

Evidence Based Design of Heuristics: Usability and Computer Assisted Assessment

by
Gavin Sim

A thesis submitted in partial fulfillment for the requirements of the
degree of PhD at the University of Central Lancashire,
Preston.

2009

Student Declaration

- I declare that while registered as a candidate for the research degree, I have not been a registered candidate or enrolled student for another award of the University or other academic or professional institution.
- I declare that no material contained in the thesis has been used in any other submission for an academic award and is solely my own work.

Signature of Candidate: _____



Type of Award: PhD

School: School of Computing, Engineering and Physical Sciences

Evidence Based Design of Heuristics: Usability and Computer Assisted Assessment

Gavin Sim

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

Abstract

The research reported here examines the usability of Computer Assisted Assessment (CAA) and the development of domain specific heuristics. CAA is being adopted within educational institutions and the pedagogical implications are widely investigated, but little research has been conducted into the usability of CAA applications.

The thesis is: *severe usability problems exist in CAA applications causing unacceptable consequences, and that using an evidence based design approach CAA heuristics can be devised.* The thesis reports a series of evaluations that show severe usability problems do occur in three CAA applications. The process of creating domain specific heuristics is analysed, critiqued and a novel evidence based design approach for the design of domain specific heuristics is proposed. Gathering evidence from evaluations and the literature, a set of heuristics for CAA are presented. There are four main contributions to knowledge in the thesis: the heuristics; the corpus of usability problems; the Damage Index for prioritising usability problems from multiple evaluations and the evidence based design approach to synthesise heuristics.

The focus of the research evolves with the first objective being to determine *If severe usability problems exist that can cause users difficulties and dissatisfaction with unacceptable consequences whilst using existing commercial CAA software applications?* Using a survey methodology, students' report a level of satisfaction but due to low inter-group consistency surveys are judged to be ineffective at eliciting usability problems. Alternative methods are analysed and the heuristic evaluation method is judged to be suitable. A study is designed to evaluate Nielsen's heuristic set within the CAA domain and they are deemed to be ineffective based on the formula proposed by Hartson *et al.* (2003). Domain specific heuristics are therefore necessary

Abstract

and further studies are designed to build a corpus of usability problems to facilitate the evidence based design approach to synthesise a set of heuristics. In order to aggregate the corpus and prioritise the severity of the problems a Damage Index formula is devised.

The work concludes with a discussion of the heuristic design methodology and potential for future work; this includes the application of the CAA heuristics and applying the heuristic design methodology to other specific domains.

Contents

Chapter 1 Introduction.....	1
1.1 Introduction.....	1
1.1.1 Structure.....	3
1.2 Overcoming Problems With CAA.....	3
1.3 The Thesis.....	5
1.3.1 Structure of Thesis.....	5
1.3.2 Literature Review.....	7
1.3.3 Methodology and Evidence Based Design.....	7
1.3.4 Survey Studies.....	7
1.3.5 Heuristic Evaluations.....	8
1.3.6 Evidence Based Design of Heuristics for CAA.....	9
1.3.7 Conclusions and Further Research.....	9
1.4 Conclusions.....	9
1.4.1 Publications Related to the Thesis.....	9
Chapter 2 CAA Overview.....	11
2.1 Introduction.....	11
2.1.1 Objectives.....	11
2.1.2 Scope.....	11
2.2 Assessment.....	12
2.3 Types of Assessment.....	13
2.4 Assessment Techniques.....	14
2.5 Variations in CAA.....	14
2.6 Adoption of CAA.....	15
2.7 CAA Software.....	16
2.8 Stakeholders within CAA.....	18
2.8.1 Student’s Goal.....	18
2.8.2 User Tasks.....	21
2.9 What can CAA Test?.....	21
2.10 Question Styles.....	22
2.10.1 Point and Click.....	23
2.10.2 Text Entry.....	23
2.10.3 Move Object.....	24
2.10.4 Draw Object.....	24
2.11 Guessing.....	25
2.12 Accessibility.....	26
2.13 Institutional Strategies for the Adoption of CAA.....	27
2.14 Security Issues with CAA.....	28
2.15 Test Design and CAA.....	30
2.16 Potential Unacceptable Consequences.....	32
2.17 Conclusions.....	34
2.17.1 Limitations.....	34
2.17.2 Contributions.....	34
Chapter 3 Evaluating Usability.....	35
3.1 Introduction.....	35
3.1.1 Objectives.....	35
3.1.2 Scope.....	35

3.2 Usability Evaluation Methods.....	36
3.2.1 Automated.....	37
3.2.2 Empirical – User Studies	37
3.2.3 Formal Inspections.....	38
3.2.4 Inspection Methods.....	39
3.2.4.1 Cognitive Walkthroughs	39
3.2.4.2 Heuristic Evaluations.....	39
3.3 Usability and Computer Assisted Assessment.....	40
3.3.1 Potential Task Based Usability Problems.....	42
3.3.2 Evaluating Usability in CAA.....	43
3.4 Conclusions.....	44
3.4.1 Limitations	45
3.4.2 Contributions.....	46
Chapter 4 Methodology and Research Design	47
4.1 Introduction.....	47
4.1.1 Objectives	47
4.1.2 Structure.....	47
4.1.3 Contributions.....	47
4.2 Selecting Research Methods	48
4.3 Research Methods.....	49
4.3.1 Types of Data.....	49
4.3.2 Techniques of Data Elicitation.....	49
4.3.2.1 Direct.....	50
4.3.2.2 Indirect	50
4.3.3 Types of Design for Monitoring Change	51
4.3.3.1 Longitudinal.....	51
4.3.3.2 Cross-Sectional	51
4.3.3.3 Sequential.....	52
4.3.4 Treatment of the Data	52
4.3.4.1 Qualitative Data	52
4.3.4.2 Quantitative Data	52
4.3.5 Reliability.....	53
4.3.6 Validity	53
4.3.7 Mixed Methods.....	55
4.4 Research Design	55
4.4.1 Purpose.....	56
4.4.2 Survey Design.....	56
4.4.2.1 Analysis of Survey Data	57
4.4.2.2 Limitations of Survey Methods	60
4.4.3 Merging of Data.....	60
4.5 Participant Selection.....	61
4.6 Ethics.....	62
4.7 Conclusions.....	64
Chapter 5 Usability Pilot Test.....	65
5.1 Introduction.....	65
5.1.1 Objectives	65
5.1.2 Scope.....	66
5.1.3 Contributions.....	66
5.1.4 Structure.....	67
5.2 Study Design	67

Contents

5.2.1 Participants.....	67
5.2.2 Apparatus	67
5.2.3 Questionnaire Design.....	67
5.2.4 Exam Design.....	69
5.2.5 Procedure	70
5.2.6 Analysis.....	71
5.3 Quantitative Results.....	72
5.4 Qualitative Results.....	74
5.4.1 Logging on.....	74
5.4.2 During the Test	75
5.4.3 Ending the Test.....	76
5.5 Conclusions.....	77
5.5.1 Usability Problems Identified in WebCT®	79
5.5.2 Methodological Limitations.....	79
5.5.3 Research Questions.....	80
Chapter 6 2nd Pilot Usability Evaluation	81
6.1 Introduction.....	81
6.1.1 Objectives	81
6.1.2 Scope.....	82
6.1.3 Contributions.....	82
6.1.4 Structure.....	83
6.2 Study Design.....	83
6.2.1 Participants.....	83
6.2.2 Apparatus	84
6.2.3 CAA Question Design	84
6.2.4 Questionnaire Design.....	85
6.2.5 Procedure	86
6.2.6 Analysis.....	86
6.3 Qualitative Results.....	87
6.3.1 Group A Results.....	87
6.3.2 Group B Results.....	87
6.3.3 Group C Results.....	88
6.3.4 Group D Results.....	88
6.3.5 Group E Results.....	89
6.3.6 Merged Problem Sets.....	89
6.3.7 Problems Reported across WebCT® and Questionmark®.....	91
6.4 Discussion.....	91
6.5 Conclusions.....	93
6.5.1 Methodological Limitations.....	94
6.5.2 Research Questions.....	94
Chapter 7 Heuristic Evaluations	96
7.1 Introduction.....	96
7.1.1 Objectives	96
7.1.2 Scope.....	96
7.2 Heuristic Evaluations.....	97
7.3 New Heuristics.....	99
7.4 Developing Heuristics.....	100
7.4.1 Developing Heuristics Based on Literature	100
7.4.2 Modification of Nielsen’s Heuristics	101
7.4.3 Primary Research.....	101

7.5 <i>Validating Heuristics</i>	101
7.5.1 No Validation of Heuristics	102
7.5.2 Validating Heuristics by using them	103
7.5.3 Comparison of new Heuristics with Nielsen’s Heuristics	103
7.5.4 Comparison Between Heuristics and User Studies.....	104
7.6 <i>Heuristic Evaluations Research Design</i>	107
7.6.1 Analysis of Heuristics Data	107
7.6.2 Limitations of Heuristic Evaluations	109
7.7 <i>An Evidence Based Design Approach to Corpus Building</i>	109
7.7.1 What is Acceptable Evidence?.....	110
7.7.2 The Research Strategy - Summary	110
7.7.2.1 Stage 1 – Pilot Studies	111
7.7.2.2 Stage 2 – Existing Heuristics	112
7.7.2.3 Stage 3 – Heuristic Evaluations	112
7.7.2.4 Stage 4 – Audit from the Literature	112
7.7.2.5 Stage 5 – Synthesis of Heuristics.....	113
7.7.3 Merging data	113
7.8 <i>Conclusions</i>	113
Chapter 8 Pilot Study of Nielsen’s Heuristics	115
8.1 <i>Introduction</i>	115
8.1.1 Objectives	115
8.1.2 Scope.....	116
8.1.3 Contributions.....	116
8.1.4 Structure	116
8.2 <i>Experimental Design</i>	117
8.2.1 Participants.....	117
8.2.2 Apparatus	118
8.2.3 CAA Questions Design.....	118
8.2.4 Procedure	118
8.2.5 Analysis.....	120
8.3 <i>Results</i>	121
8.3.1 Number of Problems Found.....	121
8.3.2 Evaluator Effects.....	122
8.3.3 Attaching Problems found to Heuristics	123
8.3.4 Severity Ratings.....	124
8.3.5 Further Aggregation of Data Sets	126
8.3.6 Problems Identified in both User Studies and Heuristics	127
8.3.7 Analysis of the Effectiveness of Nielsen’s Heuristics	128
8.3.8 Problems with Unacceptable Consequences.....	129
8.4 <i>Conclusions</i>	130
8.4.1 Methodological Limitations.....	132
8.4.2 Research Questions.....	132
Chapter 9 Additional Pilot Study using Heuristics to Expand the Corpus.....	134
9.1 <i>Introduction</i>	134
9.1.1 Objectives	134
9.1.2 Scope.....	135
9.1.3 Contributions.....	135
9.1.4 Structure	136
9.2 <i>Study Design</i>	136
9.2.1 Participants.....	137

9.2.2 Apparatus	138
9.2.3 CAA Question Design	138
9.2.4 Procedure	139
9.2.5 Analysis.....	140
9.3 Results	140
9.3.1 Number of Usability Problems Discovered	140
9.3.2 Evaluator Effect	141
9.3.3 Problem Classification for Formative Assessment	143
9.3.4 Problem Classification for Summative Assessment	144
9.3.5 Problems Identified in Both Contexts.....	145
9.3.6 Inclusion of Information about Assessment	146
9.3.7 Severity Ratings.....	147
9.3.8 Severity Reliability over Time.....	150
9.3.9 Difference in Severity Ratings Based on Context of Use.....	151
9.3.10 Further Aggregation of Data Sets	151
9.3.10.1 Problems in Both Contexts	152
9.4 Conclusions.....	152
9.4.1 Methodological Limitations.....	154
9.4.2 Further Research	154
Chapter 10 Expanding the Corpus and Developing an Aggregation Instrument	
.....	156
10.1 Introduction.....	156
10.1.1 Objectives	156
10.1.2 Scope.....	157
10.1.3 Contributions.....	157
10.1.4 Structure	157
10.2 Study Design	157
10.2.1 Evaluators	158
10.2.2 Design	159
10.2.3 CAA Question Design	159
10.2.4 Procedure	160
10.2.5 A Method for Aggregating the Data	163
10.2.6 Recoding the Data.....	165
10.3 Results.....	166
10.3.1 Stage 1 - Student Aggregation	166
10.3.2 Stage 2 - Duplication Treatment.....	167
10.3.3 Stage 3 - Researcher Reduction	167
10.3.4 Stage 4 – Applying the Damage Index	168
10.3.5 Stage 5 - Type Allocation	170
10.3.6 Unique Problems.....	170
10.4 Exploring the use of the Damage Index and Unacceptable Consequences Scale	
.....	171
10.4.1 Problems Classified with Tasks and Consequences	173
10.4.1.1 Problems with Unacceptable Consequences.....	173
10.4.1.2 Unacceptable Consequences vs. Damage Index.....	174
10.4.1.3 Problems Found in all Three CAA Environments.....	175
10.5 Conclusions.....	176
10.5.1 Methodological Limitations.....	178
10.5.2 Research Questions.....	179
Chapter 11 Synthesis of Heuristics for CAA.....	180

