An automated feedback system based on adaptive testing: extending the model

Trevor Barker

University of Hertfordshire

Key words: *Computer Adaptive Testing, Automated Feedback, Assessment*

Abstract:

The results of the recent national students survey (NSS) revealed that a major problem in HE today is that of student feedback. Research carried out by members of the project team in the past has led to the development of an automated student feedback system for use with objective formative testing. This software relies on an 'intelligent' engine to determine the most appropriate individual feedback, based on test performance, relating not only to answers, but also to Bloom's cognitive levels. The system also recommends additional materials and challenges, for each individual learner. Detailed evaluation with more than 500 students and 100 university staff have shown that the system is highly valued by learners and seen by staff as an important addition to the methods available. The software has been used on two modules so far over a two year period.

1 Introduction

Despite the reported benefits of the computer-aided assessment approach, high staff/student ratios often mean that tutors are often unable to provide learners with feedback on assessment performance that is timely and meaningful. Freeman & Lewis (1998) amongst others have reported on the importance of feedback as a motivator for student learning. Thus, there is an increasing demand for the development of software applications that would enable the provision of timely, individual and meaningful feedback to those learners who are assessed via computer-aided assessment applications.

Computer Adaptive Testing

The development of the CAT application that was the subject of this study has been reported by Lilley and colleagues (Lilley et al. 2004; 2005). The application comprised a graphical user interface, an adaptive algorithm based on the Three-Parameter Logistic Model from Item Response Theory (Lord, 1980; Hambleton, 1991; Wainer, 2000) and a database of questions. This contained information on each question, such as stem, options, key answer and IRT parameters. In this work, subject experts were employed for question calibration. The subject experts used Bloom's taxonomy of cognitive skills (Pritchett, 1999; Anderson & Krathwohl2001) in order to perform the calibration. Questions were first classified according to cognitive skill being assessed. After this initial classification, questions were then ranked according to difficulty within each cognitive level. Table 1 summarises the three levels of cognitive skills covered by the question database and their difficulty range. It can be seen from Table 1 that knowledge was the lowest level of cognitive skill and application was the highest. An important assumption of our work is that each higher level cognitive skill will

include all lower level skills. As an example, a question classified as application is assumed to embrace both comprehension and knowledge.

Table 1: Level of difficulty of questions

Difficulty *b*	**Cognitive Skill**	**Skill Involved**
$+1 \leq b \leq +3$	Application	Ability to apply taught material to novel situations
$+1 < b < -1$	Comprehension	Ability to interpret and/or translate taught material
$-1 \leq b \leq -3$	Knowledge	Ability to recall previously taught material

At the end of each assessment session, questions were re-calibrated using response data obtained by all participants who attended the session. In general terms, questions that were answered correctly by many test takers had their difficulty levels lowered and questions that were answered incorrectly by many test takers had their difficulty levels increased.

- Our research so far on the development and use of CAT systems in Higher Education has related to
  - Establishment of test conditions
    - E.g. ability to review questions, test stopping conditions
  - The reliability of CAT measures
    - Test-retest (reliability studies)
  - The fairness of the method
    - Comparison to other testing methods (validity studies)
  - Student perception of test difficulty
  - Student and staff attitude to CAT method
  - The adaptive questions database
  - Use of CAT in formative and summative tests
  - Using CAT model to provide automated feedback

Based on this research, the Graphical User Interface developed and used in this project is shown in figure 1 below.

