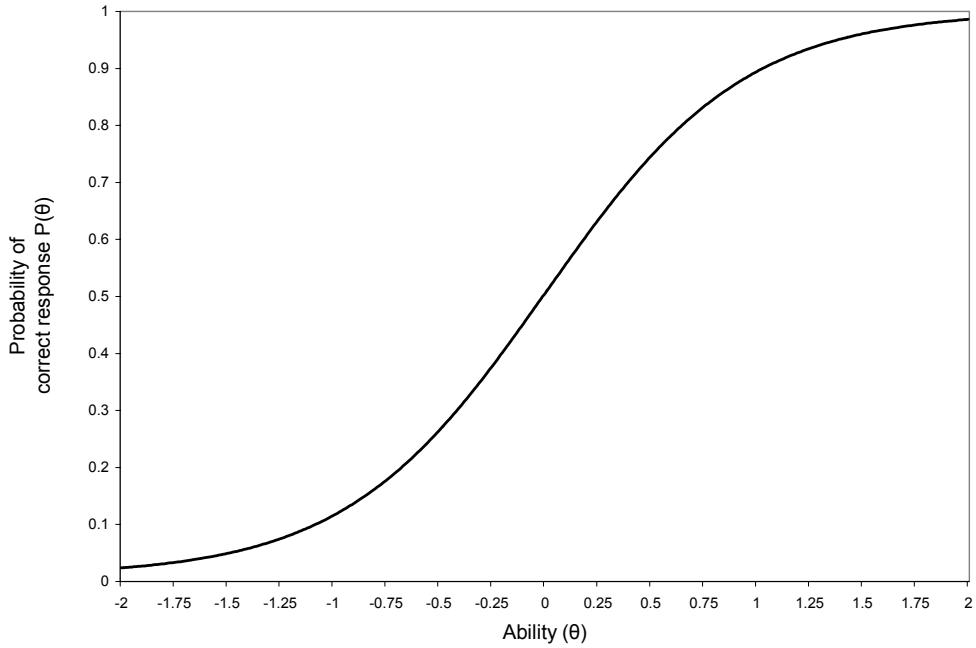

Figure 1: Example of an ICC for an item answered correctly

Given that the probability of an item been answered correctly is $P(\theta)$, the probability that this item is not answered correctly (i.e. incorrectly) $Q(\theta)$ is equal to $1-P(\theta)$. The curve shown in Figure 2 illustrates the typical curve shape of $Q(\theta)$.

