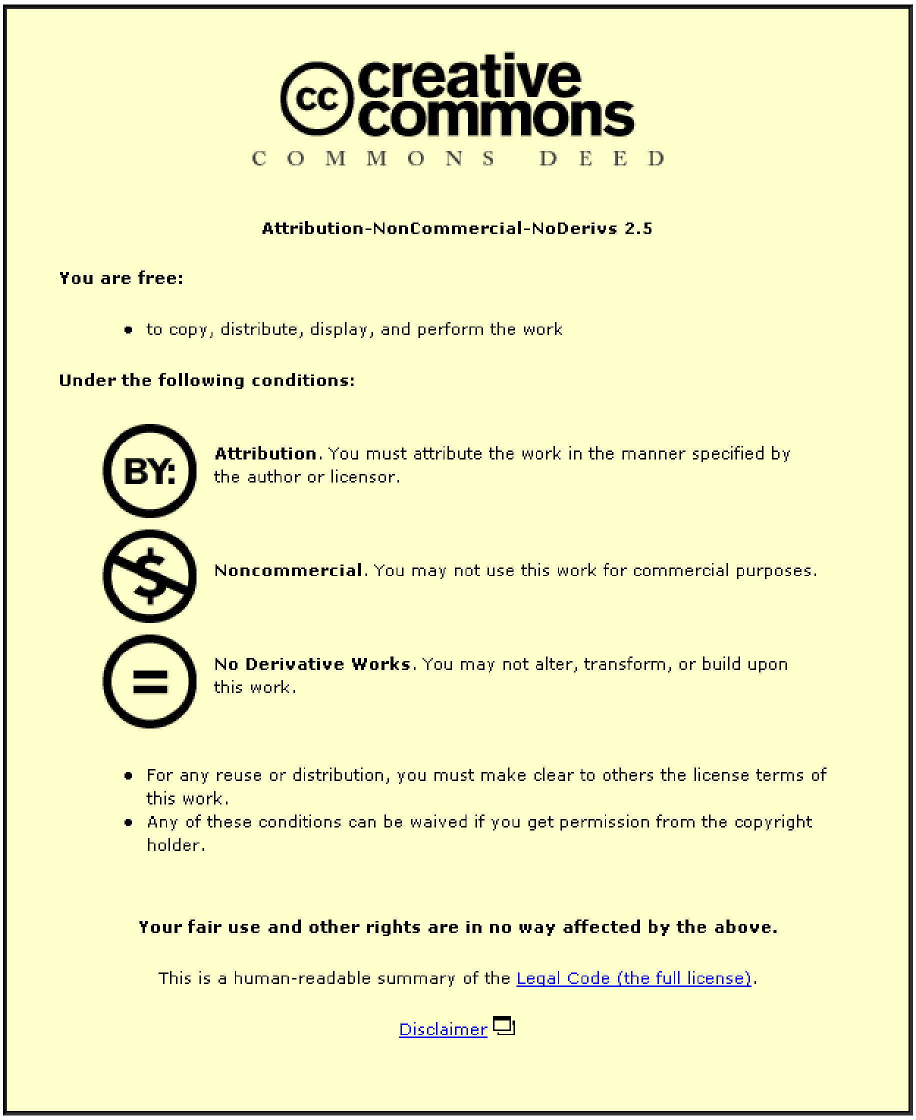
View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Loughborough University Institutional Repository

This item was submitted to Loughborough's Institutional Repository (https://dspace.lboro.ac.uk/) by the author and is made available under the following Creative Commons Licence conditions.



creative commons

COMMONS DEED

Attribution-NonCommercial-NoDerivs 2.5

You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

Attribution. You must attribute the work in the manner specified by the author or licensor.

Noncommercial. You may not use this work for commercial purposes.

No Derivative Works. You may not alter, transform, or build upon this work.

- For any reuse or distribution, you must make clear to others the license terms of this work.
- Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/

HEURISTICS FOR EVALUATING THE USABILITY OF CAA APPLICATIONS

Gavin Sim, Janet C Read and Phil Holifield

Heuristics for Evaluating the Usability of CAA Applications

Gavin Sim
School of Computing Engineering and Physical Sciences
University of Central Lancashire
Preston
PR1 2HE
grsim@uclan.ac.uk

Janet C Read
School of Computing Engineering and Physical Sciences
University of Central Lancashire
Preston
PR1 2HE
jcread@uclan.ac.uk

Phil Holifield
Faculty of Arts, Humanities and Social Sciences
University of Central Lancashire
Preston
PR1 2HE
pholifield@uclan.ac.uk

Abstract

Evaluation of usability is well researched in the area of HCI. One widely used method is a heuristic evaluation which relies on a small number of evaluators inspecting an interface to see to what extent it complies with a set of heuristics. Once a problem is identified it is categorised to a heuristic and a severity rating is attached. Severity ratings indicate the potential impact of the problem.

