Evidence based Design of Heuristics for Computer Assisted Assessment

Gavin Sim, Janet C Read and Gilbert Cockton

School of Computing Engineering and Physical Sciences
University of Central Lancashire, Preston, UK
grsim@grsim@uclan.ac.uk, jcread@uclan.ac.uk

Department of Computing, Engineering and Technology
University of Sunderland Sunderland, UK
gilbert.cockton@sunderland.ac.uk

Abstract. The use of heuristics for the evaluation of interfaces is a well studied area. Currently there appear to be two main research areas in relation to heuristics: the analysis of methods to improve the effectiveness of heuristic evaluations; and the development of new heuristic sets for novel and specialised domains. This paper proposes an evidence based design approach to the development of domain specific heuristics and shows how this method was applied within the context of computer assisted assessment. A corpus of usability problems was created through a series of student surveys, heuristic evaluations, and a review of the literature. This corpus was then used to synthesise a set of domain specific heuristics for evaluating CAA applications. The paper describes the process, and presents a new set of heuristics for evaluating CAA applications.

Keywords: Heuristics, usability, computer assisted assessment

1. Introduction

The process of conducting a heuristic evaluation has been well researched and is widely understood. Currently there are two main research areas in relation to heuristics: analysing methods to improve their effectiveness [1, 2]; and developing domain specific heuristics [3, 4]: this paper concentrates on the later.

The relevant literature contains no consensus over the most effective approach to developing domain specific heuristics. Paddison and Englefield [4] suggests two main methods for developing heuristics; examination of literature; and analysis of data from prior studies. Nielsen [5] used factor analysis and a explanatory coverage process to devise a set of 9 heuristics from a list of 249 problems. In their definition, Paddison and Englefield [4] did not especially clarify the meaning of analysing the data from prior studies, which could be interpreted as conducting primary research or carrying out a meta-analysis of other peoples' results. More clarity is found in [6], which identified three methods for developing heuristics highlighting previous research (Literature), modification of existing (Nielsen's) heuristics and from evaluation results (Primary Research). These more explicit criteria guided the research reported below.

A criticism of the approaches used to create heuristics is the method used for validating the heuristics. The raw count of the number of usability problems identified may not be an appropriate indicator of the effectiveness of a set of heuristics [7]. To validate heuristics, certain criteria are used including thoroughness [8], correctness, coverage and terminology [4]. Correctness refers to the terminology used in the specifications of the heuristics, and whether the descriptions provide sufficient information. Coverage and thoroughness are concerned with the extent to which heuristics adequately represent the domain being evaluated (both terms are used to describe the same construct). Effectiveness relates to the ability of heuristics to capture all significant problems within a domain. Ease of use is concerned with the application of heuristics by evaluators. To establish the effectiveness of a method, the formula proposed by [9] should be applied, however no new domain specific heuristics have had their effectiveness calculated by this method. For example the e-learning heuristic devised by [10] only used literature to synthesise the heuristic set. Reliance on a single method for developing heuristics may result in some important aspects being overlooked or biasing results based on the evaluator's experience.

Within the literature regarding Computer Assisted Assessment (CAA), there is a lack of consensus regarding terminology. Even so, Bull and McKenna [11] argue that CAA is the common term for the use of computers in the assessment of students and the other terminology tends to focus on broader e-learning. Therefore, the definition used here will be: CAA encompasses the use of computers to deliver, mark or analyse assignments or exams. CAA applications range from bespoke applications to off-the-shelf (ready-made) commercial systems. CAA embraces a wide variety of assessment techniques. However, this paper concentrates on devising heuristics for evaluating usability within CAA applications that incorporate *objective tests*, these are questions where the answer is predefined, for example multiple choice questions. With the increased adoption of CAA within educational institutions, there has been a rise in the number of ready-made systems available for delivery of objective tests. These include Questionmark Perception, TRIADS, and TIOA in addition to assessment tools that are incorporated into learning management systems like WebCT and Moodle.

There is limited research surrounding usability of assessment tools relative to studies investigating usability of educational technology environments [12-14]. Usability is important in CAA. Usability problems can cause users difficulties and dissatisfaction with unacceptable consequences. For example, in a multiple choice question with negative marking applied [15], inability to deselect a radio button could result in losing marks. Within CAA, student experience is also affected by pedagogy.

