

THE GENERATION OF AUTOMATED STUDENT FEEDBACK FOR A COMPUTER-ADAPTIVE TEST

Mariana Lilley, Trevor Barker and Carol Britton

The Generation of Automated Student Feedback for a Computer-Adaptive Test

Mariana Lilley, Trevor Barker and Carol Britton
University of Hertfordshire
School of Computer Science
College Lane, Hatfield
Hertfordshire AL10 9AB
M.Lilley@herts.ac.uk;
T.1.Barker@herts.ac.uk; C.Britton@herts.ac.uk

Abstract

This paper marks further progression on research previously undertaken at the University of Hertfordshire on the use of computer-adaptive tests (CATs) in Higher Education. Findings from two previous empirical studies by the authors suggested that the CAT approach was a fair assessment method, capable of offering accurate and consistent measurement of student abilities. Participants in a pedagogical evaluation of the application indicated that one of the limitations of the approach was the type of the feedback provided to students. According to the evaluators, the sole provision of a score would not help students to detect their educational needs. Providing students with a copy of all questions they got wrong did not seem an attractive option either, as it could jeopardise the re-use of these questions in future assessment sessions. Furthermore, it seemed unlikely that providing students with the questions alone, without any comment or explanation, would foster research and/or reflection skills.

This paper reports on our most recent empirical study, in which the ability estimate θ for each student in each section of the CAT test was used to generate automated feedback based on Bloom's taxonomy of cognitive abilities. The feedback was then sent directly to individual students via personal email. In the first section of this paper, we present an overview of our CAT research followed by the main characteristics of the feedback tool we designed and implemented. In the final section of this paper, we present the results a summary of how learners performed on the CAT, along with student attitude towards the automated feedback. In addition, we present our views on how the work described here can be developed further.

Introduction

The adoption of computerised adaptive testing by some high-stake, large scale examinations such as the Graduate Management Admission Test (Graduate Management Admission Council, 2002), Test of English as a

Foreign Language (Wainer, 2000), Graduate Records Examination (Wainer, 2000), Armed Sciences Vocational Aptitude Battery (Wainer, 2000) and Microsoft Certified Professional (Microsoft Corporation, 2002) suggests an increased interest in computer-assisted assessments that are capable of tailoring the test questions to the individual knowledge of each test-taker. The reasons for this increased interest range from higher levels of efficiency when assessing large numbers of candidates to higher levels of personalisation and individualisation than those supported by traditional computer-based tests (Lord, 1980; Jacobson, 1993; Carlson, 1994; Wainer, 2000; Rafacz & Hetter, 2001). Furthermore, our research to date (Lilley & Barker, 2002; Barker & Lilley, 2003; Lilley & Barker, 2003a; Lilley & Barker, 2003b) seems to corroborate the view that computer-adaptive tests (CATs) have the potential to offer a fair, accurate and consistent measurement of student ability.

A CAT prototype has been designed, developed and evaluated at the University of Hertfordshire over the past three years. The prototype was based on the Three-Parameter Logistic (3-PL) Model from Item Response Theory (IRT) and the rationale for using this particular statistical model is described in earlier work by the authors (Lilley & Barker, 2002; Lilley, Barker, Bennett & Britton, 2002; Barker & Lilley, 2003; Lilley & Barker, 2003a; Lilley & Barker, 2003b; Lilley, Barker & Britton, 2004).

The 3-PL model provides a mathematical function used to predict the probability of a student with an unknown ability θ correctly answering a question of difficulty b , discrimination a and pseudo-chance c . This mathematical function is shown in Equation 1 (Lord, 1980).

In order to evaluate the probability Q of a student with an unknown ability θ incorrectly answering a question of difficulty b , the function $Q(\theta) = 1 - P(\theta)$ is used (Lord, 1980). Within a CAT, the question to be administered next as well as the final score obtained by any given student is computed based on the set of previous responses, which is obtained using the mathematical function shown in Equation 2 (Lord, 1980).