Using a corpus of usability problems within CAA this paper reports on the development of domain specific heuristics and severity ratings for evaluating the usability of CAA applications. The heuristics are presented and the paper concludes with practical guidance on the application of the method in CAA.

Introduction

Computer Assisted Assessment (CAA) is the use of computers to deliver, mark or analyse exams. Over the past decade there has been a rise in the number of commercial and bespoke CAA applications available yet little research has been conducted on the usability of these applications. ISO 9241-11 defines usability as the extent to which a product can be used by specific users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use (ISO, 1998). Whilst usability problems have been shown to exist in CAA applications (Sim, Read, Holifield, & Brown, 2007), and research has reported on user satisfaction within CAA applications (Ricketts & Wilks, 2002) , the other two components of the ISO definition have been largely overlooked.

Within CAA it is questionable whether these components (effectiveness, efficiency and satisfaction) are appropriate measures of usability as other constructs may be required. It has been argued that the focus of HCI should be on understanding what is 'valued' by a stakeholder with the result that good products would support the stakeholder in delivering this value (Cockton, 2004). There are many different stakeholders when evaluating CAA, for example, the academics may value the time saving benefits of automated marking whilst in summative assessment students have a tendency to value the mark achieved, as apposed to any accompanying feedback(Brookhart, 1997). Ultimately the students have the most to lose as a consequence of poor usability therefore it can reasonably be argued that usability research should focus on those severe problems that may cause the user difficulties and dissatisfaction and result in unacceptable consequences. Within the context of higher education severe problems might be those that give the student grounds for appeal at an examination board. For example they may suffer a loss of time through ineffective navigation or the inability to deselect an answer could result in the question being marked wrong. The most severe consequence would be a loss of the results, however within a formative context this may be deemed less severe. These problems could be inherent within the CAA application or a consequence of poor pedagogical practices.

Evaluation of usability is well researched in the area of HCI and the methodologies are widely understood. There are two main methods; user tests and inspection methods. When evaluating the usability of CAA applications within the context of summative assessment, some user based methods are unacceptable for ethical reasons. For example, observation may be too intrusive whilst someone is conducting a summative test, think aloud (Ericsson & Simon, 1993) may be impractical as the user would not focusing on answering the questions and empirical research, comparing several interfaces, may result in interface related variance thus affecting (unfairly) the students' performance. Therefore inspection methods may be a suitable method for evaluating CAA applications.

One method that has been widely applied is the heuristic evaluation (HE) first brought to prominence by (Nielsen & Molich, 1990). The general purpose usability heuristics devised by Nielsen (1994) are the most citied and applied. The HE method uses a small number of experts, usually between three and five, to evaluate the interface against a set of heuristics. Severity ratings are also attached that indicate the potential impact of the problem. The severity ratings that are generally used were devised by Nielsen & Mack (1994) and are:

- 0= I don't think that this is a usability problem
- 1= Cosmetic problem only: need not be fixed unless extra time is available on the project
- 2= Minor usability problem: fixing this should be given low priority
- 3= Major usability problem: important to fix, so should be given high priority
- 4= Usability catastrophe: Imperative to fix so should be given high priority

Within CAA these severity ratings may be too generic as it is difficult to accurately distinguish what constitutes a 'Major Usability Problem' and a 'Usability Catastrophe' (Sim, Read, & Holifield, 2006). Domain specific severity ratings are therefore required which relate to the unacceptable consequences that the user may encounter when using a CAA application.

Nielsen's heuristics have come under increasing criticism in recent years for their poor effectiveness in certain domains. This has lead to an increase in the development of domain specific heuristics, for example accessibility heuristics (Paddison & Englefield, 2004) and heuristics for the playability of games (Desurvire, Caplan, & Toth, 2004). This paper looks at the development of domain specific heuristics and severity ratings for CAA.