Instructors have preferred pedagogies, but in some instances the technology dictates the pedagogy with respect to test design, as this is governed by the question styles available. Therefore what instructors want may not necessarily be what they get. Their preferred examination may have to be modified to accommodate the constraints of the software. For example, WebCT® in 2005 only offered a limited range of styles compared to dedicated systems such as Questionmark®, therefore, the test experience of the user is dictated by the application. However in bespoke systems, the pedagogical challenge has driven the technology. This is evident for [16], who wished to address issues of guessing within MCQ tests and devised his own system. Regardless of the approach, poorly developed software or pedagogy may have a negative impact for the test take causing problems with unacceptable consequences.

When problems are identified in a heuristic evaluation they are often given a severity rating, using a scale such as the following [17]:

0. I don’t think that this is a usability problem
1. Cosmetic problem only: need not be fixed unless time is available
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

In devising domain specific heuristics, severity rating scales are often overlooked. Domain specific heuristics are synthesized, such as those for the playability of games [18], but complementary severity rating scales are not considered. Evaluators may have difficulty in distinguishing between a problem that is rated as a Major Usability Problem and a Usability Catastrophe when using playability heuristics. When comparing evaluator’s classification of problems to severity ratings research has shown that inter-rater reliability tends to be low [19]. This may be due to the difficulty evaluators have in distinguishing the boundaries between scales. Within the CAA domain, the loss of one mark due to inability to select an answer may be difficult to classify using Nielsen’s severity scale, therefore domain specific severity ratings may need to be devised to accompany new heuristic sets.

This paper proposes an evidence-based approach to synthesis of domain specific heuristics and associated severity ratings, with the aim of providing better coverage than Nielsen’s heuristic set within the CAA domain. This evidence-based approach is also a mixed method research approach, whereby data is gathered from several studies including student surveys, heuristics and analysis of the literature to inform design of new heuristics. While Nielsen’s heuristics may be regarded as dated, and inspection methods as inadequate, the latter remains the best option for CAA, where we cannot possibly submit every authored test to user testing, or even thoroughly user test e-learning tools with CAA features before buying and installing them. Heuristics are thus essential for purchasing decisions, as well as for instructor training. Even where user testing of CAA is possible, it cannot be associated with genuine summative assessments for clear ethical reasons, nor can reliable results be expected from assessments carried out solely for the purpose of user testing, since student motivations and moods will differ between true and artificial testing contexts.

2. Evidence Based Design

Reliance on a single method for developing heuristics may result in some important aspects being overlooked or yielding biased results based on evaluator experience. A mixed method approach to developing heuristics may address shortcomings of creating and validating based on a single method. The approach described here focuses on three areas: determining the effectiveness of Nielsen’s heuristics within the domain; corpus building of usability problems within the domain; and synthesis of domain specific heuristics. Determining what constitutes acceptable evidence is a key challenge for evidence-based methods. A meta-analysis approach would use sources of evidence including guidelines, journal papers or grounded theory based on primary research, but in each instance careful attention must be paid to credibility and validity of data to ensure corpus quality.

2.1 Method

There are several heuristic sets that can be used for heuristic evaluations; two examples are the Squires and Preece e-learning heuristics [10] and Nielsen's heuristics [20]. Nielsen's heuristics have been applied to a wide variety of domains, including hypermedia browsers [21], edutainment applications [22] and to improve the hardware of musical products [23]. This suggested that the effectiveness of Nielsen's heuristics is worth evaluating in the context of CAA. Therefore in this study, the decision was made to use Nielsen's heuristics as they are the most generic and widely applied. Also Ling and Salvendy [6] suggested that it is naïve to develop domain specific heuristics without first considering Nielsen's original heuristic set. Even so, we anticipated that Nielsen's heuristic set and severity rating scale would need to be adapted and extended for CAA. Thus a series of evaluations using a mixed method approach was used to determine the effectiveness of the heuristic set.