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

Equation 1: The Three-Parameter Logistic Model

$$L(u | \theta) = \prod_{i=1}^n \left(\frac{P_i}{Q_i} \right)^{u_i} \cdot \prod_{i=1}^n Q_i$$

Equation 2: Likelihood function

The Three-Parameter Logistic Model as shown in this study supports only questions that are dichotomously scored. For instance, consider a student who answered a set of three multiple-choice questions, in which the first and second responses were incorrect and the third response was correct, such as $u_1 = 0$, $u_2 = 0$ and $u_3 = 1$. The likelihood function (see Equation 2) for this example is $L(u_1, u_2, u_3 | \theta) = (P_1^0 Q_1^1)(P_2^0 Q_2^1)(P_3^1 Q_3^0)$, or more concisely $L(u_1, u_2, u_3 | \theta) = Q_1 Q_2 P_3$. The response likelihood curve will assume a bell-shape when a student has entered at least one correct and one incorrect response. IRT suggests that the peak of this curve is the most likely value for

this student's ability θ estimate. A detailed description of IRT is beyond the scope of this paper and only a brief overview is given here. The interested reader is referred to Lord (1980) and Wainer (2000). In the next section, we provide some background information on the assessment session that was the focus of our automated feedback study.

Background on the assessment session

The CAT application consisted of a Graphical User Interface and a database comprising 109 objective questions. These objective questions were independently ranked according to their difficulty by experts and assigned a value for the b parameter (see Equation 1). Values to the b parameter were assigned according to Bloom's taxonomy of cognitive skills (Pritchett, 1999), as shown in Table 1. One of the underlying ideas within Bloom's taxonomy of cognitive skills (Anderson & Krathwohl, 2001) is that tasks can be arranged in a hierarchy from less to more complex. This hierarchy was then used within our CAT prototype to classify questions according to the proficiency level required on the part of the learner to successfully complete a task (i.e. answer a question).

Difficulty b	Cognitive skill	Skill involved
$-2 \leq b \leq -0.6$	Remember	Ability to recall taught material
$-0.6 \leq b \leq 0.8$	Understand	Ability to interpret and/or translate knowledge
$0.8 \leq b \leq 2$	Apply	Ability to apply knowledge to novel situations

Table 1: Values assigned to the *difficulty* parameter

One hundred and twenty-three second year students enrolled in a programming module of the Bachelor of Science (BSc) in Computer Science degree at the University of Hertfordshire participated in a computer-assisted assessment session using the CAT application.

The participants took the test on week 30 as part of their real assessment for the module. The test was within the subject domain of Human-Computer Interaction (HCI) and covered six different topic areas. The topic areas are listed in Table 2.

Students had 30 minutes to answer 20 questions within the subject domain. To allow tutors to monitor the fairness of the assessment, 14 of the questions administered were dynamically selected according to their performance during the test. The remaining 6 questions were selected in advance by the tutors and administered to all students. Table 2 shows how many questions were administered per topic area within the subject domain in addition to the number of questions previously selected by tutors (i.e. non-adaptive) and number of questions dynamically selected (i.e. adaptive).

Topic area	Number of non-adaptive questions	Number of adaptive questions	Total number of questions per topic area
Issues related to the use of sound at interfaces	1	2	3
Graphical representation at interfaces, focusing on the use of colour and images	1	3	4
User-centred approaches to requirements gathering	1	3	4
Design, prototyping and construction	1	2	3
Usability goals and User experience goals	1	3	4
Evaluation paradigms and techniques	1	2	3

Table 2: Main characteristics of the object questions administered during the assessment session

The test started with the 6 non-adaptive questions, followed by the 14 adaptive ones. It is important to note that the level of difficulty of the question to be administered next for each individual student was based on his or her whole set of previous responses (see Equation 2). One of our assumptions was that a given student knowledge for one topic was likely to be a good indicator of his or her knowledge for any other topic within the same subject domain. The following section presents a summary of the way how feedback on student performance was provided in prior assessment sessions.

