EVALUATING THE USER EXPERIENCE IN CAA ENVIRONMENTS: WHAT AFFECTS USER SATISFACTION?

Gavin Sim, Janet C Read & Phil Holifield

Evaluating the User Experience in CAA Environments: What Affects User satisfaction?

Gavin Sim and Janet C Read
Department of Computing
University of Central Lancashire
Preston
PR1 2HE
Tel: 01772 895162
grsim@uclan.ac.uk
jcread@uclan.ac.uk

Phil Holifield
Faculty of Design and Technology
University of Central Lancashire
Preston
PR1 2HE
pholifield@uclan.ac.uk

Abstract

This paper reports the findings of an experiment to establish students' satisfaction with various aspects of the user interface in three Computer Assisted Assessment (CAA) environments. Forty four second year undergraduate students in Human Computer Interaction participated in the study. Each student completed three tests using three different CAA software environments. Through the use of two survey instruments, user satisfaction was measured. The results highlight the fact that, in this instance, scrolling did not seem to influence student satisfaction but other attributes, such as navigational structure and question styles, appear to influence it. The students appeared to prefer different CAA environments depending on whether the context of use was for formative or summative assessment.

Introduction

With the increased adoption of CAA within educational institutions there has been a rise in the number of such systems available. Several of these are designed for use in Higher Education establishments; these include Questionmark Perception, Hot Potatoes, TRIADS and TOIA. These 'bespoke' systems are relatively new to higher education but software delivering multiple choice style questions dates back to the 1970's (Morgan, 1979) and so the concept is quite old. In addition, there are learning management systems, like WebCT and Blackboard, that have CAA tools incorporated into them.

In a commercial marketplace it is important for the vendors of CAA software to attract new customers and then to hold onto their customer base. To attract new custom, vendors often emphasise the 'features' of their products, placing great importance on the number of different question styles available. In common with many other software products, with each new version, more features and more question styles are offered. For example, TRIADS software developed by Derby University offered 17 question styles in 1999 (Mackenzie, 1999) compared to 41 in 2005 (CIAD, 2005). It has been reported that instructors and academics are often unfamiliar with many of these highly sophisticated new question styles and subsequently find it difficult to write questions that take advantage of their features (McLaughlin, Fowell, Dangerfield, Newton, & Perry, 2004).

One user group that has little influence on the design of CAA software is the student population that uses the software for assessment. This group is seldom in a position to choose which CAA software is used and yet their experience of the software is clearly important. User experience is one of the facets of usability which is generally measured by considering the effectiveness of an interface, the efficiency of the system and the user experience (ISO, 1998). It is expected that the user experience of the software would have some impact on the test performance (Bridgeman, Lennon, & Jackenthal, 2002), however, there has been very little research analysing the user experience of CAA and in particular the effect on user experience when more sophisticated questions are introduced into the test environment.

The user experience is often related to the user satisfaction of a system and is concerned with how well the system facilitates the user in achieving their goal. User experience can be ascertained by the use of surveys and observations, that rely to some extent on opinions and judgements, as well as more scientific methods, these include measures of skin sweat rate and heart rate. The most common method for evaluating user experience is, however, the written questionnaire (Johnson, Zhang, Tang, Johnson, & Turley, 2004; Van Veenendaal, 1998).

Using questionnaires to gather user opinions is problematic, studies point to the tendency of individuals to choose random answers, to report what the questioner wanted to hear, and to fail to complete questionnaires (Vaillancourt, 1973). Careful design of questionnaires can reduce these issues, paying attention to the length of the survey as well as the length of the questions and adding questions that test for reliability are known solutions (Breakwell, Hammond, & Fife-Schaw, 2000).

In this study, questionnaires were used as the means to elicit opinions from undergraduate students about the user interface for three CAA applications.