2.2 Research Design

In establishing the effectiveness of the heuristics, a corpus building strategy was adopted with the intention of collecting problems with broad coverage of the CAA domain that would lead to unacceptable consequences. This corpus was then used to synthesise a revised and extended CAA-specific heuristic set. By combining student surveys, heuristic evaluations and literature analysis, the following known factors were addressed to ensure corpus quality:

- A range of CAA applications
- Different question styles
- Context of the examination, i.e., summative or formative assessment
- Cohorts – different year groups of students
- Evaluator Effect– different experience and expertise

In each study, different questions styles were addressed. The question styles predominately used were Multiple Choice, Multiple Response and Text Entry. Using an evidence-based approach, the corpus was developed over several studies and any remaining gaps were addressed through a CAA literature review to create the final comprehensive reliable corpus of unacceptable usability problems for CAA.

2.2.1 Surveys

The first stage involved using a survey method to establish if there were severe usability problems within a CAA application that would have unacceptable consequences. Over 300 questionnaires were distributed to students following CAA exams within the first authors' institution. The initial studies were also designed to form a corpus of reported usability problems [24, 25] that could be later used to establish the effectiveness of Nielsen's heuristic set. The survey method dealt with the following corpus quality factors: question styles; CAA applications and cohorts.

2.2.2 Heuristic Evaluations

The next stage involved a heuristic evaluation (Study A) of a CAA application using Nielsen's heuristic set to establish the effectiveness of the heuristic set within the CAA domain and expand the corpus. The evaluation was a between-subjects single factor study with two conditions: formative and summative assessment. The evaluators examined the software based on the potential academic use for formative assessment to support their learning or summative assessment to award a grade.

The evaluators were 11 HCI practitioners of both genders and a diverse age range. Group X consisted of 5 evaluators and Group Y, 6. Both groups completed the same test but evaluated the application within different contexts, Group X considered summative test conditions whilst group Y considered formative ones.

Questionmark® for Windows® was used to deliver the test; this was a standalone application and did not rely upon Internet access. The application was loaded onto the evaluators' laptops which varied in specification. However, the application was designed to be portable and operate under Windows® operating systems, so the evaluators' experiences resembled real usage. The evaluation dealt with the following corpus quality factors: *question styles* and *context*.

A further two additional heuristic evaluations were performed in order to expand the corpus. The next heuristic evaluation (Study B) was designed to expand the corpus and to improve coverage of the following factors; *question styles*, *contexts* and *evaluator* effects. This study used 4 novices and 4 expert evaluators who performed a heuristic evaluation of Questionmark Perception® using Nielsen's heuristic.

After the first two heuristic evaluations, it was clear that Nielsen's heuristics could not fully cover the CAA domain, so the emphasis in this study (Study C) was on extending the corpus as a basis for synthesising a new heuristic set. The study addressed the corpus quality factors of *CAA applications* (by looking at Questionmark, WebCT and TRIADS) and *question styles* [26]. Over 90 evaluators were used, and were split into groups of between 3 and 5 and each group evaluated a single CAA application.

2.2.3 Literature Review

Two literature reviews on CAA [27, 28], were used as the initial focal point along with searches in digital libraries. The purpose of the review was to expand the corpus to incorporate additional problems that had not been identified in the previous studies. This helped to address further factors which may reduce corpus quality, such as the limited range of CAA applications evaluated. The final heuristic set needed to have appropriate domain coverage.

2.3 Coding and Filtering Problems

In order to manage the corpus, each problem was given a unique code to ensure that problems could be cross referenced with other studies. In addition, each problem was coded with a task code based on what user task they would be performing when they encountered this problem and finally a code determining whether it would lead to unacceptable consequences. The unacceptable consequences codes were based on the implications that a problem would have on a student's test performance. They are:

- Dissatisfied – a student could be dissatisfied, but this is unlikely to affect their overall test performance
- Possible – there is a possibility that the problem may affect a student's test performance
- Probable – it would probably affect a student's test performance
- Certain – It would definitely affect students' test performances