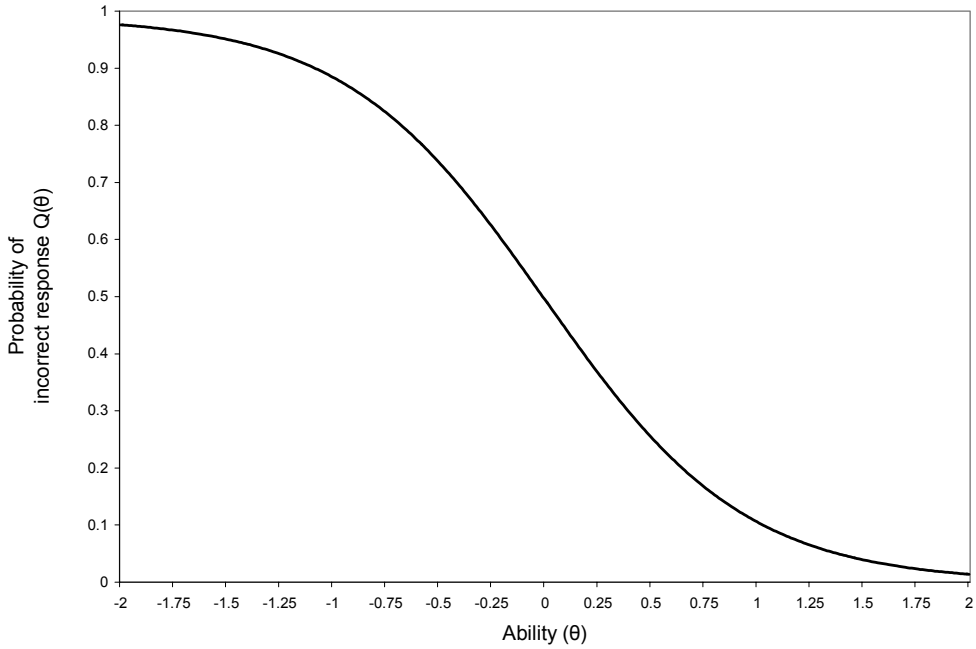

Figure 2: Example of an ICC for an item answered incorrectly

The likelihood of an examinee answering a sequence of items can be found by multiplying the ICCs for the relevant items, which is the response likelihood curve. If the examinee answers two items correctly, the likelihood of this happening is found for all values of θ by multiplying the ICCs for these two items. In this situation, the response likelihood curve would still have an S-shape, which does not provide significant information regarding the examinee's ability.

If at least one item is answered correctly and one item is answered incorrectly, the likelihood curve assumes a bell-shape as illustrated in Figure 3. The peak of a likelihood curve represents the most likely value of θ for which the particular sequence of events has occurred. Since it is the point of highest probability, this value of θ is deemed to be the most likely value of the examinee's ability.

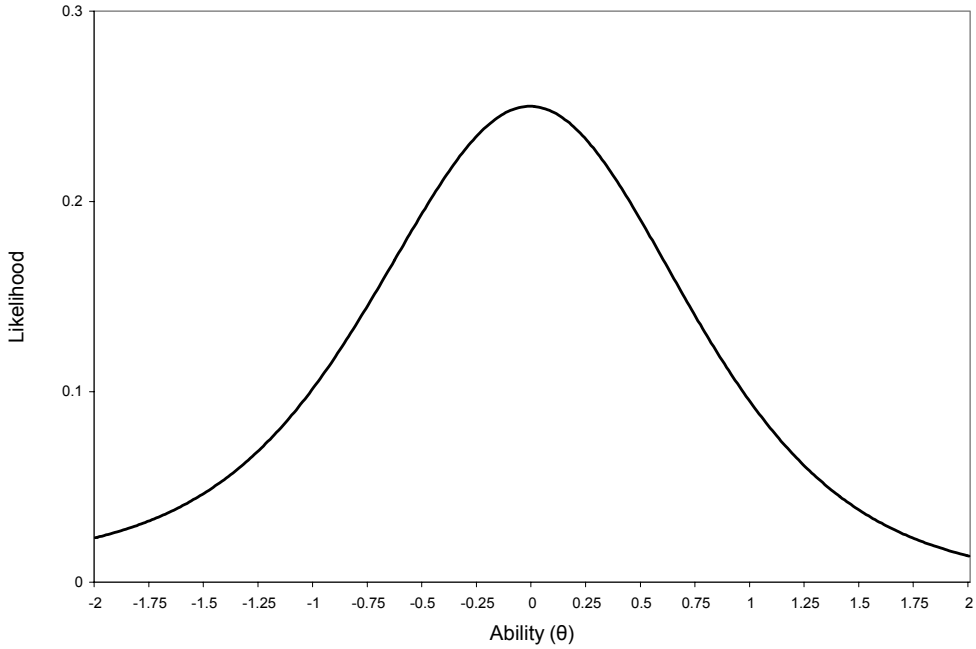

Figure 3: Example of a likelihood curve

In this paper, the explanation of IRT concepts is purposely brief, and the reader interested in further information on IRT is advised to refer to Lord (1980), Hambleton (1997) and Wainer (1990).

How the CAT Prototype works

At the University of Hertfordshire more than 200 overseas students have their English proficiency assessed every year. One of the components of this assessment is a series of multiple-choice questions where an Optical Marker Reader (OMR) is used to mark the test. Like a traditional CBT, the level of difficulty of the questions presented is the same for all the examinees who take part in a given assessment session and therefore is not tailored for the specific ability of an examinee.

In order to offer an alternative to the traditional method currently used, a high fidelity CAT prototype for testing English language was developed. The prototype consisted of a Graphical User Interface and an item bank containing 250 objective items. These objective items were either multiple-choice or multiple-response questions.

Given that the 3PL model was used both to select dynamically the questions for a particular examinee and to attempt to rate his or her ability based on his or her own responses, each item in the item bank was assigned parameters b (item difficulty), a (item discrimination) and c (pseudo-chance). As mentioned previously, these three parameters describe the ICC for the item.

Figure 4 shows a simple flowchart of how the prototype works.