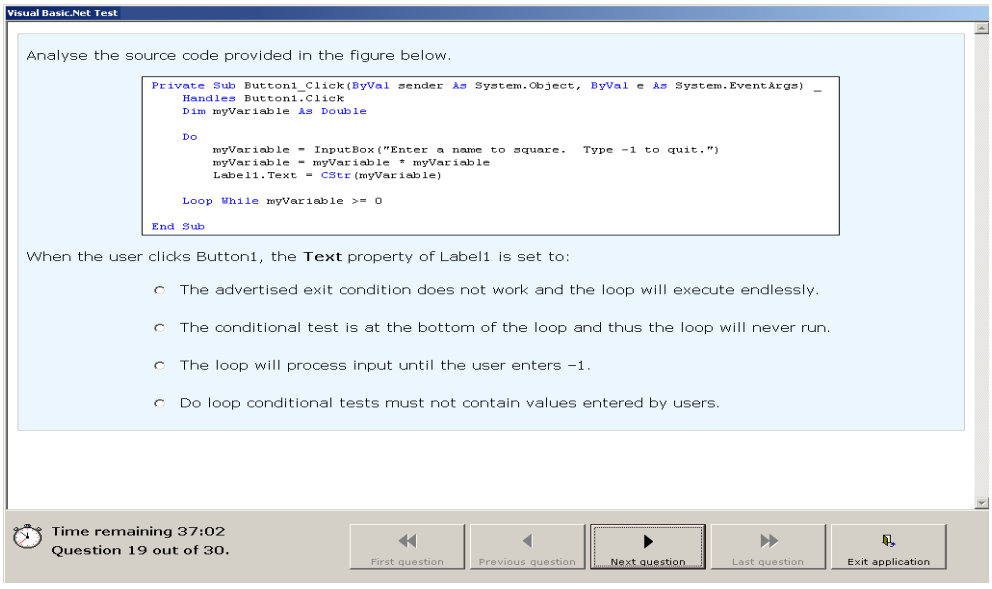


Figure 1: Graphical User Interface developed for the CAT systems employed in this study

Our approach to the provision of automated feedback

In earlier work, our research has shown that a system of automated feedback, based on student performance in a Computer Adaptive Test was useful, efficient and generally well regarded by students (Lilley and Barker 2002; 2003; 2004). Barker and colleagues (2002) noted the importance of all major stakeholders in design, implementation and evaluation of projects related to online learning.

It was one of our assumptions that a tutor-led feedback session would typically comprise the provision of an overall score, general comments on proficiency level per topic and recommendations on which concepts within the subject domain should be revised. It was then planned that the feedback would be made available via a web-based application.

Overall score

The overall score, or overall proficiency level, would be estimated by the CAT algorithm using the complete set of responses for a given test-taker and the adaptive algorithm introduced in section 2.1. Figure 2 below illustrates how this information was displayed within our automated feedback prototype.

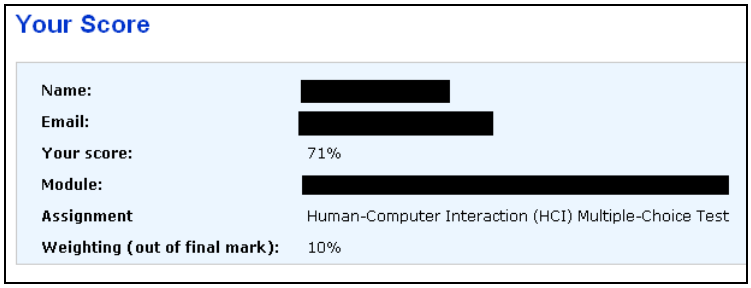


Figure 2: Screenshot illustrating how overall score was displayed within our automated feedback prototype. The student's name and module have been omitted.

Performance summary per topic

Test-takers' responses would be grouped by topic and a proficiency level calculated for each set of topic responses. Proficiency level estimates per topic would then be mapped to Bloom's taxonomy of cognitive skills. The underlying idea was to inform learners about their degree of achievement for each topic domain. Some learners reported that they would also like to compare their test performance with the performance of the whole group. This information was also made available in this section of the feedback, as illustrated in Figure 3.