The first two authors examined the raw data from each study and allocated task step and unacceptable consequence codes. Many issues identified were based on personal preferences without real adverse consequences for students, for example '*The order of the buttons previous, next, flag is not right – should be next, previous and flag*'. Ideally falsification testing would have been performed to eliminate problems from the corpus but this is not feasible within the CAA domain as it requires user testing. Therefore problems were filtered based on the unacceptable consequences scale above. Any problem judged to be 'Dissatisfied' was removed from the corpus. This would ensure that the corpus only comprised problems that would lead to unacceptable consequences. For example, an evaluator reported the problem of '*no question marks*'. This was judged to be *dissatisfied* as it would not affect the student grades and was removed from the corpus. Once the corpus was filtered, the remaining problems were merged removing duplicate problems by the first author and an educational technologist. The final corpus consisted of 34 problems which would have unacceptable consequences within the CAA domain. This corpus was then further analysed by two lecturers in HCI, two research students and the first author of this paper. A further card sorting exercise was performed. Each participant had to identify a maximum of 12 themes and match each problem independently to their own themes. Following discussion (sometimes vigorous), the themes were then merged with a summary statement. The final stage consisted of the first author and educational technologist examining the themes, re-examining Nielsen's original heuristic set and then devising an appropriate heuristic.

3. Results

A rationale for the new domain specific heuristics depends on the potential ineffectiveness of existing heuristic sets, which thus must be established.

3.1 Student Evaluations

From student post-test surveys, a total of 22 problems were reported. The problems were further filtered, with any problem judged to be 'Dissatisfied' removed. The 13 remaining problems formed the initial corpus that would be systematically extended to assess the effectiveness of Nielsen's heuristics for CAA.

3.2 Heuristic Evaluation Study A

The results from the first heuristic evaluation are presented in Table 1.

Table 1. Number of problems reported in the first heuristic evaluation

	Group X Summative	**Group Y Formative**
Raw Data	50	44
Removed	39	32
Total	11	12
Final Problem Set	17	

Raw Data is the count of problems reported by evaluators before any aggregation of the data. *Removed* is the count of problems removed after card sorting and filtering as described in Section 2.3. The raw data contained 7 problems that could not be matched to one of Nielsen's heuristics, indicating that they were not comprehensive for CAA. Evaluators were unable to attach generic severity ratings to 8 problems in both contexts; which confirms the need for context specific severity ratings.

To determine whether a problem identified in the heuristic evaluation was also reported in the user studies, an additional card sorting exercise was performed. A corpus of usability problems from student surveys in a summative context had previously been created, so this corpus was compared with the problems from the summative evaluation by the HCI experts. The first two authors performed this card sorting exercise. Each recorded problem from the heuristic evaluation was compared with the reported problems from student surveys to establish if the same problems were revealed using both methods. If both researchers agreed that the problems were the same, they were judged to match. In some instances, there was disagreement. To resolve this, there was a discussion about the problem and agreement was reached. Of 11 problems identified using heuristics in the context of summative assessment, card sorting revealed only 3 problems that had been identified in the student surveys.

Using the formula for establishing effectiveness [9] the final data was used to calculate the effectiveness of Nielsen's heuristics relative to the student data studies. Survey data is used in lieu of falsification testing, hence there 3 problems identified in both the user studies and heuristics constitute the problems found, and the 13 problems from the survey are the problems that exist, with 11 problems identified via heuristics. This gives a thoroughness score of 0.23 (3/13), a validity score of 0.27 (3/11) and an effectiveness score of 0.0621 (0.23 x 0.27). This indicates that Nielsen's heuristics are ineffective within the CAA domain. However, thoroughness is a maximum and validity is a minimum, since, had falsification user testing been performed, this would reduce thoroughness but could increase validity. This thus leaves open the issue of false alarms. Jeffries and Desurvire [29] suggest that if development resources are used to correct false alarms, then this may make the application less usable. Hence false alarms could affect the validity of the evaluation method and results [30]. Despite the ineffectiveness of the heuristics set, 17 problems remained after filtering and these were then used to expand the corpus.

3.3 Heuristic Evaluation Study B

The results of this study are reported in [31] and the results demonstrated that, in line with other studies, expert evaluators were better at finding problems than novices. An analysis of the raw data revealed that there were also 7 problems that the evaluators could not classify with an appropriate heuristic in a context of formative assessment, and 6 in a summative context. The results supported the finding from the first heuristic evaluation (Study A) that Nielsen's heuristics do not offer sufficient coverage of the CAA domain. The number of problems that were added to the evolving corpus is shown in Table 2.