Contents

<i>11.1 Introduction</i>	180
11.1.1 Objectives	180
11.1.2 Scope.....	181
11.1.3 Contributions.....	181
11.1.4 Structure.....	181
<i>11.2 Expanding the Corpus</i>	181
<i>11.3 User Studies</i>	182
<i>11.4 Existing Heuristics</i>	183
<i>11.5 Literature Review</i>	184
11.5.1 Data Set from Literature Review	185
<i>11.6 Merger of the Data Sets</i>	185
11.6.1 Merging – Phase One.....	187
11.6.2 Merging – Phase Two	192
11.6.2.1 Card Sorting.....	192
<i>11.7 Synthesis of Heuristics for CAA</i>	194
11.7.1 Mapping Problems to Heuristic Set.....	196
<i>11.8 Conclusions</i>	197
11.8.1 Methodological Limitations.....	198
Chapter 12 Conclusions	200
<i>12.1 Introduction</i>	200
<i>12.2 Research Approach</i>	200
<i>12.3 Contributions to Knowledge</i>	202
12.3.1 The Corpus of Usability Problems.....	203
12.3.2 Damage Index	203
12.3.3 Evidence Based Design.....	204
12.3.4 Heuristics for CAA	204
<i>12.4 Discussion</i>	206
12.4.1 Corpus Revisited.....	206
12.4.2 Heuristic Evaluations Revisited.....	207
12.4.3 Damage Index	208
12.4.4 Evidence Based Design Approach Revisited.....	208
<i>12.5 Future Work</i>	209
12.5.1 Heuristics	209
12.5.2 Evidence Based Design.....	210
12.5.3 Comparing Evaluation Methods	211
<i>12.6 Concluding Remarks</i>	211
Appendices	213
References	214

List of Figures and Equations

Figure 1 Motivation for the research in the thesis	1
Figure 2 The initial research strategy adopted	2
Figure 3 The structure of the thesis.....	6
Figure 4 Assessment broken down into components (after Conole and Fill).....	13
Figure 5 CAA Software and Variations.....	15
Figure 6 A Questionmark template with navigation on right (2005).....	17
Figure 7 Questionmark Template with vertical navigation (2005).....	17
Figure 8 Activity Diagram of user interaction with a CAA environment	20
Figure 9 Objective testing grouping by interaction type	23
Figure 10 Relationship between technology and pedagogy	30
Figure 11 Relationship between user experience, technology and assessment	31
Figure 12 Classes of evaluation methods (Whitefield et al, 1991).....	36
Figure 13 Overview of research structure.....	48
Figure 14 Triangulation of data	61
Figure 15 The assessment interface used within WebCT®.....	70
Figure 16 Left Group A interface and Right Group B interface.....	86
Figure 17 Nielsen and Landauer Formula	98
Figure 18 Hartson et al. formulas	102
Figure 19 Sources of evidence that feed the development of new heuristics	111
Figure 20 Completed Heuristic Evaluation Report Sheet.....	119
Figure 21: The assessment interface used within Questionmark® for Windows®...	120
Figure 22 Overlap between two studies.....	128
Figure 23 Overlap between studies based on unacceptable consequences.....	130
Figure 24 From left to right the interfaces used for formative and summative assessment.....	138
Figure 25 Problems found in both contexts	145
Figure 26 Number of problems in both contexts	152
Figure 27 From left to right Questionmark , TRIADS and WebCT.....	158
Figure 28 Reporting form used in the heuristic evaluation.....	161
Figure 29 Example of a groups aggregated problem set.....	163
Figure 30 Damage Index Formula	164
Figure 31 Problems found within each application	175
Figure 32 Three stages of investigation used to develop the corpus	182
Figure 33 Triangulation of data from the various studies.....	186
Figure 34 Merging of data within and between task step codes	187
Figure 35 Revised Damage Index Formula	189
Figure 36 Damage Index Formula.....	203

List of Tables

Table 1 Reliability of Questionnaire	69
Table 2 Results to the question ‘Did you have any difficulty accessing the test?’	73
Table 3 Mean scores for usability questions.....	73
Table 4 Mean scores for student satisfaction with WebCT	74
Table 5 Usability problems found in WebCT®.....	79
Table 6 Overview of test design for each group.....	84
Table 7 Reported Usability Problems for Questionmark.....	90
Table 8 Development and validation of domain specific heuristics	106
Table 9 Mean value of problems identified based on context	121
Table 10 Number of problems found within each of the heuristic evaluations	122
Table 11 Total number of problems found by each evaluator and their lambda value calculated on the total aggregated problems.....	123
Table 12 Number of problems classified to each heuristic.....	124
Table 13 Number of problems attached to each severity rating	124
Table 14 Mean severity ratings after the card sorting exercise	125
Table 15 Problems identified in both user and heuristic evaluations	127
Table 16 Problems with consequences attached.....	129
Table 17 The order each of the groups applied the heuristics.	136
Table 18 Shows the order the evaluators saw the question sets	139
Table 19 Number of usability problems reported after each analysis stage	141
Table 20 The average number of usability problems found based on evaluator experience	141
Table 21 Total number of problems found by each evaluator with their lambda value calculated on the total aggregated problems.....	142
Table 22 Number of formative problems classified to heuristics.....	144
Table 23 Number of summative problems classified to heuristics.....	145
Table 24 Number of problems found by each group based on context	146
Table 25 Problems classified to each of the severity ratings	147
Table 26 Formative problems where an evaluator rated the problem 0 and another rated it 3	148
Table 27 Summative problems where an evaluator rated the problem 0 and another rated it 3 or more.....	149
Table 28 Problems with consistent severity ratings over time	150
Table 29 Number of problems reported in each of the 3 applications.....	166
Table 30 Problems classified to each of the heuristics at stage 0	167
Table 31 The mean number of problems identified per application.....	167
Table 32 Number of unique problems reported by group and application	171
Table 33 The number of problems classified to each of the severity ratings	172
Table 34 Results of the damage index applied to the three environments.....	172
Table 35 Problems recoded with task and consequence	173
Table 36 Problems rating to the unacceptable consequences scale	173
Table 37 Damage index compared to consequences classification	174
Table 38 Cross tabulation of the Consequences Scale and Damage Index for WebCT	175
Table 39 Problems that appear in all 3 applications	176
Table 40 Problems remaining after filtering process.....	183
Table 41 Number of problems mapped to the task step code and merged within	188
Table 42 Problems remaining after between task step code analysis	189
Table 43 Example of the merging process for the task step code T	191

List of Tables

Table 44 Final themes merged from groups individual themes.....	193
Table 45 Initial Heuristic Set based on Retaining and Modifying Nielsen's	194
Table 46 New Heuristics synthesised not in Nielsen's set	195
Table 47 Final Heuristic Set and problems classified to each heuristic	196

Acknowledgements

There have been many people that have contributed to the work in this thesis and this section acknowledges their contribution. I have been lucky to get great support from the Department of Computing and now the School of Computing Engineering and Physical Sciences at the University of Central Lancashire throughout the course of my PhD.

My supervisory team of Dr Phil Holifield, Dr Janet C Read and Dr Martin Brown merits special mention. They have reviewed my work, encouraged me to publish and offered support and advice when necessary. They have always been there when needed and found time with little notice to offer assistance and guidance. Special thanks should go to Dr Janet C Read for her meticulous review of the draft versions of the thesis and suggestions during the writing up stage.

From the HCI community, special thanks must also go to Prof. Gilbert Cockton for his thorough review of early work, his assistance in the methodology and reviewing of draft versions of the thesis, for this I am forever indebted.

I would also like to thank the staff for enabling me to use their modules for data collection in the early stages of the research. In addition I would like to thank the staff who gave up their valuable time to perform the heuristic evaluations and merging of the data sets, without whom I would have been unable to complete this research.

Over 300 students have been involved in this research, by completing surveys and performing heuristic evaluations of the CAA applications and I am grateful for their co-operation and willingness to participate in the studies.

Finally, I have to mention my family who have had to endure laptops on holidays and weekends disrupted. Thanks Stef and Jack for putting up with me over the years.

Dedication

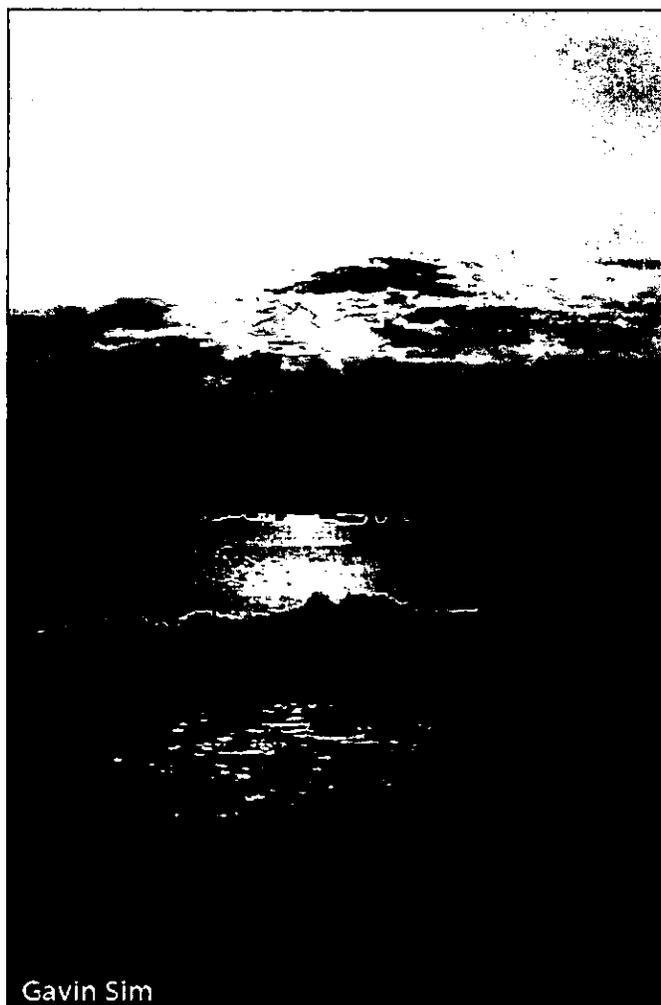
The work in this thesis is dedicated to my wife.

Without her love and support I would have found the journey lonely and difficult.

She has always been there for me,
offering encouragement and listening to my concerns.

She has had to endure some very dull conversations over several meals
and a messy house.

Thanks for sharing the journey with me.



Glossary of Terms and Abbreviations

Association for the advancement of computers in education	AACE	An international organisation that researches the use of technology in education.
Association for Computing Machinery	ACM	The worlds largest educational and scientific computing society; delivers resources that advance computing as a science and a profession.
Computer Assisted Assessment	CAA	CAA encompasses the use of computers to deliver, mark or analyse assignments or exams
Human Computer Interaction	HCI	
Joint Information Systems Committee	JISC	Organisation funded by the UK FE and HE funding bodies to provide world class leadership in the innovative use of ICT to support education and research.
Multiple Choice Question	MCQ	A question with a predefined answer which requires the person to select the correct answer from a list that includes a number of distracters.
Questionmark®		An assessment management system that enables educators and trainers to author, schedule, deliver and report on surveys, quizzes, tests and exams.
Tripartite interactive assessment delivery system	TRIADS®	TRIADS is a highly flexible interactive assessment system capable of delivering a wide variety of question styles in a wide variety of modes to facilitate the testing of higher order learning skills.
Usability Evaluation Method	UEMs	A methodology for finding or testing for usability problems in an application or prototype.
Unique problem	UPT	A usability problem that is not identified by another evaluator.

Glossary

Tokens		
WebCT®		An online proprietary virtual learning environment.

Chapter 1 Introduction

1.1 Introduction

This chapter is used to introduce the focus of the thesis, examining some of the issues associated with Computer Assisted Assessment (CAA) and the move towards technology enabled assessment practices.

The work was motivated by an interest in the area of educational technology and usability. This interest arose as a result of working on a number of projects and initially the research presented in this thesis was conducted in conjunction with these projects, see Figure 1.

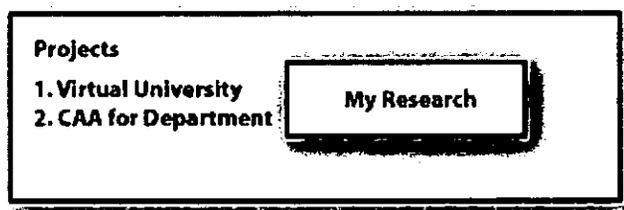


Figure 1 Motivation for the research in the thesis

When the work on this thesis began, there was growing international research focusing on computer assisted assessment (Bull & McKenna, 2001) but this work concentrated on implementation issues and pedagogical challenges such as question design, hardware requirements and institutional strategies for uptake. In 2001 there had been very little work conducted into usability and computer assisted assessment. Early studies had examined student satisfaction which showed a level of satisfaction (O'Hare, 2001), but did not examine specific interface attributes or identify potential usability problems. This led to the work reported in this thesis, the initial objective was to establish *“If severe usability problems exist that can cause users difficulties and dissatisfaction with unacceptable consequences whilst using existing commercial CAA software applications?”*. From the objective the following hypotheses were formulated:

- Usability problems exist which could have an impact on students' test results thus leading to unacceptable consequences.
- Students are satisfied with commercial CAA applications.