Method

An experiment was devised using three CAA applications that provided between them a variety of interface design characteristics which the users could evaluate. At the outset of the experiment, there were several hypotheses about the impact of certain 'features' of CAA software with respect to user satisfaction. User satisfaction was considered to be affected by the user experience of:

- Accessing and finishing the test
- Navigation within the test
- Visual layout
- Interface for answering questions

In a CAA environment, the goal of the user is to complete the assessment, progression towards this goal requires the completion of several tasks; to start the test, answer the questions, navigate between pages and end the test (Sim, Horton, & Strong, 2004). It was expected that there would be some variation between CAA applications with respect to the above constructs. The purpose was not to identify, or claim, that one application was better than another, merely to examine attributes of the interface that affect user satisfaction. This limitation was necessary as the three applications being considered could be customised to present the tests in different formats and the students were examining the interaction within the environment and so were not using the software to test their knowledge of a specific subject domain.

Choice of CAA Applications

As outlined in the introduction, there are numerous CAA applications. For this study a choice was made to focus on three software applications, S1, S2, and S3. S1 was selected as an example of a CAA application integrated into a Learning Management System (LMS) as an assessment tool. Such tools usually have limited question styles compared to more specialist CAA software, however they are widely used for assessment purposes within Higher Education (Alexander, Bevis, & Vidakovic, 2003; Cooper, 2002; Pretorius, 2004; Sayers & Hagan, 2003).

S2 is a dedicated CAA software application offering a lot more functionality and question styles than learning management systems. Many institutions have adopted such software for formative and summative assessment (Sim, Holifield, & Brown, 2004).

Finally S3 is a CAA software application offering more advanced question styles than the other two applications and is perceived to be more flexible and specialist. A demonstration version of S3 was used exhibiting a variety of sophisticated question styles.

Software Set Up

For S1 (see Figure 1), the test was set up so that all the questions were displayed on the screen at once and three question styles were used; Multiple Choice, Multiple Response and Text Entry.

For S2 the test was set up using question by question delivery and incorporated the following question styles; Multiple Choice, Multiple Response, Order, Text Entry, Matrix and Drag and Drop.

Finally within S3 four sections of the demonstration were selected to be used which incorporated a variety of sophisticated question styles such as drawing lines, assertion reason and matrix.

Figure 1: Screen shots of the three software used (from left to right S1, S2, S3)

Survey Design

The study used two survey tools. The first (Q1) was a questionnaire adapted from an earlier version (Sim & Holifield, 2004) which had previously been used to examine user satisfaction with the interface of a CAA software application. Additional questions were included in Q1 to examine the effectiveness of the software in facilitating the user in achieving their goal.

This questionnaire (Q1) consisted of 13 Likert style questions and was divided into four sub-sections. To minimize acquiescence, the tendency by some of a sample to consistently agree or disagree with a set of questions (Bryman, 2004), a mixture of positive and negative statements were incorporated into the design. There was also the opportunity for students to provide qualitative data with regards to specific features they liked about the interface.

The second survey instrument (Q2) was a variation on a repertory grid (Fransella & Bannister, 1977) and loosely based on an instrument that is used for children to measure fun (Read, MacFarlane, & Casey, 2002). This was presented to the students one week after completing the evaluations of the three applications and it required the participants to rank each application according to nine constructs. This survey also included two questions that required the students to identify which of the CAA applications would be their preferences for formative and summative assessment.

Apparatus

The students conducted the first part of the experiment in three different labs using networked PCs with flat screen monitors, full size keyboards and scrolling mice. In each lab, the hardware specification was the same.

Participants

The students that took part in the study were a convenience sample taken from an undergraduate class in HCI. A total of 44 participated in the experiment, but only 25 completed the second survey (Q2). This class comprised students from seven different computing courses and therefore had a wide range of different 'types' of student for example, networking and software engineers. The sample was predominantly male and approximately 5% of the sample did not have English as their first language. The participants did not receive any payment for taking part in the study but a draw was made at the end of the experiment and the lucky winner got a free text book. Participation was voluntary but some may have felt it was a part of their class as it took place in class time.