Method

A corpus of usability problems within a number of CAA applications was gathered through a series of evaluations (Sim, Horton, & Strong, 2004; Sim et al., 2006; Sim et al., 2007) and through an analysis of the literature (ISO/IEC23988, 2007; Sim, Holifield, & Brown, 2004). A filtering process was applied to the corpus to extract only the frequent or most severe problems. This corpus was used for the synthesis of CAA heuristics to ensure that the heuristic set would adequately represent the domain.

The tasks the user would perform were identified by (Sim, Horton et al., 2004). as:

- starting the test
- navigating
- answering the questions
- finishing the test

These tasks were used as the initial starting point for the development of the heuristics along with an analysis of Nielsen's heuristic set (Nielsen, 1994). Each of the problems within the corpus was mapped to a task and groupings were formed that enabled the synthesis of the heuristic (see table 2).

Task	**Problem**	**Consequence**	**Heuristic**
Finishing the test	P1. Accidentally finished the test	Lost results	Inform users of any unanswered questions before finishing
	P2. When all questions attempted finish appears. It exits without confirmation & doesn't check whether any flags are still set	Some questions may be left unanswered	
	P3. If trying to finish early error message doesn't inform which questions haven't been answered, could be clearer	Some questions may be left unanswered	
	P4. You could finish the test and submit your answers even if some questions hadn't been attempted - should have prompted you		

Table 1: Reported usability problems mapped to a heuristic

Two of the problems (P3 and P4) reported in table 1 refer to the same issue and were identified in two different evaluations. In this example, it can be seen that the problems reported were then used to formulate the heuristic '*Inform users of any unanswered questions before finishing*'.

In some instances problems were reported that did not belong to a specific task for example '*Staring at the screen for two hours is painful*'. A heuristic was then initially described for this which was originally called *'Environment'* and then, any further problems which fitted this category were classified to this heuristic. It later became apparent that the heuristic '*Environment*' was too generic and therefore it was modified to '*Minimise external influences to the user'*. At this stage it was important to ensure that the problems already classified would fit the new heuristic and therefore these had to be re-examined.

Heuristics

Using the process outlined above a set of heuristics for evaluating CAA applications was synthesised, these are presented in table 2.

Heuristics	**Description**
1. Use clear language and grammar within questions and ensure the	Text should be grammatically correct and make sense. It should be obvious to the

score is clearly displayed	user what the score is for a particular question and the scoring algorithm applied (e.g. if negative marking is used).
2. Ensure progress through the test is visible and understandable	Ensure that the number of questions answered and remaining is obvious and time remaining is clear.
3. Answering questions should be intuitive	Clear distinction between question styles and the process of answering the question should not be demanding. Answering the question should be matched to interface components.
4. Easy reversal of actions	It should be possible to change or remove an answer. Ensure it is possible to return to an incomplete test or question.
5. Inform users of any unanswered questions before finishing	If a user has opted to end the test ensure that they are informed of any unanswered questions.
6. Ensure appropriate interface design characteristics	Interface should match standards and design should support user tasks.
7. Visual layout - adequate spacing and visibility of questions	Ensure that there is enough spacing between the elements within the interface and scrolling is minimised within the questions.
8. Ensure appropriate feedback	System feedback should be clear about what action is required. Question feedback should assist the learning process.
9.Moving between questions and terminating the exam should be intuitive	User input to navigating between questions and returning to unanswered questions should be consistent. Options to exit should be identifiable.
10. Minimise time delays	Prevent any unnecessary delays. Ensure that there is minimal latency when moving between questions or saving answers.
11. Minimise external influences to the user	Ensure test mode does not impact on fairness and performance within the test. Prevent distractions to other users and do not penalise them due to constraints of the software e.g. spelling mistakes (unless essential)

Table 2 CAA Heuristics

These heuristics can be used by academics or educational technologists to perform a heuristic evaluation and to establish the appropriateness of the CAA application ensuring that severe problems are minimised.

Procedure

The procedure for using the heuristic set is as follows:

- Stage 1 – Recruit 5 or 6 evaluators
- Stage 2 – Train the evaluators in the use of the heuristics, in the context of the application and in HE (if not already familiar with this)
- Stage 3 – Carry out the HE
- Stage 4 – Collate results
- Stage 5 – Recommend actions

The evaluators chosen will have an impact on the results. It is known (Jacobsen & John, 1998; Slavkovic & Cross, 1999) that there is an evaluator effect and this is in the main a result of varying levels of expertise. For instance, if the evaluators are very familiar with CAA they will bring domain knowledge that influences the results, it is often suggested that the evaluators should simply be expert in the method of heuristic evaluations rather than in the domain but the recruitment of double experts, as suggested by Nielsen (Nielsen, 1994) is the ideal.