Usability is an important issue within CAA as any usability problems with the software may constitute a threat to the fairness of the assessment. For example, if the test was designed to assess students' ability to understand chemistry, but as a result of poor usability the students are required to possess a high level of I.T. skills to complete the assessment, this would be unfair for inexperienced computer users. The context of the research within this thesis will be in Higher Education therefore if marks are lost during the test as a result of poor usability this would potentially affect students' degree classification. These would be unacceptable consequences and a student may have grounds for appeal.

Figure 2 below outlines the initial research approach that will be adopted in order to answer the hypotheses.

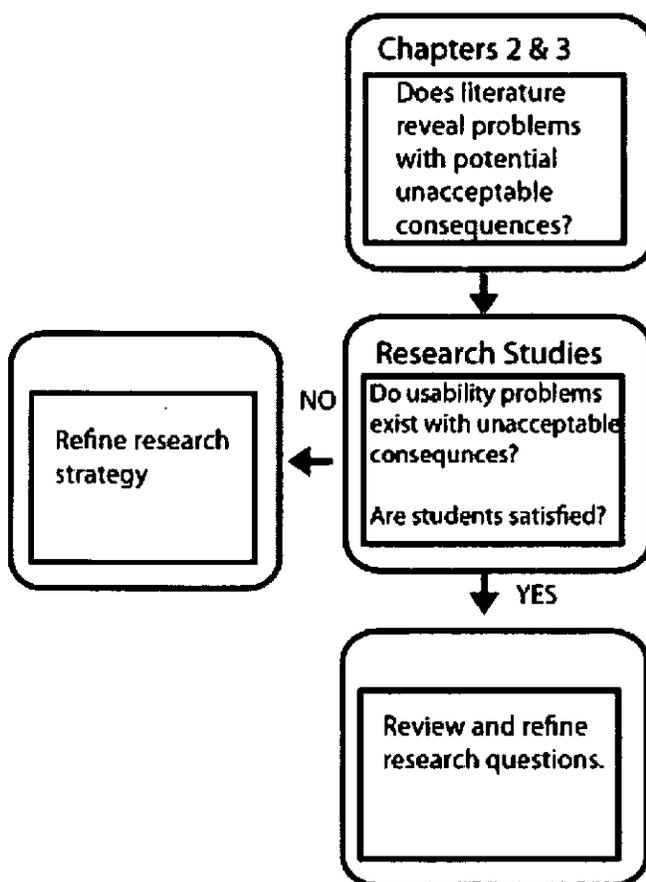


Figure 2 The initial research strategy adopted

A literature review will establish the extent to which usability has been investigated in the CAA domain and what problems have been reported. To devise a suitable research study, an analysis of usability evaluation methods will be performed which will aid the research design, in order to answer the initial hypotheses. It is anticipated that additional research questions will emerge as a result of the research studies. However, if no usability problems are revealed the focus of the research will have to change direction.

There are four main foci to the thesis, these being, assessment, CAA, usability, and evaluation techniques. To put the research in context, the thesis begins by identifying how CAA has overcome some of the problems associated with current assessment practices. CAA is then introduced by looking at the variations in question styles and applications to determine if problems identified would result in unacceptable consequences. It is envisaged that these problems are usability related therefore usability studies will be examined focusing on usability evaluation techniques that are commonly used to determine their suitability within the CAA domain.

1.1.1 Structure

The structure of the chapter is as follows; Section 1.2 introduces problems with current assessment techniques and the move towards CAA and Section 1.3 outlines the structure for the remainder of the thesis.

1.2 Overcoming Problems With CAA

Each technique of assessment presents its own difficulties, whether computer based or traditional. Many techniques, including essays, present the problem of double marking, in one study both markers agreed only 52 per cent of the time (Powers *et al.*, 2002). Additionally there are problems with cheating as Internet sites offer custom-written and off the shelf essays (Crisp, 2002). It has been suggested that exams tend to encourage surface learning (Race, 1995) and may cause increased anxiety resulting in significantly lower scores (Cassady & Johnson, 2002). The multiple choice question (MCQ) styles are used in both offline and CAA exams and raise a number of concerns, for example, grade deflation by not enabling partial credit (Baranchik & Cherkas, 2000), poorly designed questions (Jafarpur, 2003; Paxton, 2000) and guessing (Burton, 2001). However, the advantages of using computers to deliver MCQ for lecturers include automated marking (Pollock *et al.*, 2000) and for formative

Chapter 1

purposes the students have the opportunity to study at their own pace, repeat incorrectly answered questions and receive instant feedback (Loewenberger & Bull, 2003). These potential advantages of CAA have driven research into ways to overcome the difficulties.

Ultimately in an academic environment, the marks from summative assessment are accumulated to award an overall grade and there are concerns over comparability across subject domains. It has been suggested that the scientific subjects produce more First Class Degrees than the humanities because of the nature of the marking criteria in using the full range of marks and subjectivity is eliminated from the equation where there is a predefined correct answer (Horney, 2003; Yorke *et al.*, 2002). These findings would appear to be further corroborated by the Higher Education Statistics Agency (HESA) figures. Of the students graduating from UK universities in 2001/02, in Mathematical Science 25.5% passed with a First Class Degree, compared to 10.4% in Humanities (HESA, 2002) and this trend was also evident in other years for example, 1994/95 (HESA, 1995). CAA tends to use the full range of marks (like mathematics and some science subjects), therefore, the trend towards a high proportion of First Class Degrees may occur in other subject domains adopting this technique in the future.

There is pressure on lecturers not to fail students, and one study found that in professional subjects there is a tendency to leave the award of a fail to the next assessor (Hawe, 2003). Lecturers are confronted with emotional and ethical dilemmas when close working relationships with students are formed, increasing their reluctance to award a fail (Sabar, 2002). Emotional and subjectivity issues that are evident in human centred marking may be alleviated by automatic marking offered by CAA software.

It is evident that traditional methods of assessment within universities have their limitations. As a result of these limitations and also the continued increase in the use of technology to deliver curriculum, the gap between assessment methods and learning is widening. Many students are using computers to gain access to the teaching material and complete coursework or assignments however, they are required to complete the examination on paper which is a discontinuation from their learning experience. This along with the advantages of CAA has led to research into the use of technology for assessment purposes. It is important to recognise that some

of these issues discussed are still prevalent in CAA, especially issues related to MCQ, and there are new challenges.

1.3 The Thesis

As stated in Section 1.1 the objective of the research was “*To determine whether severe usability problems exist that can cause users difficulties and dissatisfaction with unacceptable consequences whilst using existing commercial CAA software application.*” How the objectives have been met is discussed in the conclusion, in Chapter 12.

Fixing the problems identified in the evaluations is beyond the scope of the research within this thesis, but has been incorporated into other methodologies such as RITE (Medlock *et al.*, 2002). The RITE method is concerned with three areas: Is it a problem? Do we understand it? Can we fix it? The first two questions fall into the scope of this research as without understanding the problem it would be difficult to determine the consequences for the user.

The thesis is structured in two parts, the first deals with *whether severe usability problems exist that can cause users difficulties and dissatisfaction with unacceptable consequences whilst using existing commercial CAA software applications*, and this is achieved by using survey methods to identify severe usability problems. The research then evolves due to limitations of survey methods and the second part focuses upon devising a set of CAA heuristics to enable educational technologists or software developers to evaluate the appropriateness of a CAA application.

The thesis identifies a number of the severe usability problems associated with CAA that would lead to unacceptable consequences. These are summarised throughout the body of this thesis. Chapters 5 and 6 use surveys to identify usability problems within CAA. Chapters 8, 9 and 10 use heuristics to devise a corpus of usability problems with unacceptable consequences. Finally, using an evidence based design approach, heuristics are synthesised for evaluating CAA applications; these are described in Chapter 11.

1.3.1 Structure of Thesis

This chapter outlines the main purpose of the thesis and provides an introduction to CAA within educational institutions, which is the domain under investigation. As

Educational Technology is such a broad area, it was felt important to focus on the terminology and highlight the limitations of traditional assessment techniques, in Section 1.2, to help emphasise the new challenges that may occur in implementing CAA. The overall structure of the thesis is displayed in Figure 3 below.

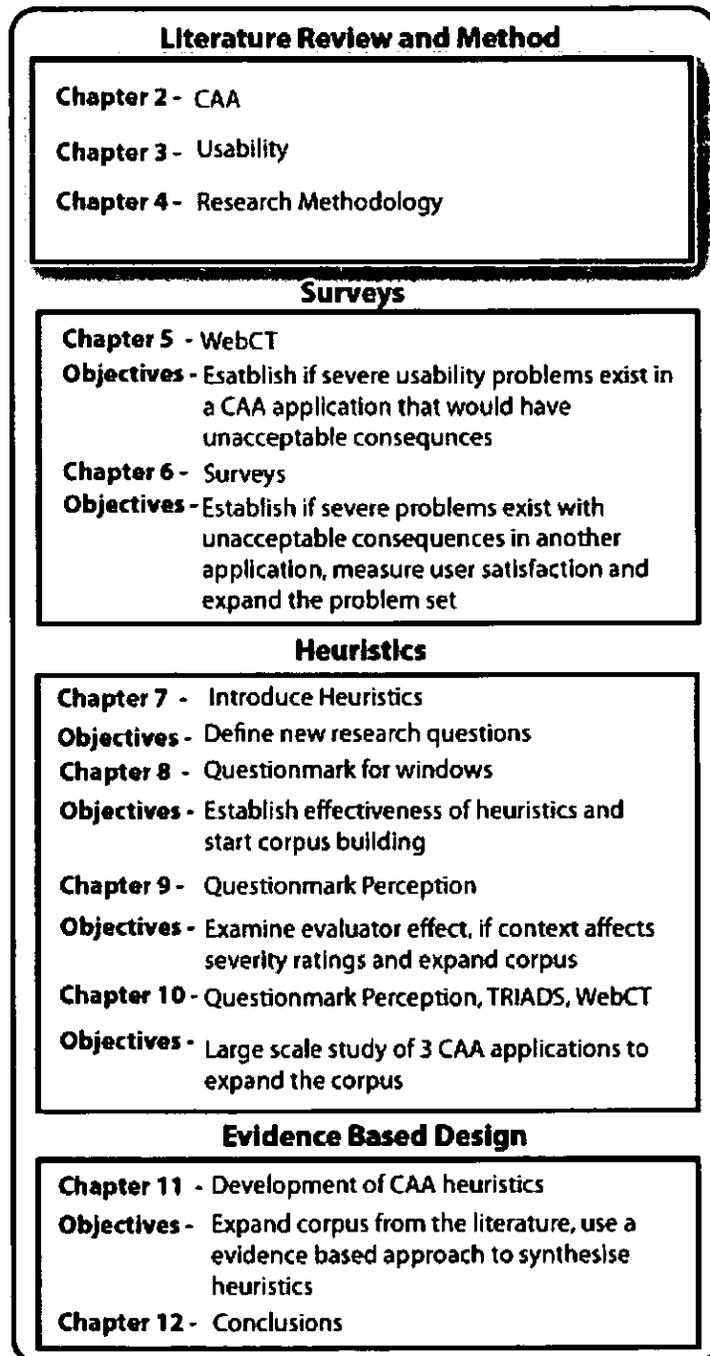


Figure 3 The structure of the thesis

1.3.2 Literature Review

As the research lies between two subject domains, Educational Technology and Human Computer Interaction (HCI), the literature review is broken into two sections with the final section looking at the merging of the two domains. Chapter 2 is used to introduce issues associated with assessment and CAA including, Types of Assessment (2.3); assessment techniques (Section 2.4); the variations of CAA (Section 2.5); its adoption (Section 2.6); software used in CAA (Section 2.7); the stakeholders within CAA (Section 2.8); testing cognitive skills (Section 2.9); question styles (Section 2.10) and issues surrounding guessing, accessibility institutional strategies and security (Section 2.11, 2.12, 2.13 and 2.14). Chapter 3 discusses how usability can be evaluated, highlighting some of the common metrics used and how they may not be applicable to CAA. It starts with a review of the literature regarding usability evaluation methods (Section 3.2) and finally usability and CAA is discussed (Section 3.3).

1.3.3 Methodology and Evidence Based Design

Chapter 4 discusses the research methodology used within this study, however, there is some overlap with the discussion in Chapter 3, as this focused on usability evaluation methods. The ethical implications of using the various methods are discussed (Section 4.6). The objectives of the research changed in Chapter 7 and a new hypothesis was formulated relating to heuristic evaluations. It was anticipated that domain specific heuristics would be required for CAA and an evidence based design approach for the synthesis of heuristics is proposed in Chapter 7 and the remaining chapters in the thesis apply this approach.

1.3.4 Survey Studies

The body of the research begins in Chapter 5; here, survey methods are used to establish user satisfaction, a component of usability as defined by ISO 9241/11 (ISO, 1998), and determine *whether severe usability problems exist that can cause users difficulties and dissatisfaction with unacceptable consequences whilst using existing commercial CAA software applications, and to identify the usability problems that may occur*. The first study focuses on the assessment tool within WebCT®, whilst the study described in Chapter 6 uses a modified survey instrument and examines

Questionmark® Perception. The limitations of using survey methods within summative test conditions are discussed in Section 6.6.2.