Procedure

The evaluation of the CAA applications took place on a single day at a single time in three identically equipped computer labs. In these labs, students worked through a series of questions in the 3 applications. The order in which they met the three packages was counterbalanced to remove any learning effects that might otherwise have affected the results. Thus, in one lab everyone started with S1, in another everyone started with S2 and in the third, everyone started with S3. The S1 application had 17 questions on football, S2 had 17 questions on Films and S3 used the default questions from the online test interface which included topics such as geology and maths. Students worked through the three applications in their own time (but were supervised). They were able to move through the three applications in their own time but had to stick to the pre determined order. As each student completed a single application, they completed the questionnaire Q1.

For the post hoc study, students were given the repertory grid activity Q2 and asked to complete it. This was done in a class a week after the initial experiment. It was not possible to link these results to the results from the experiments.

Analysis

The first questionnaire, Q1, completed after the test was scored in an ordinal way 1-5, where 5 represented Strongly Agree and 1 Strongly Disagree. If the question was negatively worded then the scoring was reversed.

The Repertory Grid (Q2), completed the week after the initial experiment, was again coded in an ordinal manner using 1-3 for each of the criteria. The last two questions on the sheet “Which of the three would you choose for: an end of year exam” and “Which of the three would you choose for: Revision purposes” were tallied according to how many students selected that software.

Friedman tests were conducted to establish whether there were any significant differences between the three software applications and Wilcoxon post-hoc tests were then preformed to determine where the difference lay.

Results and Discussion

As the results reported in this paper are predominantly gleaned from the survey instruments, a test of reliability was carried out on the major instrument, Q1; the alpha reliability of the scale is 0.888.

In Q1, the students were asked whether they had any prior experience of using the software. From the 44 participants, 17 had prior experience of S1, 20 had experience of using S2, and only 2 had used S3 before. A Mann-Whitney U Test was conducted between those who had prior experience and those without for S1 and S2. There was no significant difference between the two groups on any of the questions, therefore prior experience does not seem to influence there satisfaction of a CAA environment.

The mean scores relating to the participants answers for Q1 are displayed in table 1. Overall on the majority of questions they reported a level of satisfaction with each of the three CAA environments.

No	Question	S1	S2	S3
1	I had no problem gaining access to the test	4.21	3.84	3.53
2	I encountered difficulties starting the test	4.28	3.95	3.40
3	The interface required too much scrolling	3.44	4.19	3.81
4	The amount of scrolling was acceptable	3.37	3.95	3.60
5	It was difficult to read the text on the screen	3.86	3.84	3.09
6	The screen layout was clear	3.88	3.67	2.56
7	The screen layout was consistent	4.12	4.02	2.72
8	I liked the way the test looked	3.49	3.33	2.35
9	I would have preferred an alternative font	3.40	3.23	2.86
10	The button names are meaningful	4.02	3.88	3.33
11	I always knew where I was within the software	4.02	4.05	2.23
12	The navigation was logical	4.05	3.84	2.65
13	The navigation was clear	3.95	3.77	2.58

Table 1: The mean scores for the first questionnaire for each of the three software applications

The results from the REP grid (Q2) which was administered a week after Q1 are displayed in table 2 below.

Student Ranking	S1	S2	S3
Login	18	6	1
Navigation	12	11	2
Layout	6	16	3
Scrolling	6	8	11
Reading	12	11	2
Instructions	9	14	1
Input Answer	11	11	3
Change Answer	11	9	0
Finish Test	12	12	1

Table 2: Frequency each piece of software was ranked first by the user on a number of criteria

Accessing and Finishing the Test

The first two questions in Q1 refer to the students gaining access and starting the test. Although the mean scores suggest overall there was little difficulty in accomplishing this task S3 was significantly different to S1 for the first question ($Z=-2.882$, $p>0.01$) and S3 was significantly different to both S1 ($Z=-3.987$, $p>0.001$) and S2 ($Z=-2.293$, $p>0.05$) for the second question.