Table 2. Problems from the second heuristic evaluation

	Summative	Formative
Raw Data	50	44
Removed	13	21
Total	28	34
Final Problem Set	24	

3.4 Heuristic Evaluation Study C

There was a total of 24 groups. Each group identified problems unique to their specific CAA environments, so it is questionable whether five evaluators, as suggested by [32] would find the majority of known problems. Some major usability problems such as browser buttons terminating the exam within Questionmark® may not have been reported with a much smaller number of evaluators.

Table 3. Data from heuristic evaluation of 3 CAA applications

	Questionmark	TRIADS	WebCT
Raw Data	41	51	44
Removed	20	26	23
Total	21	25	21
Final Problem Set	38		

Many problems were unique to a particular CAA application. Only five problems were reported for all three, making it clear that to further expand the corpus, other

CAA applications would need to be evaluated. It was not possible to access all CAA applications, thus a literature review would address remaining gaps in coverage.

3.5 Literature Review

The literature revealed 24 problems identified in a range of CAA applications including TOIA and V32. They also identified issues with question styles not identified in the student surveys or heuristic evaluations.

Table 4. Themes identified and merged following the card sorting exercise

Final Theme	Individual themes
TH1. Moving through the test	Navigation x3 Clear Navigation Exiting the test
TH2. Interface / Visual Design	Bad Interface Layout Readability
TH3. Reduce Errors	Reduce errors – auto save Errors
TH4. Intuitive Input	Input Issues Answering questions Input
TH5. User Freedom	Match real world e.g. chance to review and edit objective test answers
TH6. Protecting Answers	Saving Issues
TH7. Access	Access Accessing Test
TH8. Test Design	Unclear Information in test Teacher Issues x2 Test related
TH9. Psychological / Perception	Comparability with paper Trust Stupidity Perception
TH10. Physical	Online Issues Hardware x2
TH11. System Feedback	Provide Help x2 Feedback x3 Confirm all actions Inadequate information for users Feedback and support

4. Synthesis of Heuristics

Table 4 presents themes that emerged as a result of the card sorting exercise. The final themes were then used to synthesise a set of heuristics for CAA. The first author and an educational technologist then re-examined Nielsen's heuristic set and the themes that had emerged to compare and contrast. The purpose was to guide synthesis of CAA heuristics through close attention to terminology before translating the themes into a heuristic set. For example, having a heuristic called 'access' (TH4, Table 4) would be ambiguous and not aid evaluators when using this heuristic.

Table 5. Final Heuristic Set and Descriptions, Retained and Modified Nielsen's

Theme	Heuristic	Description
Same as Nielsen's Heuristics		
TH3. Reduce Errors	H3. Error prevention and recovery	Prevent errors from affecting test performance and enable the student to recover from mistakes.
TH5. User Freedom	H5. User control and freedom	The test should match real world experience e.g. chance to review and edit
TH11. System Feedback	H11. Ensure appropriate help and feedback	System feedback should be clear about what action is required. For complex actions help should be provided.
Modification of Nielsen's Heuristics		
TH2. Interface/ Visual Design	H2. Ensure appropriate interface design characteristics	Interface should match standards and design should support user tasks.
TH4. Intuitive Input	H4. Answering question 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.

Of the 11 heuristics, 3 were based on Nielsen's original set, 2 were modifications, and 6 were new heuristics specific to CAA. The process of creating heuristics from themes was rather complex. Appropriate terminology was important to encapsulate problems in the way that breaches of a heuristic could be clearly established. For example, *Psychological and Perception* (TH9) did not allow the researcher or educational technologist to establish whether a violation of this had occurred when conducting a heuristic evaluation. This would be influenced by the evaluators' prior experience of CAA or exams and understanding of the technology. However in

Nielsen’s heuristic set, Aesthetics and Minimalist Design would give rise to similar issues, so therefore the heuristic *Design should inspire trust and not unfairly penalise* was named to capture the psychological theme.