1.3.5 Heuristic Evaluations

A substantial part of the thesis is devoted to using heuristic evaluations for extracting usability problems from three commercial CAA applications: WebCT®, Questionmark® and TRIADS®. The purpose of the heuristic evaluation studies is to extend the problem set derived from the user studies. Having established that usability problems exist in commercial CAA applications, and that, overall, users are satisfied with this form of assessment, the next stage is *to identify the severe usability problems that may occur by developing a corpus of problems*. This had partially been addressed in Chapters 5 and 6 however, using a survey method, users provided little qualitative feedback on their experience, inter-group consistency was low and the yield per evaluator was very low, therefore another method was required to identify usability problems.

Chapter 8 reports a pilot study using Nielsen's Heuristics (Nielsen, 1994a), with HCI experts evaluating Questionmark® for Windows® from two different perspectives; formative and summative assessment. Section 8.3 shows that heuristics can find usability problems within CAA applications but questions the effectiveness of using Nielsen's heuristics in Section 8.3.7.

Chapter 9 describes a study, using the same heuristics, evaluating Questionmark Perception®, which differed from the first software in being a web based delivery. This study looked at the relationship between novice and expert evaluators, and whether context (formative or summative assessment) would affect the classification of severity ratings. The problems found further expanded the corpus.

In Chapter 10 WebCT®, Questionmark® and TRIADS® are evaluated using Nielsen's heuristics (Nielsen, 1994a). This was a large scale evaluation using 98 HCI students and a between subjects design. A persistent problem was how to merge and aggregate the data from multiple evaluations and in Section 10.2.5 a Damage Index formula is reported that enables the systematic merging and prioritising of usability problems. Section 10.3 revealed that all three applications had severe usability problems that *could cause users difficulties and dissatisfaction with unacceptable consequences* however, Section 10.3.7 showed that many of these problems were

unique to the application with only a small percentage of overlap. The usability problems identified within these chapters help inform the synthesis of domain specific heuristics for CAA.

1.3.6 Evidence Based Design of Heuristics for CAA

In Chapter 7, an evidence based design approach for the creation of domain specific heuristics is proposed and in Chapter 11 this is applied to the CAA domain. Extracting evidence from three sources, a set of heuristics are synthesised for CAA and the methodology is critiqued in Section 11.8.1.

1.3.7 Conclusions and Further Research

The conclusions of the thesis are found in Chapter 12. There is a discussion about the major contributions within this research and some suggestions are offered in relation to improving the process of synthesizing heuristics using the evidence based design approach, leading to further research in the area of validating heuristic sets and comparing evaluation methods.

1.4 Conclusions

This chapter identifies the purpose of the thesis and places the work within the context of CAA. Limitations of assessment techniques are highlighted to draw attention to concerns about existing practices which technology may be able to alleviate. The first part of the thesis is made up of two chapters reviewing CAA and usability evaluation methods and the main body is subdivided into three areas: surveys, generic heuristic evaluations and an evidence based design approach to designing domain specific heuristic.

1.4.1 Publications Related to the Thesis

Whilst conducting the research presented throughout the thesis a number of publications have arisen listed below:

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

I was the primary researcher in this publication and contributed 90% of the work. This forms the basis of Chapter 2.

Chapter 2

Sim, G., & Holifield, P. (2004a). *Computer Assisted Assessment: All those in favour tick here*. Paper presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Lugano.

I was the primary researcher in this publication and contributed 90% of the work. The data from this study is used in Chapters 5 and 6.

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

I was the primary researcher in this publication and contributed 85% of the work. The data from this paper is used to expand the corpus in Chapter 6.

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

I was the primary researcher in this publication and contributed 80% of the work. This forms the basis of the research in Chapter 5.

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.

I was the primary researcher in this publication and contributed 75% of the work. This study forms the basis of the research in Chapter 9.

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.

I was the primary researcher in this publication and contributed 80% of the work. This study forms the basis of the research in Chapter 10.

Sim, G., Read, J. C., & Cockton, G. (2009). *Evidence based Design of Heuristics for Computer Assisted Assessment*. Paper presented at the 12th IFIP TC13 Conference in Human Computer Interaction, Uppsala.

I was the primary researcher in this publication and contributed 70% of the work. This study forms the basis of the research in Chapter 11 with the synthesis of the heuristic set.

Chapter 2 CAA Overview

2.1 Introduction

Within the literature regarding CAA there is a lack of universal consent regarding the terminology and its definition. Terminology that has been used in the literature include computer based testing (Chalmers & McAusland, 2002; Lloyd *et al.*, 1996); computer aided assessment (Dowsing, 1998); e-assessment (Ashton & Bull, 2004); and computerised assessment (McLaughlin *et al.*, 2004b). CAA is defined by Chalmers and McAusland (2002) as the use of computers to assess student progress, this is also supported by Bull and McKenna (2001) who 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 the activities. Therefore, the definition of CAA used in this review will be that: CAA encompasses the use of computers to deliver, mark or analyse assignments or exams.

2.1.1 Objectives

The purpose of this chapter is to provide an introduction to assessment and computer assisted assessment, to examine the ways in which it has been used within higher education, to identify potential unacceptable consequences, and to investigate some of the key challenges for implementing CAA as part of the overall assessment strategy.

2.1.2 Scope

This chapter starts with an introduction to assessment from a historical perspective in Section 2.2, Types of Assessment are discussed in Section 2.3 and Section 2.4 discusses assessment techniques. With respect to CAA a decision was made to focus the literature review by predominately focusing on objective tests administered through CAA. Section 2.5 discusses variations of CAA but the main research is constrained to objective tests. Section 2.6 discusses the adoption of CAA and is constrained to looking at the UK educational system, with Section 2.7 examining the software used there. The diverse set of users is explored in Section 2.8 along with their goals. Research relating to objective tests is vast and the literature review did

not examine areas of question interoperability and writing questions, but focussed on testing higher cognitive skills using objective tests (Section 2.9) and question styles (Section 2.10). This leads to a discussion relating to grade inflation caused by guessing and accessibility issues (Sections 2.11 and 2.12). Accessibility has to be constrained to issues specifically relating to CAA as this is a huge subject domain. The final two sections examine the implementation and security implications of incorporating CAA.

2.2 Assessment

The concept of assessment has been used within society for a considerable time. Emperor Shun in 2357 BC used written examinations that formed the basis for admission and promotion within the civil service of ancient China (DuBois, 1964). There are various definitions of assessment dependent on the context in which it is being used. Walsh and Betz (1985) define psychological assessment as the process of understanding and helping people cope with problems. In contrast an educational perspective indicates the main purpose of assessment is to measure the outcomes of learning (Burke, 2002). The Quality Assurance Agency (QAA) expects that things valued enough to be stated as learning outcomes will be assessed (Knight, 2001). McAlpine (2002) defines assessment as a form of communication. Lynch (2001) refers assessment to the systematic gathering of information for the purpose of making decisions or judgements about individuals.

Four roles of assessment have been identified: formative, summative, certification and evaluative (Horney, 2003). Whilst Rowntree (1987) identifies five purposes of assessment: selection, maintaining standards, motivation, providing feedback to teacher and students, and preparation for life. These definitions appear to overlap for example the role may be formative but the purpose maybe to motivate the student, provide feedback to both the student and the teacher. Conole and Fill (2005) devised a learning design toolkit with assessment broken down into two main components type and techniques represented diagrammatically. Figure 3 is a modification of this diagram, with CAA removed from techniques and the addition of a third component, Delivery.

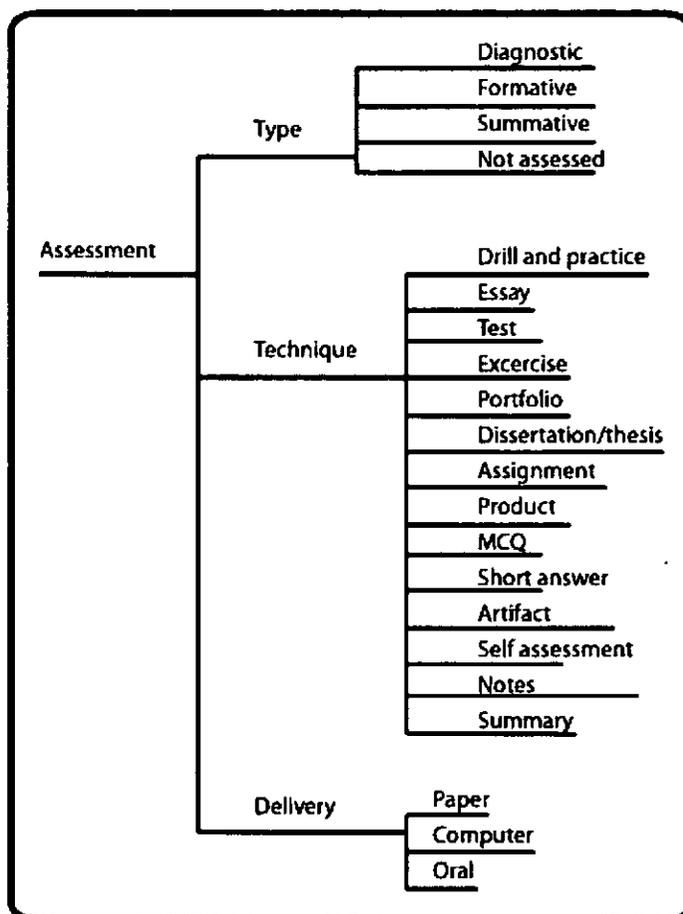


Figure 4 Assessment broken down into components (after Conole and Fill)

CAA was removed from techniques as the computer is usually used to facilitate a particular technique, for example the completion of a multiple choice test or essay and this is further discussed in Section 2.10. Delivery was added to the diagram to represent the different ways the techniques could be administered.

As assessment has many types, it is, therefore, essential to clearly define the meaning of diagnostic, formative or summative. This will be discussed in the next section in order to put the work in this thesis in context.

2.3 Types of Assessment

Diagnostic assessment is conducted either before or after an activity to ascertain the students' knowledge. This could be used to determine eligibility for employment or exemptions from a course of study. *Formative* assessment is carried out during the course to determine the effectiveness of teaching and provide feedback to the students on their performance. In contrast *Summative* assessment is conducted with

the purpose of making judgement about students' performances. Summative assessment is the area to which students attach the most importance as it determines their overall grade (Taras, 2002).

The thesis does not consider diagnostic assessment, but focuses on formative and summative as this plays a greater role within Higher Education. The user evaluations of CAA interfaces in Chapters 5 and 6 have predominately been conducted under summative assessment conditions. However, further studies have been conducted examining the interface from a formative perspective in Chapter 8 and 9.

2.4 Assessment Techniques

Academic assessment can be administered through various techniques. Fifty varied techniques have been identified and used within higher education for assessment purposes (Knight, 2001) and some of the most commonly used are exams and essays (Graham, 2004). However, this does not include all the techniques now available within CAA packages, for example, incorporating questions that make use of multimedia. Many of the techniques have inherent problems, whether administered via traditional methods or by computer, and these are discussed in the next section.

2.5 Variations in CAA

Within higher education institutions the application of CAA has occurred in a number of varied ways, these include, adaptive testing (Latu & Chapman, 2002; Mills *et al.*, 2002; Rudner, 1998), analysis of the content of discussion boards (Macdonald & Twining, 2002; Wiltfelt *et al.*, 2002), automated essay marking (Burstein *et al.*, 2001; Christie, 1999), delivery of exam papers (Sim *et al.*, 2003) and objective testing (Pain & Le Heron, 2003; Walker & Thompson, 2001). Figure 5 diagrammatically represents the variations in CAA and examples of software used within CAA.

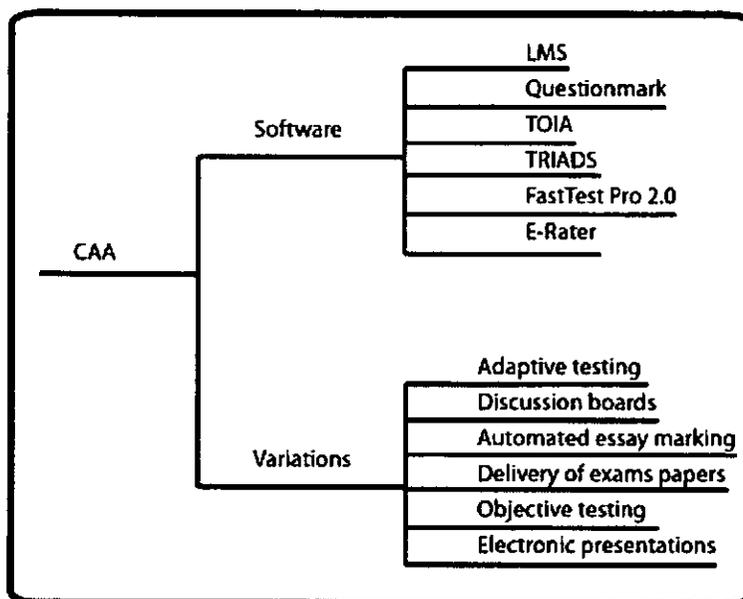


Figure 5 CAA Software and Variations

These methods vary considerably, however the focus of this review of research will centre on the issues relating to implementing objective tests via CAA as this is one of the most widely used forms of CAA within Higher Education.