Context knowledge can be improved by having the evaluators work through an example test. Creation of an example test for the evaluators should be a priority and should reflect the sorts of question styles that real users would encounter. In addition, some information about context is widely recommended, e.g. where would the application be used, what would happen to the results from the test, under what circumstances would users use the application etc.

To carry out the HE, the evaluators can be given a method and this can be conveyed using special reporting forms (Woolrych & Cockton, 2002). The design of the reporting form can also affect the results from the HE, some forms are better suited than others at assisting evaluators to both find problems and map them to heuristics. The method suggested here is that evaluators find and log problems (noting where the problem was encountered), then map these problems to one or more heuristics whilst also indicating a severity rating for each problem.

A suggested set of severity ratings for CAA are given:

- 0= I don't think that this is a usability problem
- 1= Possible effect, could cause some users to perform less well than would have performed otherwise
- 2= Minor effect, would probably affect one or more questions in the test for most users.
- 3= Major effect, would probably affect many questions in the test for most users.
- 4= Catastrophe: all work lost

The evaluators come together and aggregate their individual problem sets. During this discussion, evaluators are able to see more clearly the extent and the causes of problems and at this point there will be valuable discussions about the severity of each problem. In resolving differences, the evaluators come to shared understanding, especially in regard to the serious problems.

Finally, the problems are ranked in order of severity and, subject to their expertise, the evaluators may be able to suggest some fixes (although this is normally carried out by a different group of people).

The value of apportioning problems to the heuristics may appear small but it is the process of thinking through problems and categorising them that makes heuristic evaluation a powerful tool.

Conclusion

This work has resulted in a set of heuristics for CAA being produced along with a new set of severity ratings. Currently work is being undertaken to validate the heuristic set and to see how easy the set is to use.

Further research will be conducted to establish the effectiveness of these heuristics within the CAA domain. This may lead to modification or expansion of the original heuristic set presented here.

References

Brookhart, S. M. (1997). A theoretical framework for the role of classroom assessment in motivating student effort and achievement. *Applied Measurements in Education, 10*(2), 161-180.

Cockton, G. (2004, October). *Value-centred HCI.* Paper presented at the NordiChi, Tampere.

Desurvire, H., Caplan, M., & Toth, J. A. (2004). *Using heuristics to Evaluate the Playability of Games.* Paper presented at the CHI, Vienne.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis. Verbal reports as data*. Cambridge: MIT Press.

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

ISO/IEC23988. (2007). *Information technology - A code of practice for the use of information technology (IT) in the delivery of assessments* (No. ISO/IEC23988).

Jacobsen, N. E., & John, B. E. (1998). *The evaluator effect in usability studies: problem detection and severity judgements.* Paper presented at the Proceeding of the Human Factors and Ergonomics Society 42nd Annual Meeting, Chicago.

Nielsen, J. (1994). *Enhancing the Explanatory Power of Usability Heuristics.* Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence, Boston.

Nielsen, J., & Mack, R. L. (1994). *Usability Inspection Methods*. New York: John Wiley & Sons.

Nielsen, J., & Molich, R. (1990). *Heuristic evaluation of the user interface.* Paper presented at the SIGCHI conference on Human factors in computing systems: Empowering people, Seattle.

Paddison, C., & Englefield, P. (2004). Applying heuristics to accessibility inspections. *Interacting with Computers, 16*(2), 507-521.

Ricketts, C., & Wilks, S. (2002, 2002). *What factors affect students opinions of computer-assisted assessment?* Paper presented at the 6th 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.

Sim, G., Read, J. C., & Holifield, P. (2006). *Using Heuristics to Evaluate a Computer Assisted Assessment Environment.* Paper presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Orlando.

Sim, G., Read, J. C., Holifield, P., & Brown, M. (2007). *Heuristic Evaluations of Computer Assisted Assessment Environments.* Paper presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Vancouver.

Slavkovic, A., & Cross, K. (1999). *Novice Heuristic Evaluations of a Complex Interface.* Paper presented at the CHI 99.

Woolrych, A., & Cockton, G. (2002). *Testing a conjecture based on the DR-AR Model of Usability Inspection Method Effectiveness.* Paper presented at the 16th British HCI Group Annual Conference, London.