Similar results were obtained in Q2 with S3 appearing quite different from the other two as only one student ranked it first. In addition, in this survey a post hoc Wilcoxon revealed a significant difference between S1 and S2 ($Z=-2.562$, $p<0.01$). The high scores for S1 could have been due to the fact that the majority of students access the LMS for teaching material for their modules and so the look, if not necessarily the test environment, was familiar to them. These differences may have also been as a consequence of the amount of interaction that is required before the user gets to the first question: S1 and S2 both required 5 tasks whilst S3 required 6.

Using Q2 the students were asked about how easy it was to end the test and only one student ranked S3 the easiest whilst S1 and S2 were both ranked easiest by 12 students. This may be because of the amount of interaction for exiting the test was higher in S3 than the other two applications.

Visual Layout

Both S1 and S3 incorporated scrolling in the user interface, in S1 the questions were all displayed on the screen and for S3 the scrolling was in the instructions and results. In Q1 there were two questions that examined the effects scrolling had on user satisfaction. For question 3, S2 was significantly different to both S1 ($Z=-3.473$, $p<0.01$) and S3 ($Z=-2.215$, $p<0.05$) and a similar result was obtained for question 4. However, Q2 revealed no significant difference between the three software in relation to scrolling.

The three applications all used different font types and sizes and this was presumed to affect legibility. In the first survey question 5 asked about the legibility of the text and in the answers to this, S3 was found to be significantly different to both S1 ($Z=-3.007$, $p<0.01$) and S2 ($Z=-3.15$, $p<0.01$) and similar results occurred for question 9. Q2 also asked about legibility and it was also found that S3 was scored lower than S1 and S2. This perception about the legibility of the text within S3 may have been because, due to this application being evaluated with the ready made questions rather than the simple style questions used in S1 and S2, there was a lot more text in both the questions and the feedback than in S1 and S2. This, coupled with scrolling which is a known factor that effects on screen legibility, could have led to the poor result for S3 (Bernard, Chaparro, Mills, & Halcomb, 2003).

Questions 6, 7 and 8 in Q1 also related to the layout of the screen and again satisfaction with S3 was significantly lower than S1 and S2. For example question 6, a Wilcoxon test revealed that S3 was significantly different to S1 ($Z=-4.463$, $p<0.001$) and S2 ($Z=-4.337$, $p<0.001$) with similar results found for questions 7 and 8. These findings were all supported by the results from Q2 where S3 was ranked lower than both S1 and S2. This may have been attributed to the fact that each question in S3 used a different style and therefore there was no continuity in the interface compared to the other applications.

Navigation

The final four questions in Q1 related to the navigation of the CAA software applications and again there were differences between them. For question 10 S3 was significantly lower than S1 ($Z=-3.018$, $p<0.01$) and S2 ($Z=-2.485$, $p<0.05$) this was also found to be the case for the other three questions relating to navigation.

The results from Q2 in relation to navigation revealed a significant difference $\chi 2=21.68$, $p<0.001$ and post hoc tests revealed that S3 was ranked significantly lower than S1 ($Z=-3.273$, $p<0.01$) and S2 ($z=-3.855$). There was no difference between the navigation of S1 and S2. The low results for S3 may have been due to the linear navigational structure, students being required to select an option then work through the questions in order. There was little freedom to move between questions or skip a question and return to it later.

Answering the Questions

Q2 asked the students about inputting an answer and there was again a significant difference between the three applications $\chi^2=19.76$, $p<0.001$. The post hoc test revealed there was no difference between S1 and S2 however, S3 was significantly lower than S1 ($Z=-3.855$, $p<0.001$) and S2 ($Z=-2.805$, $p<0.05$). These results were similar to the results relating to instructions and it is possible that because the level of interaction was more complex, students found the process of answering questions more difficult within S3.

Preference for Software Depending on Context

The final two questions in the survey asked the students which software they would choose for summative and formative assessment, the results are shown in Table 3.