2.6 Adoption of CAA

Over a decade ago it was stated that many universities were using technology in their assessment strategies (Stephens & Mascia, 1997) but institutional barriers still exist that prevent more widespread adoption (Conole & Warburton, 2005), this is further discussed in Section 2.13. However, the majority of students today, who enter higher education directly from schools and colleges, are likely to have been exposed to Information Technology as part of the UK National Curriculum. Government initiatives are also driving the adoption of CAA, for example, the Department for Education and Skills was developing, in conjunction with the Qualifications and Curriculum Authority (QCA), a key stage 3 communication technology test that would be administered using CAA. It was envisaged that the test would become statutory within 2008, however, on the 14th October 2008, the Secretary of State announced that the key stage 3 national curriculum tests are no longer statutory for 2009 (NAA, 2008). Despite this, pilot studies conducted within schools for the delivery of summative assessment via the web (Ashton *et al.*, 2003; Nugent, 2003) and for basic key skills tests in both Learn Direct and army centres

(Sealey *et al.*, 2003) indicate that CAA can successfully assess students and provide timely feedback regarding class and individual progress. Chapman (2006) surveyed the 115 awarding bodies recognized by the QCA within the UK regarding their usage of CAA. With a response rate of 81%, 38% of the awarding bodies currently used CAA to deliver up to 60% of their assessments. It is expected that this uptake is likely to continue, therefore, putting greater pressure on higher educational institutions to adopt CAA, and there may be a certain expectation from students that this will be one of the assessment methods they encounter. Therefore, it could be argued that for many students CAA may become a more widely used method of assessment in schools, further education institutions and universities.

There is also evidence to suggest students find CAA an acceptable assessment technique and prefer this to other forms of assessment (Ricketts & Wilks, 2002a; Sambell *et al.*, 1999; Croft *et al.*, 2001; Sim & Holifield, 2004a). This along with the other drivers, such as Government policy, may see in an increase adoption of CAA in Higher Education Institutes.

2.7 CAA Software

With the increased adoption of CAA within educational institutions there has been a rise in the number of commercial and bespoke CAA applications that enable the construction and administration of objective tests. These (objective tests) are tests where the answer is predefined, as seen in multiple choice questions. Commercial applications include Questionmark Perception®, I-Asses®, TRIADS® and Hot Potatoes®, whilst two examples of bespoke applications are TOIA, developed through JISC funding with a number of UK universities, and CASTLE, developed at the University of Leicester. Many universities have adopted learning management systems that incorporate assessment tools such as WebCT®, Blackboard® and Moodle®. Although all these applications enable the delivery of objective tests they vary in the number of question styles available, for example WebCT® in 2005, only offered a limited number of questions styles, nine in total, compared to dedicated systems such as Questionmark® which offered 18 in 2005. They also differ in layout and interface design attributes, even within the same application a number of different templates may be available. Academics or educational technologists may not question the suitability of these templates if they are not experienced within HCI, making the presumption that the software manufacturer has evaluated the

Chapter 2

appropriateness. For example, Figures 6 and 7 show the same test displayed in two different templates within Questionmark®. They are considerably different in design and the interaction would be different, with the navigation placed on the right in Figure 6 and across the bottom in Figure 7. The navigation in the second image may be more difficult than the first due to horizontal scrolling.

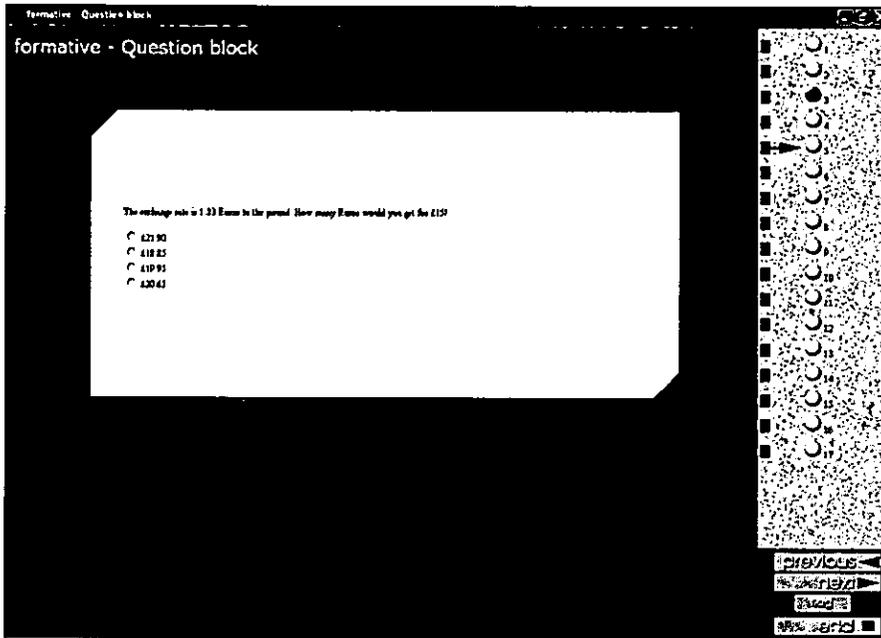


Figure 6 A Questionmark template with navigation on right (2005)

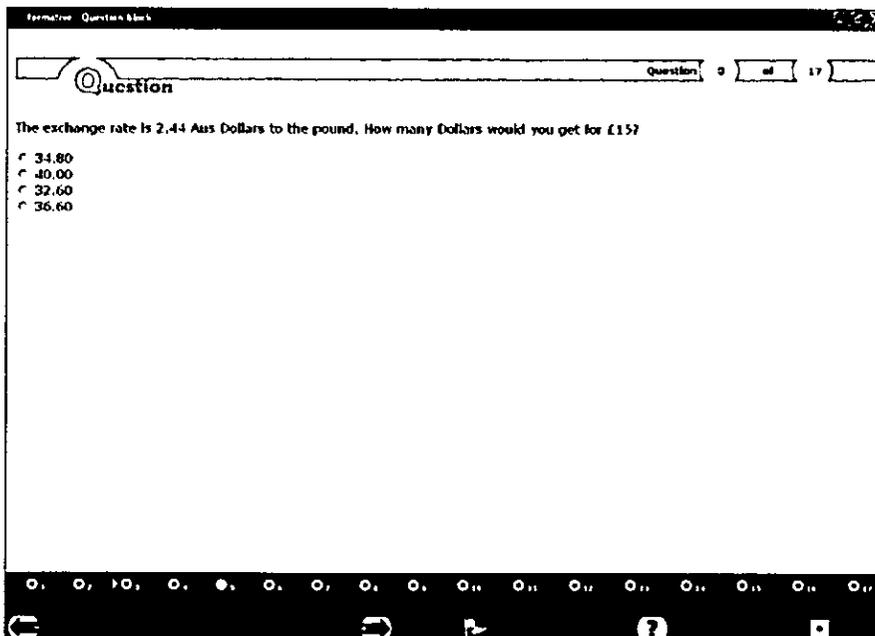


Figure 7 Questionmark Template with vertical navigation (2005)

2.8 Stakeholders within CAA

Smythe and Roberts (2000) identified nine potential user groups within a CAA environment, each with different requirements, however they did not clearly specify these requirements. These are:

- Author
- Psychometrician
- Assessor
- Scorer
- Candidate
- Invigilator / Proctor
- Administering Authority
- Administrator
- Tutor

The groups range from academics authoring the questions, invigilators starting the exams and students participating in the test. However, not all these stakeholders will be prevalent in every exam, for example, within the authors' institution no psychometricians have been involved in authoring the questions. With so many different user groups and requirements, ensuring the CAA application is useable may be difficult; this is discussed further in Chapter 3.

Of the nine different user groups, it is the students using the software to complete the exam that have the most to lose as a consequence of poor usability as their grades may be affected. The remaining chapters in the thesis will, therefore, address usability from the student's perspective as the end user. The concern is that if software cannot be used intuitively this can often lead to an increase in the rate of errors (Johnson *et al.*, 2000) and this could be detrimental to student results.

2.8.1 Student's Goal

The overall goal of the student is to complete the assessment, and this can be broken down into several tasks: start the exam, answer the questions, navigate between pages and end the exam (Sim, Horton, & Strong, 2004). It is feasible that students

Chapter 2

may encounter severe usability problems that can cause difficulties and dissatisfaction with unacceptable consequences whilst completing these tasks. Based on the authors experience of using Questionmark® and WebCT® in the studies reported in Chapters 5 and 6 a UML activity diagram was synthesised, demonstrating the user interaction with a CAA application see Figure 8.

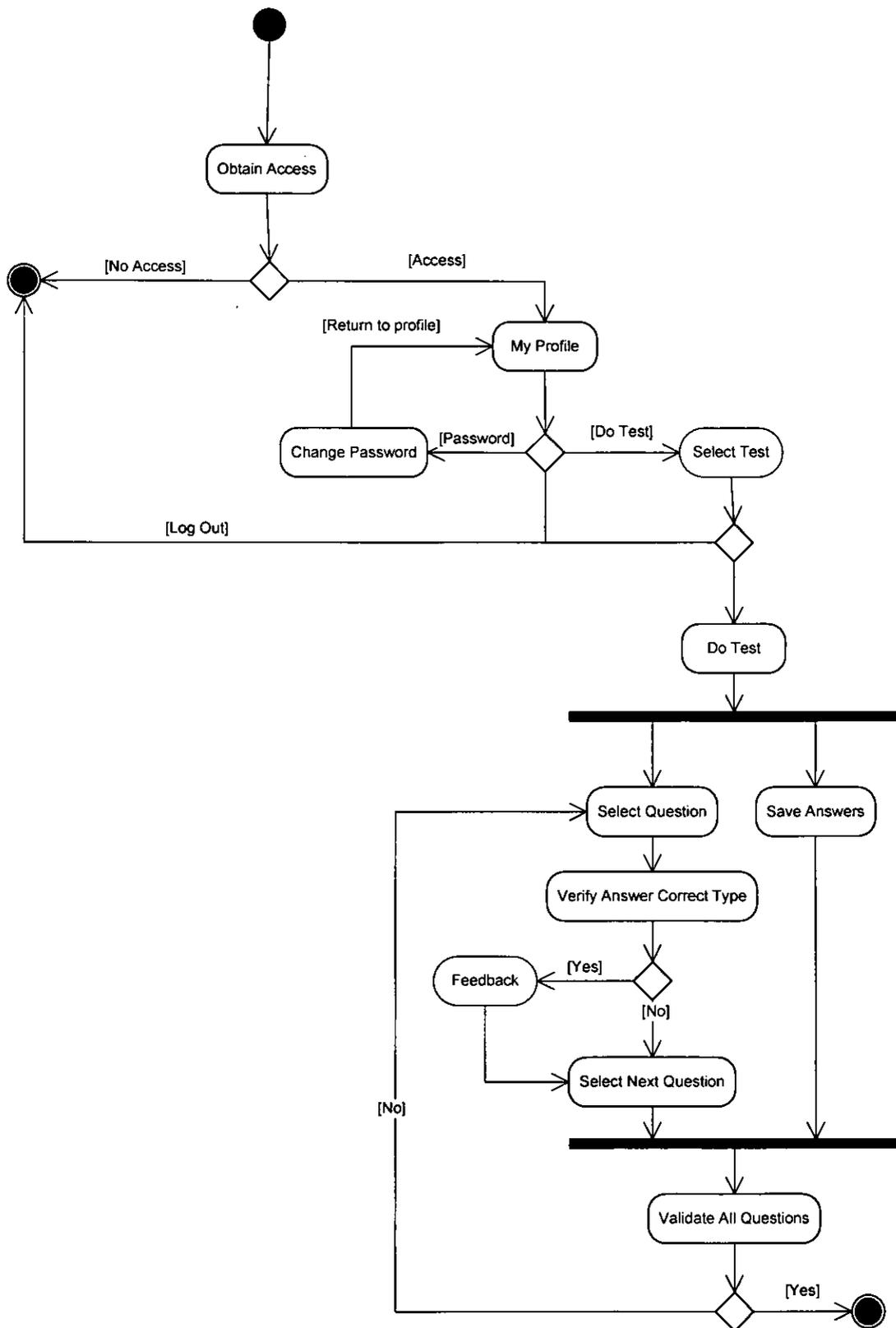


Figure 8 Activity Diagram of user interaction with a CAA environment

2.8.2 User Tasks

Unless the application is open access, the first stage requires users to enter a password to gain access. This password would then be validated to gain access to the users profile.

Once in the application users usually have the option to change their password or select a test that has been allocated to them. At either of these stages it is feasible for them to leave the application. It is possible that they may have selected a non active test and, therefore, may also exit at this stage, or if the test is active, they can do the test.

The next stage is a continual process of selecting a question, answering the question, (this is then validated by the system and, if feedback is enabled at this stage, this feedback is displayed), selecting the next question and saving their answer. The save function after each question is optional in some systems but a recommended practice in case of system failure (Ricketts & Zakrzewski, 2004).

Once users decide to terminate the exam then the questions are usually validated to ensure that all questions have been answered, allowing users to return to the test if they desire.