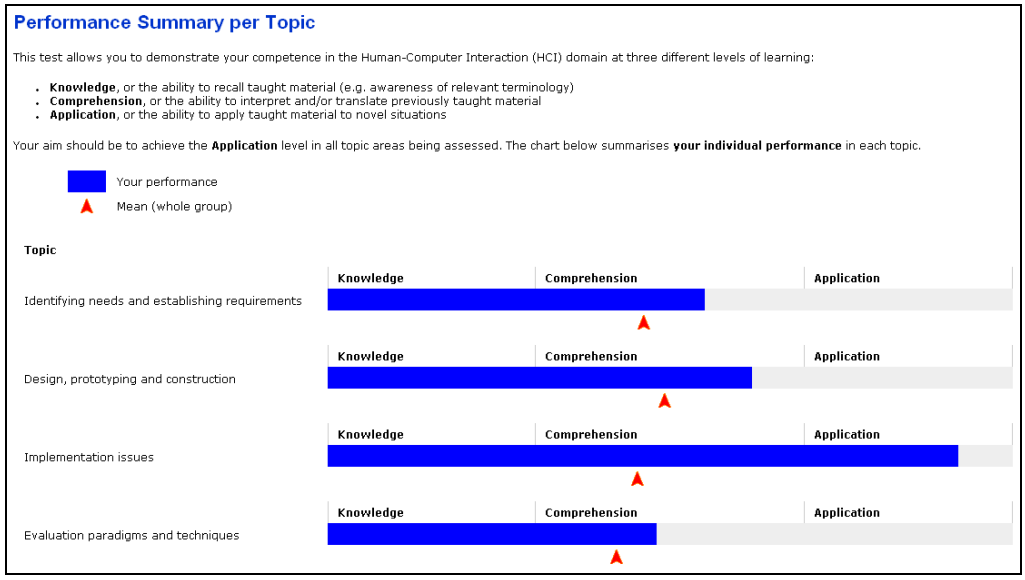
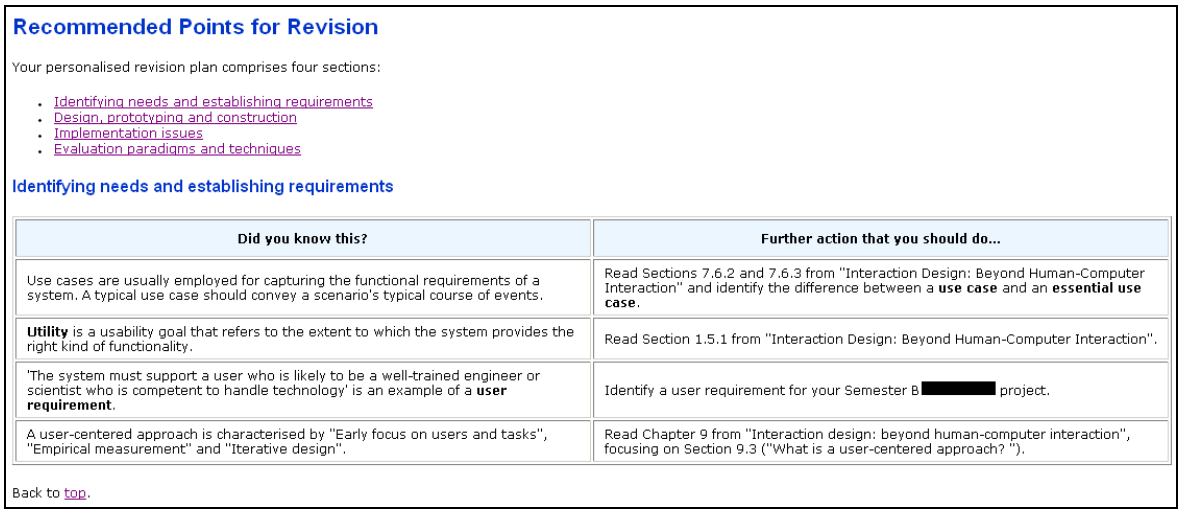


Figure 3: Screenshot of screen containing information regarding performance per topic.

Recommended points for revision

An important assumption of our feedback tool was that tutors providing feedback on an objective test during a face-to-face session were likely to provide students with directive feedback rather than simply indicating what the correct options for each question were. As an initial attempt to mimic some aspects of how a subject domain expert would provide learners with recommendations on how to increase their individual proficiency levels, a database of feedback sentences was designed and implemented. This database comprised statements relating to each one of the questions. For each individual student, only those questions answered incorrectly were selected. Figure 4 illustrates the approach to directive feedback employed in this study.