Feedback provided for the first and second assessment sessions

To investigate the feasibility of sending scores directly to individual students via personal email, this approach was used to provide learners with their scores in two previous sessions of assessment. Figure 1 illustrates the template created in Microsoft Word.

To: <<Student_Name>>

Your score for the Visual Basic Theory Test 1 was <<Student_Score>>%.

*This is an automated message from
The Programming_Module team*

Figure 1: Template used in pilot study

The values used for the Student_Name and Student_Score fields were retrieved from the actual CAT database. The emails were generated using the Mail Merge facility provided by Microsoft Word. Student scores were sent via email within one week after test completion.

Although students seemed pleased to receive their scores via email, informal feedback received from some students indicated that the score on its own provided learners with very little – if any – help in determining which part of the subject domain they should revise next or which topic they should prioritise. Their views were in line with the opinion of the experts who participated in the pedagogical evaluation of the CAT prototype (Lilley & Barker, 2002). In this pedagogical evaluation, the experts reported that the score provided by the CAT prototype was unlikely to help students to identify their educational needs.

A simple, but unattractive, potential solution was to provide students with a copy of all questions they got wrong. A major limitation of this approach was the lack of explanation or comment on their performance. A further practical limitation of the approach was increased exposure of the objective questions stored in the database. This exposure could, in turn, jeopardise the re-use of these questions in future assessment sessions. It is important to emphasize that the re-use of questions is one of the perceived benefits of computer-assisted assessments (Freeman & Lewis, 1998; Harvey & Moge, 1999). Furthermore, it seemed unlikely that providing students with the answers to the questions they did not get right would foster research and/or reflection skills.

In the next section of this paper we present a summary of student performance on the HCI test followed by an overview of how the feedback for the HCI test was generated is presented.

Automated feedback using Item Response Theory

Tables 3 and 4 show a summary of student performance for the HCI test.

Mean CAT Level	Mean % Correct responses (Non-adaptive mode)	Mean % Correct Responses (Adaptive mode)
-1.236	41.33	42.17

Table 3: Summary of overall performance (N = 123)

Topic area	Mean Ability Level	Mean % Correct responses (Both modes)
Issues related to the use of sound at interfaces	-0.70213	42.09
Graphical representation at interfaces, focusing on the use of colour and images	-1.25967	37.3
User-centred approaches to requirements gathering	-1.19623	42.22
Design, prototyping and construction	-0.49066	34.16
Usability goals and User experience goals	-0.77508	40.37
Evaluation paradigms and techniques	-0.97738	37.44

Table 4: Summary of performance per topic (N = 123)

Regarding the feedback for those students who took the HCI test, it was envisaged by tutors that all students should receive a feedback document containing three sections: overall score, a summary of performance in each topic and a list of topics for revision. The generated feedback document should then be sent to student personal email accounts as a Word document attachment (file extension .doc).

Overall score

This section of the feedback document contained the overall score for the test. The inclusion of the overall score for the test was simple, as this data was stored in the CAT database immediately after students completed the HCI test.

Feedback according to topic

The aim in this section of the feedback document was to provide students with a summary (up to 50 words) of their performance in each topic area. To this end, all responses for each individual student were select from the CAT database. Student responses were then grouped by topic and an ability level was calculated using the functions shown in Equations 1 and 2. It is important to note that the ability level was calculated for each group of responses (i.e. set of student responses for a given topic).