2.9 What can CAA Test?

There is concern in the literature, relating to CAA, in its ability to test higher cognitive skills across subject domains (Daly & Waldron, 2002; Paterson, 2002). The higher cognitive skills are often associated with 'Analysis, Synthesis, and Evaluation', as defined in Bloom's Taxonomy (Bloom, 1956). However, a revised taxonomy takes into consideration the 'Knowledge Dimension' (Anderson & Krathwohl, 2001) and this has also been used in CAA research for classification of questions (King & Duke-Williams, 2002; Mayer, 2002) who suggest that higher cognitive skills can be assessed through CAA.

Paterson (2002) indicated that it is not feasible to test the higher-level cognitive skills using CAA within mathematics. Bloom *et al.* (1971) states that, in the majority of instances, Synthesis and Evaluation promote divergent thinking and answers cannot be determined in advance, which is a requirement for objective tests. Heinrich and Wang (2003) argue that objective testing is still not sophisticated

enough to examine complex content and thinking patterns. However, other research in linguistics and computer programming concluded that higher-level skills can be assessed via CAA through innovative approaches (Reid, 2002; Cox & Clark, 1998). In the study by Reid (2002) a new language was devised and students were required to apply linguistic techniques in order to answer Multiple Choice Question (MCQ): it has been suggested that CAA tests of higher-level skills are more complex and costly to produce (Dowsing, 1998) and this may be because more innovative approaches are needed. If more innovative question styles are adopted then this could increase the complexity of the interaction and interface layout, affecting the overall ease of use. For example, an assertion/reason combines elements of multiple-choice and true and false question types and this style is believed to assess higher level skills.

2.10 Question Styles

Objective testing has been used within assessment for a considerable period of time (Wood, 1960) and computer programs delivering MCQ date back to the 1970s (Morgan, 1979). With the evolution of technology and research, more sophisticated question styles have emerged enabling diverse assessment methods. The question styles delivered by the TRIADS® software, developed at Derby University, are evidence of this evolution, offering seventeen question styles in 1999 (Mackenzie, 1999) and thirty nine in 2003 (CIAD, 2003). Faculty members at the University of Liverpool using TRIADS® found that this presented an additional problem, as they were unfamiliar with the new question styles and lacked confidence in writing suitable questions (McLaughlin *et al.*, 2004a). To overcome these problems, staff development in writing suitable questions and following recommended guidelines is suggested. For example, generic guidelines developed by Haladyna (1996) or Herd and Clark (2002) present examples of the various questions styles used in Further Education whilst examples used within Higher Education can be found at <http://www.caacentre.ac.uk>. Complex question styles and poorly written questions may result in usability problems for students.

Although there are a large number of possible formats for CAA questions, it is possible to classify them into four distinct groups based on the human interaction technique required (CIAD, 2003), this is represented by Figure 9.

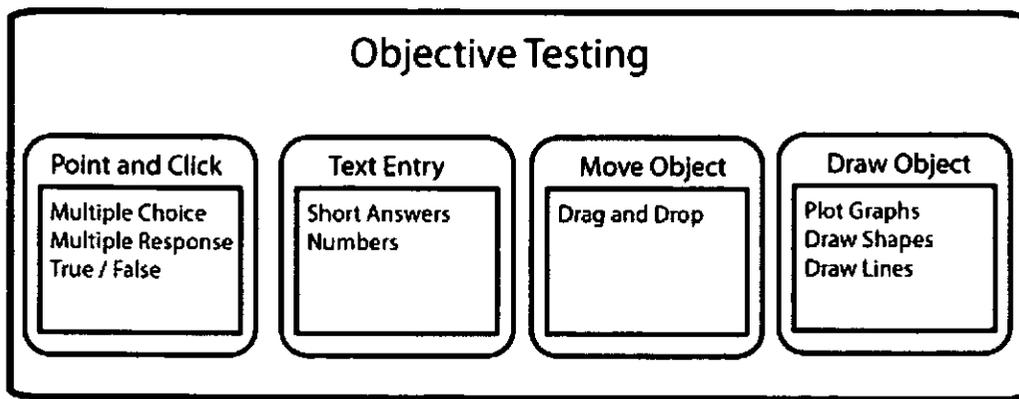


Figure 9 Objective testing grouping by interaction type

These groups are defined as point and click, move object, text entry and draw object. In using these questions styles, problems may arise with unacceptable consequences.

2.10.1 Point and Click

Point and click includes MCQ and Multiple Response Question (MRQ) items, which have both been used within assessment practise for a considerable time and as a result are often transformed into CAA (Ricketts & Wilks, 2002). Ebel (1972) suggests that any understanding or ability that can be tested by means of any other technique, for instance essays, can also be tested by MCQ. More complex MCQ questions can be devised through assertion-reasoning, resulting in the testing of higher cognitive skills (Bull & McKenna, 2001). This question style presents a MCQ in two parts, there is an initial statement which students have to determine whether it is true or false, followed by a number of statements relating to the cause. Both MCQ and MRQ have inherent problems that have previously been discussed in Section 1.2. There are concerns about relying on true and false style questions as they often lead to emotional reactions and students perceive them to be unfair (Wood, 1960). Davies also argues that the quality of MCQ is dependent on the quality of the distracter and not the question (Davies, 2002). Many of the problems discussed are in relation to test design and do not examine the implication of user interaction with such question styles.

2.10.2 Text Entry

Text entry, within objective testing, centres upon the input of short predefined answers, such as factual knowledge, or syntax in computer programming. An

advantage of this format is that students must supply the correct answer, removing the possibility of guessing (Bull & McKenna, 2001), and this style has been found to be the most demanding format for students (Reid, 2002). There are problems associated with text entry within some subject domains, such as mathematics, as mathematical expressions cannot easily be included in most commercial software (Croft *et al.*, 2001; Paterson, 2002). Students are also concerned that they may be penalised for inaccurate spelling (Sim *et al.*, 2006a), this may affect their overall satisfaction with the technique. Also the time saving benefit may be reduced if lecturers need to manually check for spelling errors.

2.10.3 Move Object

Move object style questions focus on the movement of objects to predetermined positions on the screen. They are a variation of the MCQ format and are good for assessing students understanding of relationships (Bull & McKenna, 2001). For example, in computing they could be used for the labelling of entity relationship diagrams or in linguistics, students could be presented with a poem and would be required to move the highlighted words to the appropriate word class. One problem is that when the number of moveable objects is equal to the number of targets, if students know all but one answer they will automatically get full marks (Wood, 1960). There may also be accessibility concerns as often this style of question relies on the mouse for answering the question. Move object is not supported in a number of CAA applications for example WebCT®, Hot Potatoes and CASTLE. However, it is possible to embed Flash objects into TOIA which would enable this type of interaction but would require technical expertise to generate the questions.

2.10.4 Draw Object

Draw object is associated with drawing simple objects or lines. For example, students may be required to plot graphs which can be automatically marked. This style of question is found to be a high discriminator between strong and weak candidates (Mackenzie, 1999). There is little evidence in the literature concerning the effectiveness of this format, but this might be due to the fact that commercial software such as Questionmark® and I-Assess® do not have this style in their templates. If the input method is the mouse then similar problems may arise to the move object style of question, Section 2.10.3.

2.11 Guessing

A number of the question styles associated with CAA can lead to artificially high marks through guessing (Bush, 1999). This has implications for setting the pass mark of the test. For example, setting a pass mark of 40 per cent then basing the assessment on true/false answers would be inappropriate, as guessing alone would give an average of 50 per cent (Harper, 2002). The problems of guessing may be addressed through various marking schemes, such as post test correction (Bull & McKenna, 2001), negative marking (Bush, 1999), increasing the number of questions or combining the results from several tests (Burton & Miller, 1999), or increasing the number of distracters and the pass mark (Mackenzie & O'Hare, 2002). It has been suggested that negative marking is not generally implemented in the UK (McAlpine, 2002) and that post test correction is only suitable with a single question style because the formulae would vary depending on the number of distracters (Harper, 2003).

Statistical analysis has resulted in various methods being developed to assist in test construction in order to reduce the effects of guessing. An empirical marking simulator to assist in scoring and test construction, based on a base level guess factor, has been developed (Mackenzie & O'Hare, 2002). This program examines the mark distribution and measurement scale for a set of random answers, enabling tutors to establish the effects of guessing on their assessment. Also statistics to award a score for partial credit through a formula based on a mean uneducated guessers' score has been investigated (McCabe & Barrett, 2003). This allows MCQ to be unconstrained, similar to MRQ styles, enabling students to provide more than one answer and their score is weighted depending on the number of choices. For example, a MCQ with one correct answer, four possible options and a score of 3, if a student includes the correct answer by selecting 2 options they would only score 2 ($2=3-1$). Davies used a combination of predetermining the students' confidence in answering the question prior to seeing the distracters and negative marking, resulting in students perceiving this to be a fairer test of their abilities (Davies, 2002).

These techniques of partial credit and confidence base questions are not readily available as question styles in the majority of CAA applications and, therefore, are

not evaluated in the studies reported in Chapters 5-10. Students would need careful training in the process of answering the questions as it was found, using negative marking, students were concerned they did not fully understand the scoring associated with each question style (Sim *et al.*, 2007). It could be argued that these techniques may be unnecessary if the tests are well constructed (Bull & McKenna, 2001).

2.12 Accessibility

Within the UK, institutions need to comply with the Special Educational Needs and Disability Act when preparing both teaching and assessment material (SENDA, 2001). The number of students in UK higher education registering a disability in 2000 was 22,290 (Phipps & McCarthy, 2001) and in 2008 at undergraduate level there were 51,275 (HESA, 2008) and this has implications for CAA. For example, a student with dyslexia may exert more cognitive effort in interpreting the question, therefore, ensuring the language is appropriate is a necessity (Wiles & Ball, 2003). In addition extra time may be required to complete the test which may necessitate the publishing of two different assessments, one with a longer duration. Feedback from one dyslexic student regarding CAA indicated that he or she thought it provided a more level playing field in which he or she can demonstrate their knowledge (Jefferies *et al.*, 2000). Students with visual or physical impairment may struggle to answer move object and draw object style questions without the aid of assistive technology. They may need specially adapted input software and hardware such as, touch screens, eyegaze systems, or speech browsers.

There are guidelines for supporting accessibility in general teaching, however there is little evidence that guidelines for inclusive and accessible design in CAA are emerging (Wiles, 2002). For example, when multimedia elements, such as video are used within the assessment, it may necessitate the provision of an alternative paper based version for students with sensory impairment. The introduction of an alternative, in this instance paper, poses the problem of ensuring comparability (Al-Amri, 2007; Bennett *et al.*, 1999). When identical tests are presented on a computer and paper they are not comparable (Clariana & Wallace, 2002) because there are numerous variables that impact on a student's performance when questions are presented on a computer. These variables include the monitor (Schenkman, Fukuda, & Persson, 1999), the way text is displayed on screen (Dyson & Kipping, 1997) and

the problems of obtaining a feel for the exam when only a single question is presented (Liu *et al.*, 2001).

The Web Accessibility Initiative (<http://www.w3c.org/WAI/>) has produced useful guidelines for promoting online accessibility which may be applicable to CAA but this initiative does not address the issue of comparability between questions.

2.13 Institutional Strategies for the Adoption of CAA

The greatest barrier to the adoption of CAA by academics is shortage of time, to both develop questions and learn the software (Warburton & Conole, 2003). This may have contributed to the fact that the adoption of CAA has usually resulted from the impetus of enthusiastic individuals rather than strategic decisions (Daly & Waldron, 2002; O'Leary & Cook, 2001). The perceived benefits of CAA of freeing lecturers' time can be elusive if no institutional strategy or support is offered (Stephens, 1994), successful implementation may be left to chance (Stephens *et al.*, 1998) and CAA may be developed in an anarchic fashion (McKenna & Bull, 2000). Research conducted at the University of Portsmouth indicates that there is no time saving benefit for courses with less than twenty students (Calleary, 1997). In order to utilise the features within software packages staff training and development is necessary (Boyle & O'Hare, 2003) and this may not be feasible without institutional support.

Institutions adopting CAA are faced with the difficulty of evaluating and deciding upon the most appropriate CAA software. Without an institutional strategy, individual departments may adopt their own systems as evident at the University of Bristol, which is utilising a range of CAA software within different departments (O'Leary & Cook, 2001). This can result in students having to cope with a number of different user interfaces and CAA formats, increased licence costs and problems offering administrative and technical support. Even if an institution has a clear strategy there are also problems in determining the selection criteria for software used to deliver assessment and there is a lack of analysis within the literature (Valenti *et al.*, 2002). Sclater and Howie (2003) contributed to this literature by defining the ultimate online assessment engine. This was achieved through a process of examining the user requirements of the system, establishing the stakeholders and

their functional requirements. This research may aid institutions identify their needs and establish an appropriate evaluation methodology.

The following guidelines for an institutional strategy have been formulated by Loughborough University and the University of Luton (Stephens *et al.*, 1998):

- establish a coordinated CAA management policy for CAA unit(s) and each discipline on campus;
- establish a CAA unit; establish CAA discipline groups/committees;
- provide funding; organise staff development programmes;
- establish evaluation procedures;
- identify technical issues;
- establish operational and administrative procedures.