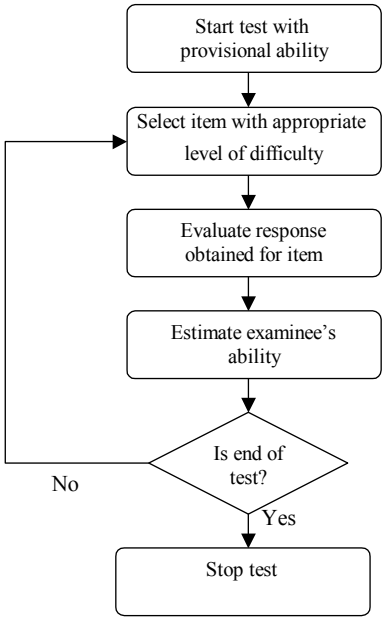

Figure 4: A flow chart illustrating the CAT

The test starts with a provisional ability for the examinee. This provisional ability can be obtained either randomly or based on the most recent estimate of the examinee's ability.

If a particular question is too easy for an examinee, it can be assumed that θ is greater than the item difficulty ($\theta > b$) and therefore the probability of this question being answered correctly by this examinee is relatively high. Likewise, if a particular question is too difficult for an examinee, it can be assumed that θ is less than the item difficulty ($\theta < b$) and therefore the probability of this question being answered correctly by this examinee is relatively low. At $\theta = b$, the mathematical functions provided by IRT can offer maximum information about the examinee's ability (Wainer, 1990). As a result, once a provisional ability has been established, the examinee is supplied with an item from the item bank for which the difficulty b is the closest value to the provisional ability θ. Figure 5 illustrates how a given item is presented to an examinee within the prototype.

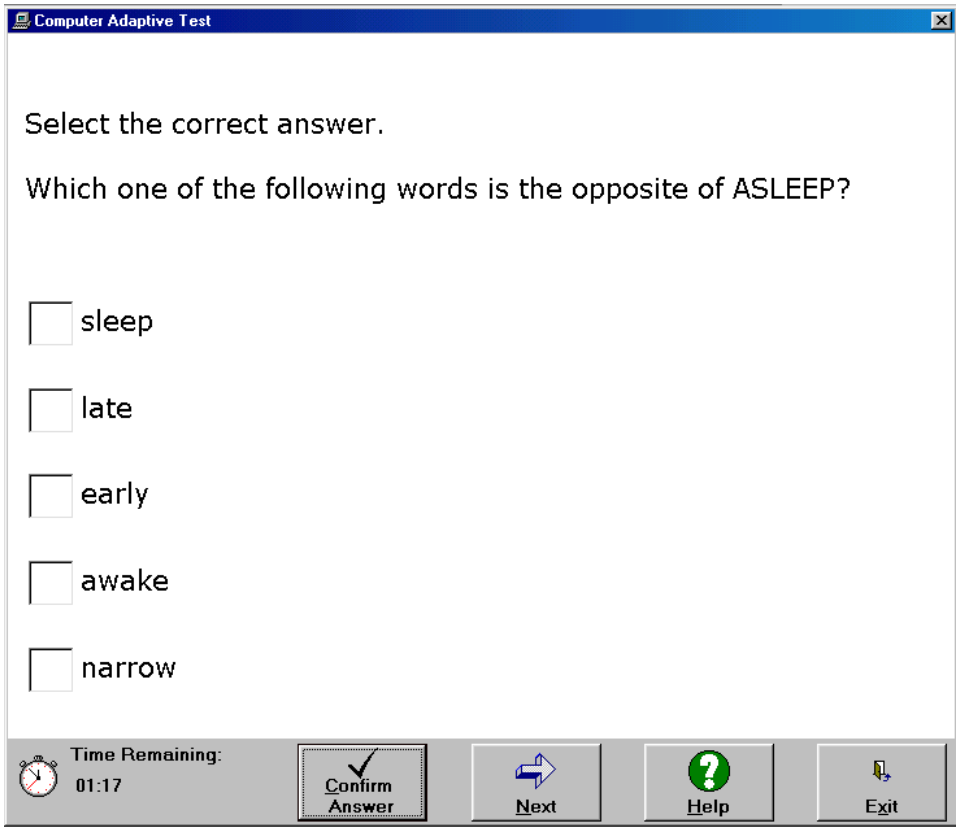

Figure 5: Example of how an item is displayed within the prototype

The examinee's response to the item presented would be evaluated as either correct or incorrect. If the response is evaluated as correct, the ICC curve used in the likelihood multiplication will be a $P(\theta)$ curve. Otherwise, the $Q(\theta)$ curve will be used.