Ability Level	Number of correct responses	Feedback sentence
$-2 \leq b \leq -0.6$	All responses for this topic were incorrect	None of your responses provided in this section of the assignment were correct. We strongly recommend that you start reviewing user-centred approaches to usability goals as soon as possible.
$-2 \leq b \leq -0.6$	One or more correct responses	In this section of the assessment, you demonstrated awareness of relevant terminology related to Usability goals and User experience goals. We recommend that you now concentrate on identifying which Usability goals are most likely to be relevant for your Semester B project.
$-0.6 \leq b \leq 0.8$	One or more correct responses	Your performance in this section of the assessment suggests an understanding of the role of Usability goals and User experience goals within the system development process. With the importance of Usability goals and User experience goals in mind, start planning how you are going to apply these concepts to your Semester B multimedia project.
$0.8 \leq b \leq 2$	One or more correct responses	You showed knowledge and understanding of fundamental principles related to Usability goals and User experience goals. Your performance in this section of the assessment suggests an ability to apply these principles to your multimedia project.
$0.8 \leq b \leq 2$ <i>and levels for remaining topics areas are different from</i> $0.8 \leq b \leq 2$	One or more correct responses	This is the section of the assignment in which you performed best. You showed knowledge and understanding of fundamental principles related to Usability goals and User experience goals. Your performance in this section of the assessment suggests an ability to apply these principles to your multimedia project.
$0.8 \leq b \leq 2$	All responses for this topic were correct	You have answered all questions in this section of the assignment correctly. You showed knowledge and understanding of fundamental principles related to Usability goals and User experience goals. Your performance in this section of the assessment suggests an ability to apply these principles to your multimedia project.

Table 5: Feedback sentences for "Usability"

In addition to an algorithm capable of calculating an ability level based on IRT principles, the automated feedback application comprised a database of feedback sentences. The sentences database consisted of 87 records.

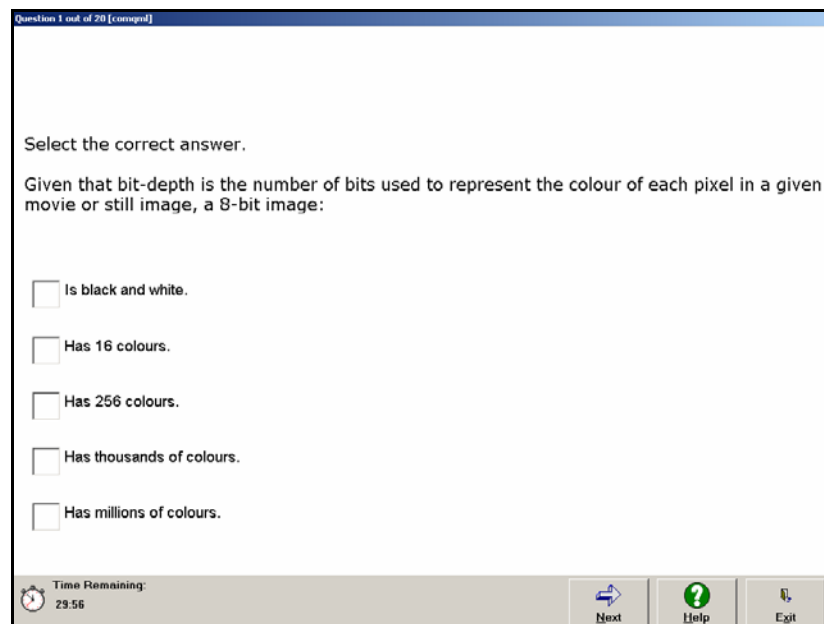
Thirty-six records were sentences related to topic areas and the remaining 51 records were feedback sentences related to questions.

Given that all test questions were calibrated using Bloom's taxonomy of cognitive skills, the sentences created for this section of the feedback were also based on Bloom's taxonomy. Using the topic "Usability" as an example, Table 5 illustrates how these sentences were structured.

Feedback according to question

This section of the assignment comprised a list of points for revision, based on the questions answered incorrectly by each individual student.

Each question in the database had a feedback sentence associated with it. This feedback sentence did not reproduce the question itself. Instead, the feedback sentence listed specific sections within the recommended reading and/or additional learning materials. The same feedback sentence could be used for more than one question in the database. For instance, consider the questions shown in Figures 2 and 3. Both questions had the same feedback sentence associated with, as shown in Figure 4.