BS7988 is a British Standard Code of practice that has been introduced governing the use of information technology in the delivery of assessments (BS7988, 2002). The guidelines have various implications for the delivery of assessments, for example, it is recommended that students take a break after 1.5 hours which has an impact on the invigilation process. If this recommendation is followed, procedures need to be established to prevent collusion between students during the break, or the tests need to be split into two separate sections. One of the difficulties for many institutions using CAA arises through the lack of resources to accommodate large cohorts of students sitting the exam simultaneously (Mackenzie *et al.*, 2004). This problem can be alleviated through institutional support by using library facilities and therefore, to fully utilise the benefits of CAA, an institutional strategy would appear necessary to increase the chance of successful implementation. These benefits are evident within a number of institutions with strategies, such as, Ulster (Stevenson *et al.*, 2002), Derby (Mackenzie *et al.*, 2002), Coventry (Lloyd *et al.*, 1996) and Loughborough (Croft *et al.*, 2001) where they have successfully embedded CAA into their teaching strategy.

2.14 Security Issues with CAA

The move from traditional teaching environments and examination settings, presents additional issues relating to security. Frohlich (2000) states that in traditional

Chapter 2

environments it is possible to ensure the security of the exam papers and scripts, this includes the transportation to and from the exam venue. However, even under this system breaches in security do occur, for example AQA had to replace 500,000 English and English Literature exam papers after a box had been tampered with (Curtis, 2003).

Tannenbaum (1999) defines security in computer systems as consisting of procedures to ensure that persons or programs cannot access material for which they do not have authorisation. This is essential within a CAA environment as questions and student details are stored in a database and usually the test data is sent over a local network or the Internet. Before computers were connected to the Internet it was relatively easy to have effective security measures (Mason, 2003), but transmission of sensitive data over an insecure network requires additional security measures to be implemented.

Using computers in the delivery process, encryption techniques can be used to ensure the security of the questions and answers when transmitting data over the Internet (Sim *et al.*, 2003). To increase security, examinations can be loaded on to the server at the last minute (Whittingham, 1999) and if email is used to submit results there is a potential risk due to the lack of authentication (Hatton *et al.*, 2002). In using CAA four security requirements have been identified by Luck and Joy (1999), these being: all submissions must be logged; it must be verified that a stored document used for the assessment is the same as the one used by each student; a feedback mechanism must inform students that their submission has been received; and the identity of each student must be established.

With the majority of CAA software, students and administrators are required to have passwords and these are often the weakest link in terms of protection (Hindle, 2003). Although an unlikely event, students could get access to the administrator password and change their results or gain access to the questions. Other concerns are authentication and invigilation of the students. These are security issues that institutions have always had to deal with but are particularly problematic in remote locations (Thomas *et al.*, 2002). At present students enrolled on distance learning courses overseas need to sit exams in a specific location, such as the British Council Offices, to enable authentication and invigilation. Research is being conducted to

overcome these problems, but unless solutions are found, geographical barriers will remain as students need access to the test centres.

During the test computers need to be locked down, removing the possibility of accessing other content and secure browsers have been developed to enable this, such as Questionmark Secure (Kleeman & Osborne, 2002). There are operational risks associated with CAA that have security implications, such as the server crashing, and these risks need to be identified and procedures established to minimise them (Zakrzewski & Steven, 2003).

There are software standards for security, for example, the British Standards on Information Security Management BS7799, which has also been adopted as an International Standard ISO 17799. In addition, when data from the test has been collected, institutions within the UK should abide by the Data Protection Act 1998 (Mason, 2003). If security measures are in place there is no evidence to suggest that the integrity of the examination is more compromised by delivery over the Internet than by paper. However, increasing the security procedures may make it more complex for students to access the application and have an adverse affect on usability and accessibility.

2.15 Test Design and CAA

The instructor would have a preferred assessment technique but in some instances the technology dictates the technique with respect to test design, as this is governed by the question styles available. Therefore, what the instructor wants may not necessarily be what they get. Figure 10 is the author's own summary of the relationship between the user (the student) and the system.

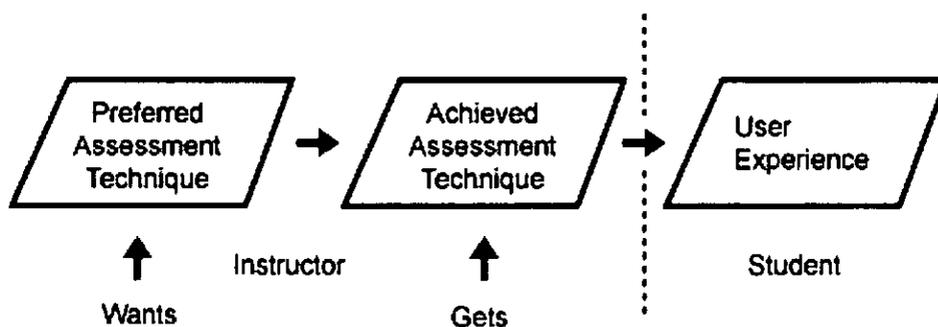


Figure 10 Relationship between technology and pedagogy

For example WebCT® only offers a limited range of question styles compared to dedicated systems such as Questionmark®, therefore, the experience of the test taker is dictated by the application. However, in bespoke systems the preferred assessment technique has driven the technology. This is evident in (Davies, 2002) who wished to address the issues of guessing within MCQ tests and devised his own system. In either of these approaches, poorly developed software or test design may have a negative impact for the test taker.

Figure 10 is further expanded upon in Figure 11 by showing how usability encapsulates the areas.

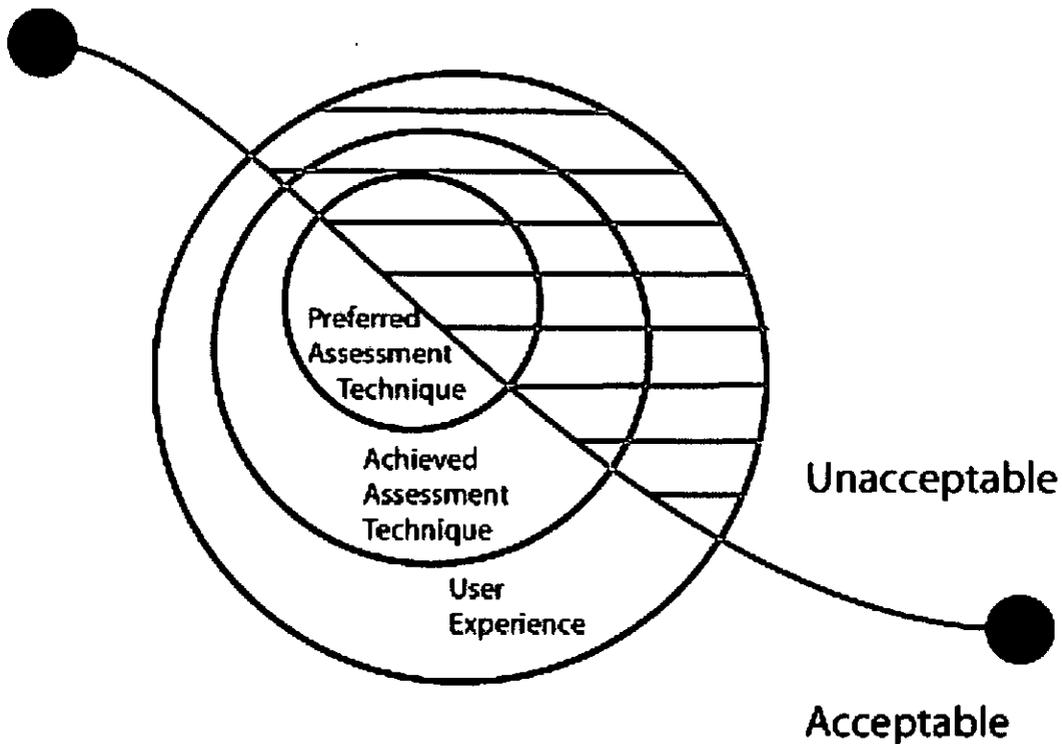


Figure 11 Relationship between user experience, technology and assessment

It is possible to have assessment without technology and vice versa. In figure 11, the preferred assessment technique is the assessment strategy the tutor wants, whilst achieved assessment technique is the assessment integrated into a CAA application. For example, the instructor may wish to embed video into the test and ask a series of questions relating to the video this would be preferred assessment. However, this may not be feasible, the technology presents a barrier therefore, the assessment may need to be modified and this would be the achieved assessment. The user

experience is then influenced by the achieved assessment technique. Usability evaluations can be performed on either the technology or the educational technology. The assessment can have an influence on usability, for example, poorly written questions will adversely affect user satisfaction and could lead to unacceptable consequences. Therefore, within CAA it is important to understand how the various interactions between assessment and technology can affect the user. The research in this thesis is concerned with identifying the problems that would lead to unacceptable consequences represented by the shaded area in Figure 11. Unacceptable consequence within this thesis is defined as a problem that may affect test performance. There may be cases whereby problems do not have any adverse consequences and just leave the user dissatisfied with the application or experience, for example, if the user did not like the colour of the interface, this would be judged as acceptable. This definition will form the basis of the coding of usability problems discussed in Chapter 4.

2.16 Potential Unacceptable Consequences

As stated in Chapter 1, the objective of the thesis is *To determine whether severe usability problems exist that can cause users difficulties and dissatisfaction with unacceptable consequences whilst using existing commercial CAA software applications*, through analysis of the CAA literature there are a number of situations that may arise that would lead to unacceptable consequences.

- **Situation:** Server Crashing (Zakrzewski & Steven, 2003)
- **Consequence:** Loss of exams or marks
- **Situation:** Computer crashes (Ricketts & Zakrzewski, 2004)
- **Consequence:** The students have to complete the exam again if the answers have not been saved or login again continuing from the point where the crash occurred and complete the remaining part of the exam on paper. This could potentially increase stress for students, it may disrupt other students and cause time delays in completing the exam.

In addition to these a number of potential unacceptable consequences have been identified by analysing the user tasks depicted by the activity diagram Figure 8.

Chapter 2

These have not been cited in the literature but were judged to be plausible situations arising from the use of CAA.

- **Situation:** Cannot access the system
- **Consequence:** Exams may have to be rescheduled for another day or a paper based version may have to be completed leading to comparability issues between versions. There could also be an increase in stress which may affect overall performance.
- **Situation:** Exam not active
- **Consequence:** Delays for the students in starting the exam which may disrupt their concentration or increase their stress.
- **Situation:** Not able to change answer once saved
- **Consequence:** Question marked wrong, depending on the scoring algorithm applied, this could lower the students overall mark. The student could fail the test or be classified to a lower grade.
- **Situation:** Student accidentally forgets to save the answer
- **Consequence:** Questions are not marked or they have to re-answer the question once they realise it has not been saved. They could lose marks for the questions which again may affect the grade. In answering the question again they may not have sufficient time to complete the other questions which would affect their grade.
- **Situation:** Student exits without any validation
- **Consequence:** Some of the questions may not have been answered. They could lose marks for any unanswered questions thus affecting their grade.
- **Situation:** Student accidentally exits the Exam
- **Consequence:** They could lose marks for any unanswered questions thus affecting their grade.

It is unclear at this stage whether the situations identified above would arise and whether they are usability related. Further research would need to be performed to establish this.

2.17 Conclusions

This chapter has highlighted some of the complexities of using CAA and has discussed some of the different research perspectives from security to question styles. It is evident that user satisfaction of CAA could be hindered by factors associated with test design as well as the technology.

The knowledge gained in this literature review of CAA has been used to refine the research and help the design of the studies in Chapters 5 to 10.

It is evident that there is a close link between the assessment and technology with respect to CAA. It is unclear as to whether it is feasible to evaluate the technology in isolation of the assessment as they are dependant on one another. Usability evaluation is discussed in the next chapter.

2.17.1 Limitations

The discussion in this chapter has mainly focussed on issues and challenges in CAA within the context of objective testing. Section 2.5 offered a discussion about question styles and examples were discussed from a number of domains, other challenges and issues may arise in domains not covered in this review, such as CAA integration in the arts and humanities.

The work in Section 2.9 on institutional strategies is based on UK experiences; it may be that in different countries the adoption of CAA has been approached differently.

2.17.2 Contributions

The literature review has demonstrated that there is a diverse array of technology available for CAA as represented by Figure 5. There has been considerable research conducted into the integration of CAA within institutions predominately focusing on the software from a pedagogical and technological perspective such as; question design, time saving benefits, instant feedback and the infrastructure required to deliver the assessments. It is evident from Section 2.16 that in using CAA unacceptable consequences may arise for the end user and these may be attributed to poor usability in the application. Many of the decisions made in relation to utilising the system may affect the stakeholders in an adverse manner, such as inappropriate use of question styles and security procedures may affect the ease of use.

Chapter 3 Evaluating Usability

3.1 Introduction

In an early definition of usability Shackel (1986) identified four dimensions that are important: effectiveness, learnability, flexibility and attitude. Although these four constructs are still relevant, the more widely adopted definition is ISO 9241-11 which 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). Although this is widely cited, other constructs have emerged to define characteristics of usability. In a study of the usability of E-encyclopaedias, the evaluation aimed to measure the level of interaction and added value to gauge the usability of the applications (Wilson *et al.*, 2004). Applied to CAA, the students tend to place emphasis upon the grade they obtain and, therefore, this construct may be the factor they place greatest value on. This is supported by research described in Chapter 2 which suggested that students are assessment driven, therefore, any factor that affects test results may have a negative impact on perceived usability.