The new provisional ability for the examinee can only be estimated when the examinee has answered at least one item correctly and one item incorrectly. This happens because the examinee's response likelihood curve is formed from the product of all the ICCs of items answered during the current test, and if the examinee answered all the items presented correctly, the examinee's response likelihood curve is composed by the product of various $P(\theta)$ and, therefore, the curve does not have a peak value. If the examinee answered all the items presented incorrectly, the examinee's response likelihood curve is calculated by the product of various $Q(\theta)$ and, consequently, the curve does not have a peak value.

When the examinee's response likelihood curve does not have a peak value, the next item to be presented would be an item with a higher difficulty when the previous question was answered correctly or with a lower difficulty otherwise. Nonetheless, when the examinee's response likelihood curve is formed by the product of at least one $P(\theta)$ and one $Q(\theta)$, the curve would typically have a peak. The value of the X-axis at the curve's peak is taken to be the new provisional ability θ.

The item to be administered next is determined by the estimated θ for a particular stage. The prototype locates in the items bank which item has not

been administered yet and has the closest value of b to the current value of θ. The item that best matches this condition is administered next. The value of the estimated ability θ is refined as new responses are evaluated.

The process of presenting items, evaluating the responses and dynamically selecting the next item to be administered is repeated until a target time limit is reached. This marks the end of the test. The examinee is then presented with a summary of how many items were answered correctly and incorrectly as well as the examinee's estimated level of ability. At present there are four possible levels: (1) Beginner, (2) Intermediate, (3) Upper-Intermediate and (4) Advanced.

Results

In order to gather data regarding the prototype's usability, a Heuristic Evaluation based on structured expert reviewing has been undertaken (Molich and Nielsen, 1990). This evaluation involved a group of eleven experts, formed by both lecturers in Computer Science and in English for Academic Purposes. After watching a short presentation, the experts were asked to undertake both a Heuristic Evaluation and an evaluation of the prototype's usefulness as a pedagogical tool, according to the guidelines provided.

In the Heuristic Evaluation described here, different elements of the interface are analysed by the experts and compared to usability principles (the heuristics). Each one of the eleven experts has independently rated ten usability standards from 1 (Poor) to 5 (Excellent). Table 1 summarises the results of the Heuristic Evaluation, where all the usability principles evaluated obtained a mean score equal or greater than 3.9 on the 1 to 5 scale.

	Poor				Excellent	
Usability Principle	**1**	**2**	**3**	**4**	**5**	**Mean**
Visibility of the system status	0	0	1	6	4	4.3
Match between system and the real world	0	0	1	4	6	4.5
User control and freedom	0	0	3	5	3	4.0
Consistency	0	0	0	5	6	4.5
Error Prevention	0	0	1	6	4	4.3
Recognition rather than recall	0	0	1	3	7	4.5
Flexibility and efficiency of use	0	0	5	2	4	3.9
Aesthetic	0	1	1	6	3	4.0
Feedback and errors	0	0	1	6	4	4.3
Help and documentation	0	2	0	6	3	3.9

Table 1: Summary of the Heuristic Evaluation

Given that in a Heuristic Evaluation five evaluators could detect 75% of the usability problems within a system (Molich and Nielsen, 1990), the scores obtained from the eleven evaluators involved in the evaluation process would suggest that there are no major usability problems within the prototype described here.

Based on the results illustrated in Table 1, it has been interpreted that the prototype's current state and available actions are made explicit to the users through simple dialogue. The location and meaning of buttons and

associated actions remain unchanged, thus improving consistency. These buttons make use of metaphors (icons), thus minimising the users' memory load. As a result, the users do not need to be familiar with system-oriented jargons or remember long sequences of commands in order to satisfactorily operate the prototype. Although the design attempts to prevent the user from making errors, when they do occur the interface is error tolerant and error messages are constructive, making recovery easy for the user.