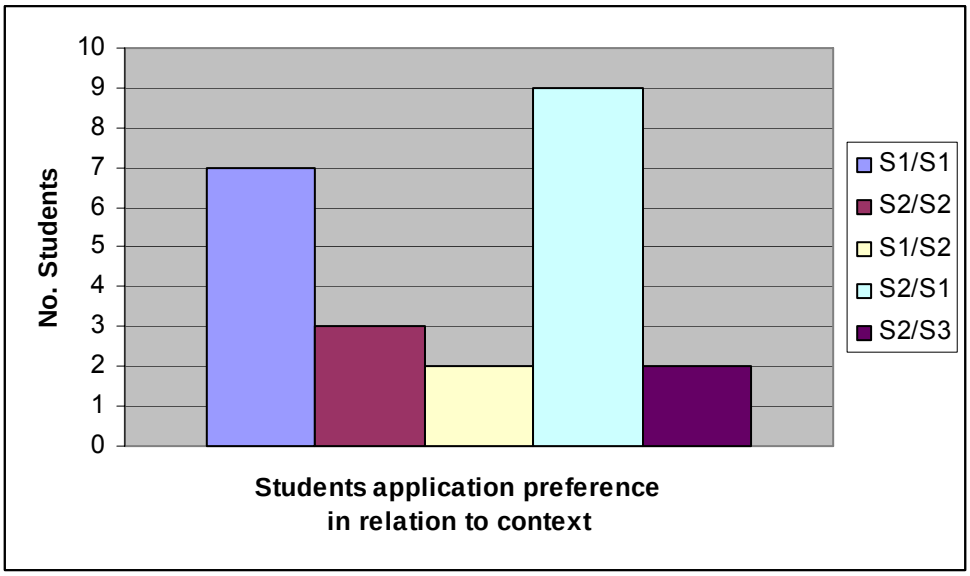

Table 3: Students application preference in relation to context

Of the 23 students completing this section only 10 stated they would use the same application for both contexts. The remaining 13 had a different preference depending on the context of the assessment. For example, 9 students stated their preference for summative assessment would be S2 and S1 for formative assessment. This would suggest the nature of assessment also influences students' perception about the suitability of a CAA environment and it is not just simply looking at the interface attributes.

Conclusions and Further Work

For developers of CAA environments or academics customising templates, this research has highlighted a number of interface characteristics that affect user satisfaction within a CAA environment. For S1 and S2 prior experience had no bearing on user satisfaction, it was not possible to examine this for S3 due to the limited number of students who had prior experience. It may be that

for more complex interaction prior experience is necessary to improve overall satisfaction as there is a greater learning curve.

There does not appear to be a single attribute that influences students' preference for a particular CAA environment. Other research has highlighted scrolling as an attribute that affects students attitude (Ricketts & Wilks, 2002) but in this study S1 required the most scrolling yet students indicated that they would still select this system for formative or summative assessment. With regards to navigation, students appeared to prefer the ability to navigate freely and were less satisfied with the linear structure presented in S3.

Increasing the number of question styles did not seem to affect attitudes between S1 and S2, however the complexity of the questions within S3 may have affected the students satisfaction. Further work may be needed to determine whether there is a complexity threshold within CAA environments in relation to question styles, and if so, whether once this threshold is passed, there is a related decline in overall user satisfaction.

When selecting and evaluating a CAA environment, context appears to be a significant factor that needs to be considered. Students appear to prefer different systems depending on whether the software is being used for formative or summative assessment. In this study, the majority of students selected S1 for formative assessment but this may be because they associate this application (which is part of a LMS) with their learning, considering S2 and S3, both more specialised and more suitable for assessment. There was a mixed response in relation to summative assessment with students opting to use either S1 or S2.

This study has highlighted the complexity of trying to do a comparative study of three CAA environments. In this instance it was not possible to customise S3 as a demo version was used; this undoubtedly influenced the results as apportioned to individual applications and so the results presented here cannot be used to indicate a preference or otherwise for a particular application. The intention of the study was to examine the interactions within general CAA environments.