The screenshot shows a question interface with a blue header bar containing the text "Question 1 out of 20 [completed]". Below the header, the text reads "Select the correct answer." followed by "Given that bit-depth is the number of bits used to represent the colour of each pixel in a given movie or still image, a 8-bit image:". There are five radio button options: "Is black and white.", "Has 16 colours.", "Has 256 colours.", "Has thousands of colours.", and "Has millions of colours.". At the bottom, there is a grey bar with a clock icon and "Time Remaining: 29:56", and three buttons: "Next" (with a right arrow), "Help" (with a question mark), and "Exit" (with a door icon).

Figure 2: Example of question

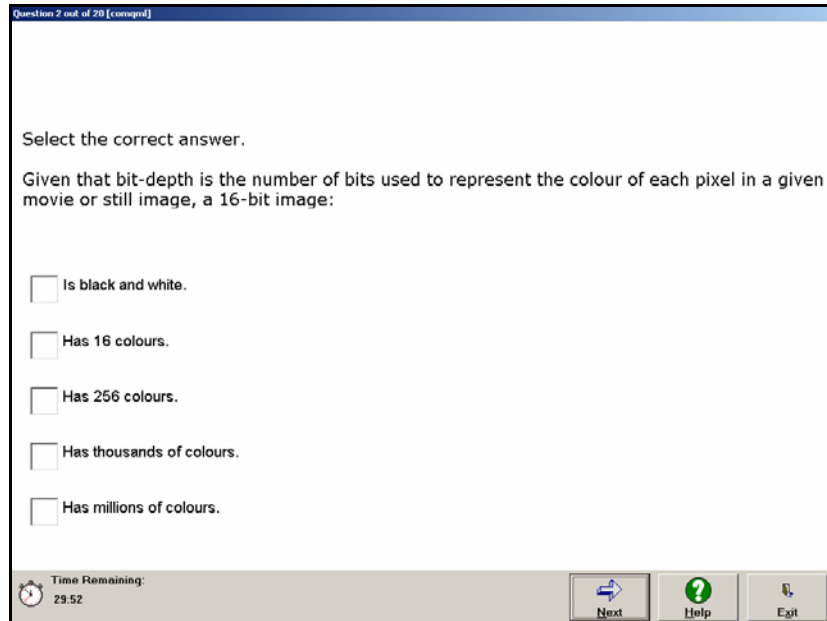


Figure 3: Example of question

Do some independent research on bit depth (the number of bits per pixel allocated for storing indexed colour information in a graphics file). As a starting point, see http://www.microsoft.com/windowsxp/experiences/glossary_a-g.asp#24-bitcolor. See also Chapter 5 from “Principles of Interactive Multimedia”, as section 5.6.4 introduces important aspects related to the use of colour at interfaces.

Figure 4: Example of feedback sentence related to questions regarding bit-depth

For each question a student answered incorrectly, the respective feedback sentence was added to the section named “Based on your test performance, we suggest the following areas for revision”. Figure 5 illustrates one actual feedback document sent to a student. To investigate student attitude towards the feedback format adopted, we invited all students who took Test 3 to tell their views on the feedback format adopted.