The usability principles "flexibility and efficiency of use" and "help and documentation" obtained the lowest mean score. Despite the fact that the items presented are adapted to the level of ability of the user, the way in which the information is presented is identical for all items and therefore for all levels of difficulty. In other words, the interface cannot be adjusted according to individual user characteristics and this fact would explain the lower score for the "flexibility and efficiency of use". One of the evaluators reported that it is usually more difficult to read on a computer monitor than on paper, and this factor becomes more evident when the items presented become more difficult. As for the usability principle "help and documentation", the evaluators recognised that the prototype offers a satisfactory context-sensitive help. However, they highlighted that it is not possible to obtain information on how the test is executed before it is started.

After carrying out the Heuristic Evaluation, the experts have been asked to rate ten statements from 1 (Unlikely) to 5 (Likely) to gather data on the prototype's usefulness as an educational tool. Table 2 summarises the results obtained in this section of the evaluation.

	Unlikely				Likely	
Pedagogical Measure	**1**	**2**	**3**	**4**	**5**	**Mean**
CAT would enable lecturers to mark summative assessments more quickly.	1	1	1	2	6	4.0
CAT would enable lecturers to mark summative assessments more accurately.	1	1	1	4	4	3.8
CAT as summative assessment tool would enable lecturers to detect students' educational needs.	1	0	7	1	2	3.3
Students would be receptive to using CAT in a summative assessment environment.	0	1	3	4	3	3.8
CAT as summative assessment tool would enable students to detect their educational needs.	4	0	4	2	1	2.6
CAT as formative assessment tool would enable lecturers to detect students' educational needs.	1	1	1	5	3	3.7
Students would be receptive to using CAT in a formative assessment environment.	0	0	2	5	4	4.2
CAT as formative assessment tool would enable students to detect their educational needs.	2	3	3	2	1	2.7
Students' interaction with the system would be simple and clear.	0	0	1	4	6	4.5
Students would find the system easy to use.	0	0	0	1	10	4.9

Table 2: Summary of the Pedagogical Evaluation

The results obtained indicate that the lecturers considered that the prototype would be valuable in terms of both speed and accuracy. However, the experts suggested that the use of objective items to assess the examinee's abilities of synthesis and evaluation is restricted, and this opinion is shared by Pritchett (1999) and Ward (1980). The evaluators also emphasized that the accuracy of the score given to an examinee relies on the correctness of the item parameters used in order to estimate the examinee's ability and therefore without an adequately large and calibrated items' bank the use of a CAT is limited.

The experts believed that the prototype would give greater assistance in a formative rather that in a summative assessment environment. They suggested that formative assessments provide the lecturers with more information regarding the students' strengthens and weaknesses, since they are typically undertaken on a regular basis.

Regarding the prototype's ability to help students to detect their own potential educational needs, both summative and formative assessment environments received a mean score lower than 3. The low scores are related to the fact that the students are unaware of the adaptive process and therefore possibly unable to learn from their mistakes.

The evaluators considered that the students would more receptive to use a CAT in a formative rather than in a summative assessment environment. These results suggest that lecturers foresee problems regarding the score method used within CAT. In a CAT, the final score given to an examinee is calculated based on the number of questions answered correctly and incorrectly, as well as on the level of difficulty of these questions. As a result, examinees who answered the same number of questions correctly would almost certainly have different final scores, and this could bring uncertainties about the "fairness" of the assessment.

The prototype was tested by the Head of English Language Teaching Department at the University of Hertfordshire in September 2001 and, in his opinion, the prototype would have potential for two uses. Firstly, as a tool to support the process of testing English proficiency of overseas students. Secondly, as a tool to be used by the overseas students to improve their fluency in the English language.

Conclusion and Future Work

There is a drive towards the use of computer-based assessment within Managed Learning Environments (MLEs) in Higher Education (HE) because of investment in computer infrastructure as a possible solution to the problem of increased student numbers and also to reduce the demands on lecturers' time. It is our conviction that CAT is likely to be an important tool in this process. The evaluation presented here represents part of an on-going investigation into the value of CATs in HE. CATs are more difficult to construct than traditional CBTs because of the need for an adaptive algorithm and a larger and calibrated question bank. It is important therefore, to understand both the limitations and the opportunities afforded by the CAT

approach and to investigate how such a powerful tool might be applied to assessment and learning in a broader context.