Recommended Points for Revision

Your personalised revision plan comprises four sections:

- Identifying needs and establishing requirements
- Design, prototyping and construction
- Implementation issues
- Evaluation paradigms and techniques

Identifying needs and establishing requirements

Did you know this?	Further action that you should do...
Use cases are usually employed for capturing the functional requirements of a system. A typical use case should convey a scenario's typical course of events.	Read Sections 7.6.2 and 7.6.3 from "Interaction Design: Beyond Human-Computer Interaction" and identify the difference between a **use case** and an **essential use case**.
Utility is a usability goal that refers to the extent to which the system provides the right kind of functionality.	Read Section 1.5.1 from "Interaction Design: Beyond Human-Computer Interaction".
'The system must support a user who is likely to be a well-trained engineer or scientist who is competent to handle technology' is an example of a **user requirement**.	Identify a user requirement for your Semester B [redacted] project.
A user-centered approach is characterised by "Early focus on users and tasks", "Empirical measurement" and "Iterative design".	Read Chapter 9 from "Interaction design: beyond human-computer interaction", focusing on Section 9.3 ("What is a user-centered approach? ").

Back to top.

Figure 4: Example of 'Recommended Points for Revision' for the topic 'Identifying needs and establishing requirements'. The module name has been omitted.

Extending the model for general use

In order that the CAT system and the associated feedback could be used more generally by students and teachers on a wider range of modules,, it was necessary to simplify the CAT and feedback systems described above. The most important features of the system that needed modification for wider use were:

- The question database
- The feedback database
- The feedback delivery system

The question database

In order that the system could be made as simple as possible for use by inexperienced tutors, a simplified calibration system for the questions database was employed. It was recommended that in the first place, a test be divided into topic areas. It was then required that the questions in each topic area of the test should cover the required syllabus at each difficulty level. To ensure adequate coverage, approximately four times the number of questions to be delivered in the test had to be written. Previous work has shown that this provided sufficient coverage of the topic area at all difficulty levels. Calibration of the question data base involved ranking the questions in order of difficulty. Two tutors ranked the questions individually and the mean of their ranking was used to calibrate the test items. Questions were also ranked according to Blooms three lower levels, knowledge, understanding and application. Using experienced tutors to undertake the ranking ensured face validity to the calibration. Earlier research has shown that this method produced a relatively stable and valid question database.

The feedback database

It was intended that learners receive feedback on the following as described in the preceding sections.

- Overall proficiency level;
- Performance in each topic;
- Recommended topics for revision
- Cognitive level (according to Bloom)

In order to simplify the writing of feedback for more general use of the system, the following was recommended:

For each question in the question database tutors should write:

- General comment about Bloom's level achieved in this topic.
- A general comment / information about the topic area covered in the question
- A statement to be presented if the question was answered correctly
- A statement to be presented if the question was answered incorrectly
- A link to course material / information where the topic was covered in the MLE
- A link forwards to related / more challenging materials
- Links to remedial materials

The feedback delivery system

The precise nature and complexity of the feedback would therefore be determined by the tutor. The CAT feedback systems would then deliver the feedback individually to learners. Previously in our research, web-based systems were used to deliver feedback. In order to simplify this for more general use, the CAT feedback application was modified in order to produce a Microsoft Word document that was then sent via email to each individual learner, using simple electronic mail merge. In principal this could be achieved within minutes of the test finishing. However results of the test were inspected carefully prior to release of results and feedback to ensure that the test was fair and that no errors had occurred.

Discussion

The CAT feedback system described above has several benefits. It has been tested and evaluated by staff and students and shown to be effective and valued. Feedback is individual, as it is based on an adaptive test. No two students would be expected to have the same test, thus feedback would also be individual. Feedback is also set at the most appropriate level for each individual learner and is also related to Bloom's cognitive domain.

Substantial investments in computer technology by Higher Education institutions and high staff/student ratios have led to an increased pressure on staff and students to incorporate electronic methods of learning and teaching. This includes a growing interest in the use of computer-aided assessment, not only to make the technological investment worthwhile but also to explore the opportunities presented by the computer technology available. It is our experience that - given the great deal of computerised objective testing that currently takes place – using adaptive tests is an interesting, fair and useful way of providing such assessment (Barker & Lilley, 2003; Lilley et al., 2004). Not only is this motivating for learners, who are challenged appropriately - i.e. not discouraged by questions that are too hard, or de-motivated by questions that are too easy - but also the information that it provides can be used in interesting and useful ways. For instance, it can be used in the presentation of remedial work for students or, as in our case, for the provision of personalised feedback.