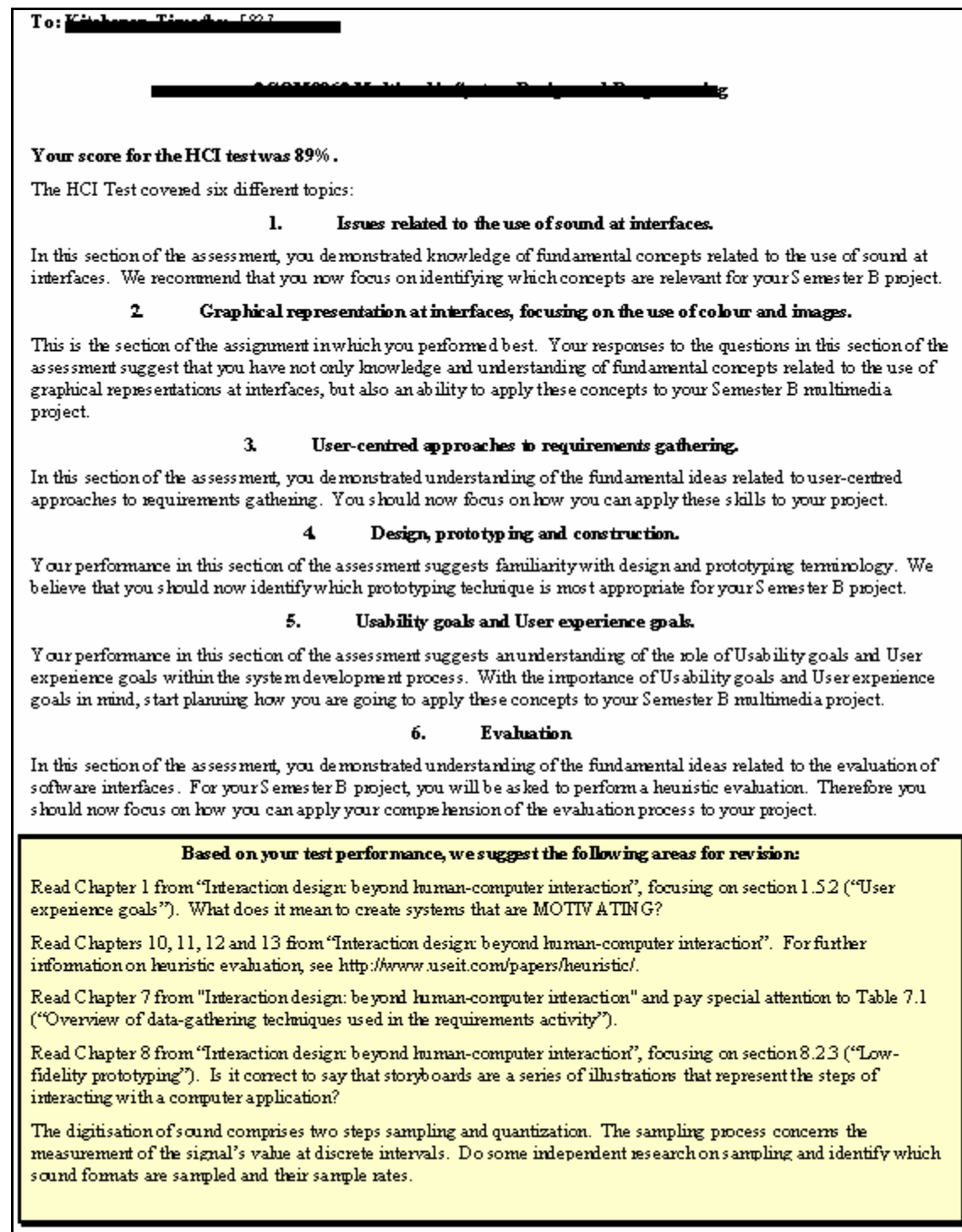


Figure 5: Example of automated feedback generated by the application described in this study

Student attitude towards the feedback format used

In an email, we invited all students who participated in Test 3 to express their views on the feedback format used. Fifty-eight students replied to our email.

In our initial email, we asked the students to classify the feedback received as "very useful", "useful" or "not useful". Students were also asked to present one positive and one negative aspect of the feedback provided. Student answers are summarised in Tables 6, 7 and 8.