The recent extensive use of CBT on the University of Hertfordshire's MLE has allowed the identification of some practical problems involved when large numbers of students are assessed on-line at the same time. These include network slowing when users simultaneously access central databases of questions, and not least the practical management of large numbers of students taking tests in a given area. It will be important in the future to perform large scale testing of our prototype to make sure that network performance is adequate. Initial work with the prototype indicates that this will be no more of a problem with CAT than CBT. An important practical benefit of CAT is that it would be unlikely for any individual undergoing CAT to be answering the same question as any other, reducing the risk of unauthorised collaboration between students and making management of sessions easier.

It is important that CAT does not hinder assessment by introducing extraneous variables, such as cognitive overhead, due to the computer interface. Bly and Rosenberg (1986), for example, have investigated such issues in computer applications. Khan (1995) suggests that limited capacity of human information processing is the reason for cognitive overhead. Excess orientation, navigation and user-interface adjustment place added strain on the user leading to cognitive overhead. Khan showed that cognitive overhead must be kept to a minimum if performance in an application is to be high. The prototype described in our study performed well in all usability tests and no major usability problems were identified by experts. It is therefore assumed that cognitive overhead was low and introduced no barriers to assessment.

Table 2 above shows that in general, the idea of CAT was quite well received by the evaluators, though its use as a tool for tutors and students to detect educational needs in formative and summative assessment was rated lowest. This is an important possible limitation to the CAT approach. Barker and colleagues (Barker et al., 2002) have investigated some of the issues involved in differentiating summative assessment according to learners' ability. Although they found that student performance was improved and that there were several other benefits, including motivation, tutors showed some reluctance to the approach especially for summative assessment. Barker and colleagues' use of a co-operative method of setting assessment levels (where learners and tutors agreed assessment levels, based on performance in a computer application) overcame some of these problems. It will be interesting to see if co-operation could be used within CAT thus allowing tutors and students to accept more responsibility for setting question level than is afforded by a purely automatic adaptive approach.

Muldner and colleagues (1997) suggest that it is possible to capture some aspects of student learning using an assessment-based student modelling approach by adapting the presentation of instruction in a computer-based learning application according to how well learners answered questions within the application. Barker and colleagues (Barker et al., 2002) used a slightly different approach, using a combination of computer adaptation and co-

operation to determine how information was presented and also the level of assessment undertaken by learners. The use of CAT in such approaches, based on adaptive student modelling, would enable instructional designers to tailor the presentation of information very accurately to the optimum level of student ability.

Finally, the current prototype will be improved in several ways, for example allowing tutors greater control in configuring aspects of the application, such as types of questions supported. The use of a wide range of questions is important in developing good approaches to learning (Felder and Brent 1994). As assessment is important as a formative tool in learning, it will be vital to extend the CAT model to support a wider range of question types, possibly involving work on and off the computer, in order to include higher levels of student cognitive ability. Adaptability to different students' abilities could be improved by implementing a second stop condition, where the test would be stopped either when a specific standard error for θ has been met or when a certain time duration has been reached, whichever happens first. Both the support of a wider range of questions and the implementation of different stop conditions will be part of future work on the prototype.

References

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), 207-241.

Bly, S. A. and Roseberg, J. K. (1986). A comparison of tiled and overlapping windows, In: *Proceedings of CHI '86*. ACM Press, New York, 101-106.

Felder, R. M. and Brent, R. (1994). *Cooperative Learning in Technical Courses: Procedures, Pitfalls and Payoffs*. NSFDUE Grant DUE-9354379, October, 1994.

Freedle, R.O. and Duran, R.P. (1987). *Cognitive and Linguistic Analyses of test performance*. New Jersey: Ablex Publishing Corporation.

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

Harvey, J. and Mogey, N. (1999). *Pragmatic issues when integrating technology into the assessment of students* in Brown, S., Race, P. and Bull, J. Computer-Assisted Assessment in Higher Education. London: Kogan Page.

Khan, P. (1995). Visual Cues for Local and Global Coherence in the WWW. *Communications of the ACM*, 38(8), 67-69.

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

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

Muldner, T., Muldner, K. and van Veen, C. M. (1997). Experience from the design of an authoring environment. *Journal of Educational Multimedia and Hypermedia,* 6(1), 114-132.

Pritchett, N. (1999). *Effective Question Design* 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.