Feedback must be timely to be useful. Our experience is that when large-scale computerised objective testing is used in a formative context, results are usually returned quickly, because of automated methods of marking. Feedback, however, is often slow and delivered by the time the course has moved on and it is of less use or, in some cases, feedback is absent. This experience was largely confirmed by the results obtained in the current study. It is time consuming to produce individual feedback for hundreds of students. When feedback is provided, it is usually little more than a final score, generic worked examples and a list of questions answered correctly and incorrectly. Automated methods are therefore likely to be useful in this context, as evidenced by the tutors' attitude reported in this study. The matching of adaptive testing and automated feedback provides an opportunity to individualise feedback to a far greater extent. We argue that the automated feedback approach proposed here, which is based on adaptive testing, is appropriate for identifying learners' strengths and weaknesses for each topic area covered by the test. Automated feedback as proposed in this study is also related to Bloom's levels, thus providing meta level information for learners about the depth of their approach in each of the topic areas. This information would be difficult to obtain with standard objective testing.

Other approaches to the provision of feedback to groups of learners, such as in-class sessions where all questions from an objective test are presented by a tutor, are likely to remain as

important feedback methods. Such in-class approaches offer high quality information about the test and each of the questions, often providing learners with an opportunity to work through the questions. They do not, however, address the individual needs of many of the learners. Explaining a question that is set at a difficulty level that is too low for most learners will not be of interest for the majority of the group. Similarly, it can be argued that discussing questions that only one or two learners are capable of answering will not be the most efficient way of employing tutors' and learners' time. We suggest that not only is the automated feedback based on adaptive testing a fast and appropriate method, but that it also provides information to learners that would be difficult to obtain elsewhere, given the decrease in the number of face-to-face sessions, the increase in staff/student ratios and the growing trend in the use of electronic resources for the delivery of courses, assessment, student feedback and support. It has also been simplified and is suitable for use on a wide range of modules. So far we have successfully used the CAT / feedback system in computer programming, English language, civil engineering, human-computer interaction and multimedia. We intend to further simplify the CAT feedback system for even wider use.

Our research has shown that learners and tutors accept and value the automated feedback approach proposed in this study. In the future we intend to apply this method more widely, for example in providing feedback for written assignments. We also intend to use the wealth of information about learners' proficiency levels provided by the adaptive testing approach to develop useful student models. Such student models will, in turn, be employed to generate profiles that could be used in a wide variety of learning contexts.

References

Anderson, L. W. & Krathwohl, D. R. (Eds.) (2001) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York.

Barker, T. & 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.

Freeman, R. & Lewis, R. (1998) *Planning and Implementing Assessment*, Kogan Page, London.

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

Lilley, M. & 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 University, United Kingdom, pp. 169-184.

Lilley, M. & Barker, T. (2003) "Comparison between Computer-Adaptive Testing and other assessment methods: An empirical study", *Proceedings of the 10th International Conference of the Association for Learning Technology (ALT-C)*, University of Sheffield, United Kingdom.

Lilley, M. & Barker, T. (2004). "A Computer-Adaptive Test that facilitates the modification of previously entered responses: An empirical study", Proceedings of the 2004 Intelligent Tutoring Systems Conference, *Lecture Notes in Computer Science 3220*, pp. 22-33.

Lilley, M., Barker, T. & Britton, C. (2004) "The development and evaluation of a software prototype for computer adaptive testing", *Computers & Education Journal 43(1-2)*, pp. 109-123.

Lilley, M., Barker, T. & Britton, C. (2005) "The generation of automated learner feedback based on individual proficiency levels", Proceedings of the 18th International Conference

on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems, *Lecture Notes in Artificial Intelligence 3533*, pp. 842-844.

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

Pritchett, N. (1999) "Effective question design" In S. Brown, P. Race & J. Bull (Eds.), *Computer-Assisted Assessment in Higher Education*, Kogan Page, London.

Wainer, H. (2000) *Computerized Adaptive Testing (A Primer)*, Lawrence Erlbaum Associates, New Jersey.

Author:

Dr Trevor, Barker
Department of Computer Science
University of Hertfordshire
Hatfield
Hertfordshire
AL10 9AB
UK
Email: t.1.barker@herts.ac.uk