Positive aspects about the automated feedback	Very useful	Useful	Not useful	Total
Clear and easy to read; section breakdown according to topic areas	7	6	0	13
It is more helpful than score only	2	2	0	4
It was sent via email	1	2	0	3
Provides clear indication on points for revision	19	18	0	37
None	0	1	0	1
Total	29	29	0	58

Table 6: Summary of positive aspects of the automated feedback according to students

Negative aspects about the automated feedback	Very useful	Useful	Not useful	Total
Meaning of word "adequate" is not clear	0	1	0	1
Copy of the test questions were not provided	5	9	0	14
Impersonal, prefers face-to-face	1	0	0	1
Document type or layout	8	6	0	14
Insufficient personalisation	2	4	0	6
Does not provide clear indication on how many questions answered correctly per topic area	3	0	0	3
Does not include a summary of all scores so far	0	1	0	1
None	10	8	0	18
Total	29	29	0	58

Table 7: Summary of negative aspects of the automated feedback according to students

It can be seen from Table 8 that 14 students considered the document type or layout to be one of the limitations of the feedback format used. Table 8 summarises the limitations of the format used according to these students.

Negative aspects about the format used	Very useful	Useful	Not useful	Total
Feedback provided was too long	2	2	0	4
Marks are at the top of the document rather than at the bottom	3	0	0	3
Paragraph alignment for topic area headings is centred rather than left	2	1	0	3
Favour text only (TXT) rather than word (DOC) format	0	1	0	1
Favour Portable Document Format (PDF) rather than Word (DOC)	1	2	0	3
Total	8	6	0	14

Table 8: Summary of problems with document layout and/or type

Discussion and future work

Like Denton (2003), it is our belief that the potential benefits of automated feedback have not yet been fully explored by academic staff, even by those who are already making use of computer-assisted assessment tools.

In this paper, we present our initial ideas on how Item Response Theory (IRT) can be used to provide students with personalised, meaningful feedback. In summary, the prototype application introduced in this study comprised an ability estimation algorithm based on the Three-Parameter Logistic Model from IRT and a feedback sentences database. Feedback sentences were selected from the latter database based on the ability level estimated and questions answered incorrectly. For each individual student only those sentences that applied to his or her test performance were selected. These selected feedback sentences were then added to a new Word document and sent to his or her personal email account.

The importance of feedback and reflection upon performance has been emphasised by Felder and others (Felder, 1993; Felder & Brent, 1994; Freeman & Lewis, 1998). They have shown that not only is feedback important in formative assessment, but it is also important for motivation and engagement for learners. Strange as it may seem on occasions, learners like to be assessed and value comments on their performance. The investment of effort by learners necessitates comment from teachers. People like to work

for other people. As class sizes increase and more use is made of online formative and summative assessment methods, it becomes increasingly difficult to provide individual feedback in HE. At the very least we have shown that our automated feedback method identifies areas of weakness and strength and provides useful advice for individual development. Student attitude to this approach was positive in general. Students still value a human contribution to feedback, but they also realise that this is becoming rarer in their academic lives.

We are planning to develop the work presented here in several ways. Firstly, by creating one distinct feedback sentence per question. It is anticipated that these sentences should resemble the actual question more than the current comments do. In so doing, we expect to address one of the concerns expressed by some learners in this study (e.g. "would it be possible to attach the question and the correct answers from the test?").

Secondly, it is envisaged that the overall layout of the document will be reviewed in order to facilitate the location of information on the feedback sheet. This is due to the fact that some learners reported that they did not intuitively locate their overall score in the feedback document. The distribution of the feedback document as a PDF rather than Word (DOC) file is also being considered.

To increase personalisation of the feedback, we are intending to compare learner performance in previous assessments with his or her performance in the most recent (i.e. current) assessment. This strategy is likely to provide students with more meaningful and personalised information on their progress than that offered at present.

Finally, in terms of the CAT prototype developed for this study, there was an assumption that performance in one topic area within a subject domain is the best indicator of performance in a related topic area in the same domain. This was practically important for us in deciding the value of student ability θ in order to present the first question in the new topic area. It was, in our opinion, a reasonable working assumption, especially as Bloom's levels are thought to be relatively stable for individuals and our CAT levels were based on these. The assumption, however, needs to be investigated more fully. It is possible that students might have differing abilities in quite similar topic areas. In this case, if tests are short, then participants may not achieve an appropriate CAT level in that area. Making tests longer reduces the efficiency of the test and requires larger question banks. A future focus of our work will be to investigate the starting conditions for topic areas within a CAT and changes in standard error for an individual within an area.