Table 6. Final Heuristic Set and Descriptions, New Heuristics not in Nielsen’s Set

New Heuristics		
TH1. Moving through the test	H1. Navigating within the application and terminating the exam should be intuitive	Navigation should be intuitive enabling the user to identify where they have been, where they are and where they want to go. Options to exit should be identifiable.
TH6. Protecting Answers	H6. Prevent loss of input data	When answers are input the data should not be lost or corrupted.
TH7. Access	H7. Accessing the test should be clear and intuitive	Students should not encounter any difficulty in accessing the test.
TH8. Test Design	H8. Use clear language and grammar within questions and ensure the score is clearly displayed.	Text should be grammatically correct and make sense. It should be obvious to students what the score is for a particular question and the scoring algorithm applied (e.g. if negative marking is used). Question feedback should assist the learning process.
TH9. Psychological/ Perception	H9. Design should inspire trust and not unfairly penalize	Students should feel confident that the system will not fail. Ensure test mode does not impact on fairness and performance within the test. For example it should be clear if marks would be lost for incorrect spelling.
TH10. Physical	H10. Minimise external factors which could affect the user	Ensure that there is minimal latency when moving between questions or saving answers. Also ensure delivery platform is secure and robust.

With the initial CAA heuristic set synthesized, the first author and educational technologist then went through the process of cross-checking every problem in the final corpus against the new CAA heuristics. Each heuristic was numbered from 1 to 11 and every problem was successfully mapped to a heuristic. A decision was made to let a problem be mapped to more than one heuristic. During this process, one heuristic was extended to enable incorporation of the problem *Recovery from errors*. Heuristic H3 *error prevention* was extended to become *error prevention and recovery*.

To ensure that the heuristics offered better coverage than Nielsen's heuristic set it was important that each problem could be classified to at least one heuristic, which was achieved.

5. Conclusions

This paper has presented an evidence based approach to the development of a set of heuristics for CAA. Through a process of student surveys and heuristic evaluation a corpus of usability problems was created. However, the process of creating a corpus was time consuming and required a significant amount of time for filtering and merging the data sets. In total, over 300 problems were reported in the various studies and this corpus was filtered and merged to leave 34 problems that would lead to unacceptable consequences for students in CAA.

The heuristic set that was synthesized offers enhanced coverage of the CAA domain. Further studies are required with the heuristic set to establish ease of use, with a focus on the adequacy of the terminology. In addition unacceptable consequences have potential for use within severity rating scales. The reliability of this will be evaluated and compared to establish if inter-rater reliability is greater than when using Nielsen's scale. However, the validity of the new heuristics cannot be validated against user testing (as, e.g., in [1]), as user testing is not well suited to the CAA domain, as argued early in this paper. Claims for the adequacy of the new CAA heuristics are thus based on their systematic inspectable derivation from relevant examples based on over 300 reported usability problems for real world CAA applications (in contrast, the 249 problems used in [5] were far more heterogeneous). The whole process of derivation is inspectable, focused well grounded and diverse, having involved a good range of HCI and e-learning expertise. Given this, we are confident that our new set of CAA heuristics can reliably support CAA authors in the elimination of potential unacceptable usability problems through well informed procurement of CAA applications and revisions to specific objective test designs.

6. References

1. Cockton, G., A. Woolrych, and M. Hindmarch. *Reconditioned Merchandise: Extending Structured Report Formats in Usability Inspection*. in *CHI 2004*. 2004. Vienna.
2. Coyle, C.L., et al. *Heuristic Evaluations at Bell Labs: Analyses of Evaluator Overlap and Group Sessions*. in *CHI*. 2007. San Jose: ACM.