There are several extensions to this work, it would be useful to ask students why they chose a particular application for formative and summative assessment, to determine what features they consider to be the most necessary and to investigate the effects of multiple question styles.

References

Alexander, M., Bevis, J., & Vidakovic, D. (2003). *Developing Assessment Items using WebCT.* Paper presented at the World Conference on E-Learning in Corporations, Government, Health and Higher Education, Phoenix.

Bernard, M. L., Chaparro, B. S., Mills, M. M., & Halcomb, C. G. (2003). Comparing the effects of text size and format on the readability of computer-displayed Times New Roman and Arial text. *International Journal of Human-Computer Studies, 59*(6), 823-835.

Breakwell, G. L., Hammond, S., & Fife-Schaw, C. (2000). *Research methods in psychology* (second ed.): Sage.

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2002). *Effects of Screen Size, Screen Resolution and Display rate on Computer-Based Test Performance.* Paper presented at the Annual meeting of the national council on measurement in education, New Orleans.

Bryman, A. (2004). *Social Research Methods* (Second Edition ed.). Oxford: Oxford University Press.

CIAD. (2005). *TRIADS Question Styles*. Retrieved 22/11/05, 2003, from http://www.derby.ac.uk/ciad/triadstyles.html

Cooper, C. (2002). *Online Assessment using Blackboard an issue paper*: University of Wales Institute Cardiff.

Fransella, F., & Bannister, D. (1977). *A manual for repertory grid technique*. London: Academic Press.

ISO. (1998). *Ergonomic requirements for office work with visual display terminals (VDTs) -- Part 11: Guidance on usability*: ISO 9241-11.

Johnson, T., R., Zhang, J., Tang, Z., Johnson, C., & Turley, J., P. (2004). Assessing informatics students satisfaction with a web based courseware system. *International Journal of Medical Informatics, 73*(2), 181-187.

Mackenzie, D. (1999). *Recent Developments in the Tripartite Interactive Assessment Delivery System (TRIADS)*. Retrieved 13/06/02, 2002, from http://www.derby.ac.uk/ciad/lough99pr.html

McLaughlin, P. J., Fowell, S. L., Dangerfield, P. H., Newton, D. J., & Perry, S. E. (2004). Development of computerised assessments (TRIADS) in an undergraduate medical school. In D. O'Hare & D. Mackenzie (Eds.), *Advances in computer aided assessment* (pp. 25-32). Birmingham: SEDA.

Morgan, M. R. J. (1979). MCQ: An interactive computer program for multiple-choice self testing. *Biochemical Education, 7*(3), 67-69.

Pretorius, G. (2004). *Objective testing in an E-Learning Environment: a Comparison between two systems.* Paper presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Lugano.

Read, J. C., MacFarlane, S. J., & Casey, C. (2002). *Endurability, Engagement and Expectations: Measuring Children's Fun.* Paper presented at the Interaction Design and Children, Eindhoven, The Netherlands.

Ricketts, C., & Wilks, S. J. (2002). Improving Student Performance Through Computer-Based Assessment: insights from recent research. *Assessment & Evaluation in Higher Education, 27*(5), 475-479.

Sayers, H. M., & Hagan, N. S. J. (2003). Supporting and Assessing First Year Programming: The use of WebCT. *Italics, 3*(1), 1-11.

Sim, G., & Holifield, P. (2004). *Piloting CAA: All aboard.* Paper presented at the 8th International Computer Assisted Assessment Conference, Loughborough.

Sim, G., Holifield, P., & Brown, M. (2004). Implementation of computer assisted assessment: lessons from the literature. *ALT-J, 12*(3), 215-229.

Sim, G., Horton, M., & Strong, S. (2004). *Interfaces for online assessment: friend or foe?* Paper presented at the 7th HCI Educators Workshop, Preston.

Vaillancourt, P. M. (1973). Stability of children's survey responses. *Public opinion quarterly, 37*, 373-387.

Van Veenendaal, E. (1998). *Questionnaire based usability testing.* Paper presented at the European Software Quality Week, Brussels.