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*. New York: Longman.
- Barker, T. & Lilley, M. (2003). Are Individual Learners Disadvantaged by the Use of Computer-Adaptive Testing? In *Proceedings of the 8th Learning Styles Conference*. University of Hull, United Kingdom, European Learning Styles Information Network (ELSIN), pp. 30-39.
- Carlson, R. D. (1994). Computer-Adaptive Testing: a Shift in the Evaluation Paradigm. *Journal of Educational Technology Systems*, 22(3), pp 213-224.
- Denton, P. (2003). Evaluation of the 'Electronic Feedback' marking assistant and analysis of a novel collusion detection facility In *Proceedings of the 7th Computer-Assisted Assessment Conference*. Loughborough University, United Kingdom, pp. 127-134.
- Felder, R. M. & Brent, R. (1994). *Cooperative Learning in Technical Courses: Procedures, Pitfalls and Payoffs*. NSFDUE Grant DUE-9354379 October 1994.
- Felder, R. M. (1993). Reaching the Second Tier - Learning and Teaching in College Science Education. *JCST* March- April. 286-290.
- Freeman, R. & Lewis, R. (1998). *Planning and implementing assessment*. London: Kogan Page.
- Graduate Management Admission Council (2002). *Computer-Adaptive Format* [online]. Available from <http://www.mba.com/mba/TaketheGMAT/TheEssentials/WhatIsTheGMAT/ComputerAdaptiveFormat.html> [Accessed 21 Mar 2004].
- Harvey, J. & Mogey, N. (1999). Pragmatic issues when integrating technology into the assessment of students In S. Brown, P. Race & J. Bull. *Computer-Assisted Assessment in Higher Education*. London: Kogan Page.
- Jacobson, R. L. (1993). New Computer Technique Seen Producing a Revolution in Educational Testing. *The Chronicle of Higher Education*, 40(4), 15 September 1993, pp. 22-23, 26.
- Lilley, M. & Barker, T. (2002). The Development and Evaluation of a Computer-Adaptive Testing Application for English Language In *Proceedings of the 6th Computer-Assisted Assessment Conference*. Loughborough University, United Kingdom, pp. 169-184.
- Lilley, M. & Barker, T. (2003a). An Evaluation of a Computer-Adaptive Test in a UK University context In *Proceedings of the 7th Computer-Assisted Assessment Conference*. Loughborough University, United Kingdom, pp. 171-182.
- Lilley, M. & Barker, T. (2003b). Comparison between computer-adaptive testing and other assessment methods: An empirical study In *Research Proceedings of the 10th Association for Learning and Teaching Conference*.

The University of Sheffield and Sheffield Hallam University, United Kingdom, pp. 249-258.

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., Bennett, S. & Britton, C. (2002). How computers can adapt to knowledge: A comparison of computer-based and computer-adaptive testing In *Proceedings of the International Conference on Information and Communication Technologies in Education*. Badajoz, Spain: Junta de Extremadura Consejería de Educación, Ciencia y Tecnología, pp 704-708.

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

Microsoft Corporation (2002). *Exam and Testing Procedures* [online]. Available from <http://www.microsoft.com/traincert/mcpexams/faq/procedures.asp> [Accessed 21 Mar 2004].

Pritchett, N. (1999). Effective Question Design In S. Brown, P. Race & J. Bull. *Computer-Assisted Assessment in Higher Education*. London: Kogan Page.

Rafacz, B. & Hetter, R. D (2001). ACAP Hardware Selection, Software Development, and Acceptance Testing In W. A. Sands, B. K. Waters & J. R. McBride. *Computerized Adaptive Testing: from Inquiry to Operation*. Washington, DC: American Psychological Association.

Wainer, H. (2000). *Computerized Adaptive Testing (A Primer)*. 2nd Edition. New Jersey: Lawrence Erlbaum Associates.