3. Korhonen, H. and E.M. Koivisto. *Playability Heuristics for Mobile Games*. in *MobileHCI*. 2006. Helsinki: ACM.
4. Paddison, C. and P. Englefield, *Applying heuristics to accessibility inspections.* Interacting with Computers, 2004. **16**(2): p. 507-521.
5. Nielsen, J. *Enhancing the Explanatory Power of Usability Heuristics*. in *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*. 1994. Boston: ACM.
6. Ling, C. and G. Salvendy, *Extension of Heuristic evaluation method: a review and reaapraisal.* International Journal of Ergonomics and Human Factors, 2005. **27**(3): p. 179-197.
7. Gray, W.D. and M.C. Salzman, *Damaged Merchandise?* Journal of Human-Computer Interaction, 1998. **13**(4): p. 203-262.
8. Somervell, J. and D. McCrickard, S., *Better discount evaluation: illustrating how critical parameters support heuristic creation.* Interacting with Computers, 2005. **17**(5): p. 592-612.
9. Hartson, H.R., T.S. Andre, and R.C. Williges, *Criteria for Evaluating Usability Evaluation Methods.* International Journal Human Computer Interaction, 2003. **15**(1): p. 145-181.
10. Squires, D. and J. Preece, *Predicting quality in educational software: Evaluating for learning, usability and the synergy between them.* Interacting with Computers, 1999. **11**: p. 467-483.
11. Bull, J. and C. McKenna, *Blueprint for computer-assisted assessment*. 2001: Loughborough University. 168.
12. Parlangeli, O., E. Marchigiani, and S. Bagnara, *Multimedia systems in distance education: effects of usability on learning.* Interacting with Computers, 1999. **12**(1): p. 37-49.
13. Piguet, A. and D. Peraya, *Creating web-integrated learning environments: An analysis of WebCT authoring tools in respect to usability.* Australian Journal of Educational Technology, 2000. **16**(3): p. 302-314.
14. Berg, G.A., *Human-Computer Interaction (HCI) in Educational Environments: Implications of Understanding Computers as Media.* Journal of Educational Multimedia and Hypermedia, 2000. **9**(4): p. 347-368.
15. Sim, G., M. Horton, and S. Strong. *Interfaces for online assessment: friend or foe?* in *7th HCI Educators Workshop*. 2004. Preston: LTSN.
16. Davies, P. *There's no confidence in multiple-choice testing.* in *Sixth International Computer Assisted Assessment Conference*. 2002. Loughborough.
17. Nielsen, J. and R.L. Mack, *Usability Inspection Methods*. 1994, New York: John Wiley & Sons.
18. Desurvire, H., M. Caplan, and J.A. Toth. *Using heuristics to Evaluate the Playability of Games*. in *CHI*. 2004. Vienne: ACM.
19. Jacobsen, N.E. and B.E. John. *The evaluator effect in usability studies: problem detection and severity judgements*. in *Proceeding of the Human Factors and Ergonomics Society 42nd Annual Meeting*. 1998. Chicago: HFES.
20. Nielsen, J., *Usability Engineering*. 1994: Morgan Kaufmann.

21. Connell, I.W. and N.V. Hammond. *Comparing usability evaluation principles with heuristics: problem instances versus problem types.* in *Human-Computer Interaction - INTERACT ‘99*. 1999. Edinborough: Amsterdam: IOS Press.
22. Embi, Z.C. and H. Hussain, *Analysis of local and foreign edutainment products- and effort to implement the design framework for an edutainment environment in Malaysia.* Journal of Computers in Mathematics and Science Teaching, 2005. **24**(1): p. 27-42.
23. Fernandes, G. and C. Holmes. *Applying HCI to music related hardware.* in *CHI 2002*. 2002. Minneapolis, Minnesota.: ACM.
24. Sim, G. and P. Holifield. *Computer Assisted Assessment: All those in favour tick here.* in *World Conference on Educational Multimedia, Hypermedia and Telecommunications*. 2004. Lugano: AACE.
25. Sim, G. and P. Holifield. *Piloting CAA:All aboard*. in *8th International Computer Assisted Assessment Conference*. 2004. Loughborough.
26. Sim, G., et al. *Heuristic Evaluations of Computer Assisted Assessment Environments*. in *World Conference on Educational Multimedia, Hypermedia and Telecommunications*. 2007. Vancouver: AACE.
27. Conole, G. and B. Warburton, *A review of computer-assisted assessment.* ALT-J, 2005. **13**(1): p. 19-33.
28. Sim, G., P. Holifield, and M. Brown, *Implementation of computer assisted assessment: lessons from the literature.* ALT-J, 2004. **12**(3): p. 215-229.
29. Jeffries, R. and H. Desurvire, *Usability Testing vs Heuristic Evaluation: Was there a contest?* SIGCHI Bulletin, 1992. **24**(4): p. 39-41.
30. Lavery, D. and G. Cockton. *Representing Predicted and Actual Usability Problems*. in *Workshop on Representation in Interactive Software Development*. 1997. Queen Mary and Westfield College, University of London.
31. Sim, G., J.C. Read, and P. Holifield. *Evaluating the user experience in CAA Environments: What affects user satisfaction?* in *10th International Computer Assisted Assessment Conference*. 2006. Loughborough.
32. Nielsen, J. *Finding usability problems through heuristic evaluation.* in *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1992. Monterey: ACM.