3.1.1 Objectives

The purpose of this chapter is to provide an introduction to evaluation techniques, to examine the different methods for evaluating the usability of applications and determine the methods that are appropriate to CAA.

3.1.2 Scope

There is a vast amount of research that has been published within the Human Computer Interaction and Educational Technology Community over the last two decades. Research into usability dates back to the 1970's and the key challenge was to identify the appropriate literature from both domains. The literature review predominately used digital libraries including the ACM publications, AACE, Ingenta and educational technology journals such as the British Journal of Educational Technology. These provide a significant proportion of the key literature within both domains. For example, the early literature on heuristic evaluations was published by the ACM.

3.2 Usability Evaluation Methods

Hartson *et al.*, (2003) state that a usability evaluation method (UEM) refers to a method used to perform a usability evaluation of an interaction design at any stage of its production. Usability metrics that may be measured include time taken to complete a task, number of errors, time lost to errors, time to recover from errors and the number of users who successfully completed the task. In evaluating these constructs Whitefield *et al.*, (1991) proposed a classification of UEM based on the two resources available during the evaluation process; the users, and the computer see Figure 12. Each of these two resources could be either real or representative. For the computer, real means having a physical computer system, whilst an example of representational could be the use of a paper prototype. The real users are actual users whilst representational can be descriptions of the users (for example personas) or domain experts simulating real users.

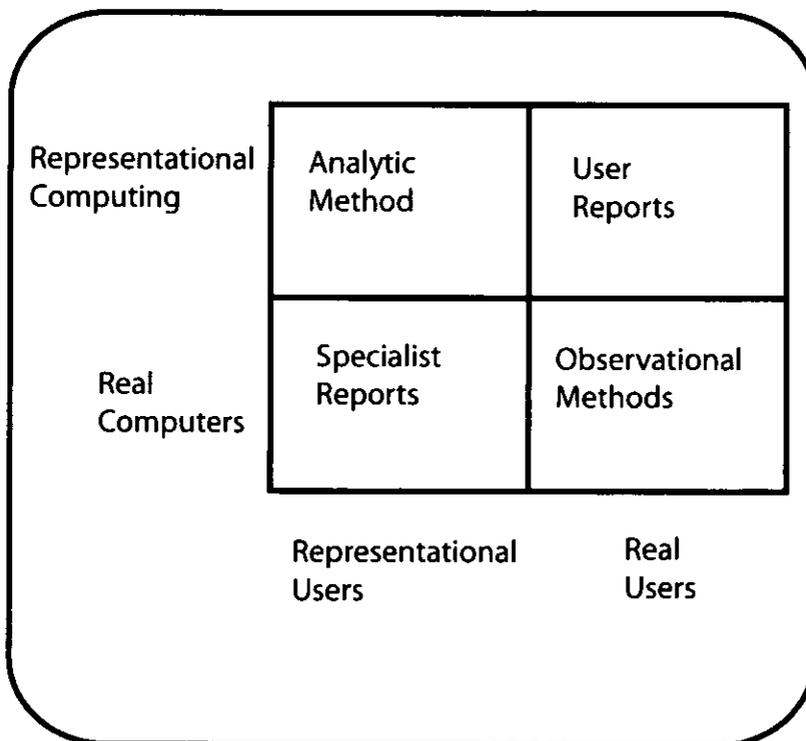


Figure 12 Classes of evaluation methods (Whitefield et al, 1991)

An alternative to the classification displayed in Figure 12 is proposed by Nielsen and Mack (1994) who identified four methods for evaluating user interfaces: automatically through the use of evaluation software; empirically through user studies; formal inspection methods based on models and formulas to identify

problems; and informally based on experts evaluating interfaces based on guidelines. Automated evaluations would not fit into the model proposed by Whitefield *et al.*, (1991) as there is no category for the computer performing the evaluation. The four methods outlined by Nielsen and Mack (1994) are discussed in the next sections.

3.2.1 Automated

Ivory and Hearst (2001) discuss the state of the art of automated evaluation techniques, such as the capture of user activity through the recording of key strokes or the recording of the time events occur. The research showed that automated approaches can work which is further supported by Byrne *et al.*, (1994) and Hibert and Redmiles (2000), this is contradicting Nielsen and Mack (1994) who claim that automatic methods do not work. This approach may only be effective within certain domains and conditions and in certain situations it may not work. For example within CAA capturing key strokes may not be a suitable measure of effectiveness as the length of answers may vary in short answer questions and it may not reveal specific problem. Therefore automated approaches were judged to be unsuitable for answering hypotheses defined in Chapter 1, as it would not reveal the usability problem and could not measure user satisfaction.

3.2.2 Empirical – User Studies

Empirical usability research requires user participation and uses a number of different approaches and styles. User testing is widely recognised in the field of HCI as the most reliable method to achieve usability in a software system (Woolrych & Cockton, 2001) however, issues can still arise such as inappropriate analysis which undermines the validity of the results (Cairns, 2007). There are two distinct evaluation styles, those performed under laboratory conditions and those in the users actual working environment (Dix *et al.*, 2004).

Evaluation techniques under laboratory conditions may include: observations, error logging, eye gaze and screen capture. However, it may be unethical to have students participating in a summative CAA exam within the confines of a laboratory if the process interfered with their concentration or they were uncomfortable in the unnatural surroundings. For example, if observations were used within a traditional computer laboratory then students may find the process distracting having someone

making notes as they progress through the exam. To overcome this screen capture software could be used (this may interfere with the CAA application) or the session could be videoed. These methods would present practical concerns, such as maintaining the integrity of the exam, as the usability laboratory within the university can only accommodate one student at a time, there would be problems obtaining a large enough sample as the exam would have to be scheduled over a number of sittings. Empirical methods within a laboratory were judged to be impractical and potentially unethical to perform, therefore these methods were deemed inappropriate. Within the context of CAA one of the easiest methods to adopt, alleviating both ethical and practical concerns, would be a post-test survey, this approach has been used successfully within this domain (Patterson & Bellaby, 2001; Ricketts & Wilks, 2002).

Surveys have been used successfully to evaluate the usability of applications (Frokjaer *et al.*, 2000; Van Veenendaal, 1998) but one of the key concerns in using survey methods is ensuring the reliability of the scale (Sapsford, 1999), particularly if using Likert Scale questions. Due to the successful use of survey methods within both the HCI and CAA domain, this method was deemed to be a viable option for answering hypotheses defined in Chapter 1. It is anticipated that surveys could be successfully used to measure student satisfaction post test and students could report any problems they encountered through open ended questions. The implications of using a survey based approach are discussed in the next chapter.

3.2.3 Formal Inspections

Formal methods are based around models of user interaction with systems such as GOMS and keystroke per character. It is suggested that formal methods are difficult to apply and do not scale up to complex interfaces (Nielsen & Mack, 1994). However, there have been a number of usability studies using formal methods such as GOMS which is an abbreviation of the components of the model Goals, Operators, Methods and Selection Rules (Card *et al.*, 1993; John & Kieras, 1996). It is concerned with the cognitive processes required to achieve a goal. These methods are normally implemented as part of the software development life cycle (Gunn, 1995) and do not appear to be as widely used on existing systems compared to other methods. If these methods were adopted they would not reveal user satisfaction

which was one of the hypotheses in Chapter 1 and therefore formal inspection methods were dismissed.

3.2.4 Inspection Methods

Informal methods rely on the judgement of the evaluator to predict problems and techniques have been developed including:

- Standards Inspection (not discussed below) (Wixon *et al.*, 1994).
- Cognitive Walkthroughs (Wharton *et al.*, 1994)
- Heuristic Evaluations (Nielsen & Molich, 1990)

3.2.4.1 Cognitive Walkthroughs

Cognitive walkthroughs were originally proposed by Polson *et al.*, (1992) and Wharton *et al.*, (1994) and require a detailed review of the sequence of actions the user will perform to complete a known task. The evaluators would then go through each of the steps documenting any likely usability problems. This may be a suitable method for evaluating CAA applications but problems may arise in predicting how users will interact with the system and establishing the sequence of events. A number of CAA applications were discussed in Chapter 2 and these share similar features in that once the user logs on they tend to offer free movement around the system but the interaction may vary depending on the question styles. Therefore, determining the correct sequence of events may be difficult and as a consequence, the evaluation may be ineffective. Using this method may result in some aspects of the system being overlooked which may occur in all inspection methods.

3.2.4.2 Heuristic Evaluations

This technique uses a small number of expert evaluators to examine the interface and judge its compliance to a number of usability principles. The evaluators would independently examine the interface recording any usability problems encountered and then merge their individual lists into an aggregate list of problems. The most widely used and cited heuristics are Nielsen's (Nielsen, 1992; Nielsen & Molich, 1990). There is uncertainty to the suitability of this method for use within CAA as Nielsen's heuristics have come under criticism in recent years for their unsuitability within certain domains, for example E-learning (Evans & Sabry, 2003) and Accessibility (Paddison & Englefield, 2004) resulting in domain specific heuristics

being synthesised. No domain specific heuristics have emerged within the CAA domain. Heuristic evaluations may be appropriate for identifying usability problems that would lead to unacceptable consequences with a CAA application, this is one of the objectives of the thesis, however a limitation of heuristics is it would not be able to measure student satisfaction the other objective. If survey methods were ineffective at identifying specific usability problems then this method may be an alternative to develop a corpus of usability problems.

3.3 Usability and Computer Assisted Assessment

There is limited research surrounding the usability of assessment tools compared to the studies investigating the usability of general educational technology environments (Berg, 2000; Parlangeli *et al.*, 1999; Piguet & Peraya, 2000). These studies used varied usability evaluation techniques such as experimental design, surveys or examining the interfaces within the context of usability heuristics such as Shneiderman's Eight Golden Rules of Interface Design (Shneiderman, 1998) or Nielsen's heuristics (Nielsen & Mack, 1994). However, within educational technology some of the metrics that have been used for evaluating usability, identified in Section 3.2, may not necessarily be applicable. Masemola and De Villiers (2006) highlight this fact in relation to the time taken to complete a task, this is one measure of usability that could be applied, but the time spent interactively learning is not a suitable usability measure for e-learning as a user may go over the content several times and not experience any difficulties with the application.

The same may be true within CAA, as task completion time may not be an appropriate metric for evaluation. A diligent student who carefully reads the questions before answering and then reviews the questions at the end may take considerably more time to complete, but this is not an indication of difficulties or complexity of the task, it indicates they have been thorough in completing the test.

The British Standard 7988-11 for the use of information technology in the delivery of assessment offers some guidelines on usability (BS7988, 2002), however, this has since been superseded by an International Standard (ISO/IEC23988, 2007). The recommendations can be difficult to interpret, with statements such as the *navigation should be simple and clear*. For example, a CAA exam consisting of 100 questions may be more difficult to navigate and cause disorientation compared to an

exam with only 20 questions. Futurelab commissioned a literature review relating to e-assessment and only a small section concerned usability (Ridgway & McCusker, 2004), this simply stated *people using an assessment system – notably students and teachers – need to understand and be sympathetic to its purpose*. A JISC project looking at advanced e-assessment techniques (Ripley *et al.*, 2009), usability was not a factor that was considered in the analysis of the applications. Fulcher (2003) analysed interface design guidelines applied to CAA within the context of a system lifecycle development. These guidelines are still ambiguous and lead to usability problems. For example, *Each page should have a clear title at the top of the page that relates to a map of the test*, indicating that users should know exactly where they are within the test. Both Questionmark® and TRIADS® comply with this guideline but present the information in different formats. For example, Questionmark® displays your location as 6/20 indicating that you are on question 6 of 20, whilst TRIADS® would display 30% of test complete which would require more cognitive resources to interpret. Following these guidelines would offer some assistance in developing usable CAA systems but may not prevent usability problems arising if interpreted incorrectly. Although these guidelines are to assist in the development of CAA applications there is no evidence to suggest they have been utilised to develop a usable CAA application.

In the development of a bespoke CAA application Lilley *et al.*, (2004) claim to have evaluated the application using Nielsen's heuristics (Nielsen, 1994b) however, they did not follow the traditional methodology. No list of usability problems were produced with associated severity ratings. Here they used 11 evaluators who independently assessed different elements of the prototype. They rated the interface based on a Likert Scale for each of the 10 usability guidelines in Nielsen's heuristic set. The results reported scores between 3.9 and 4.5 so they presumed there were no major usability problems, however, the results would have been compromised because of the evaluator effect and the fact they were examining different interface components.

There have been a number of studies looking at students attitudes towards CAA (O'Hare, 2001; Ricketts & Wilks, 2002) which both used survey instruments to gather data. Although they concluded that students were satisfied with the assessment process, these surveys only examined students within a limited number

of subject domains and, therefore, the research cannot be confidently generalised. The research in this thesis will thus use survey methods to establish if students are satisfied with CAA in the computing domain expanding the existing knowledge.

3.3.1 Potential Task Based Usability Problems

As identified in Section 2.8.1, the first task starting the exam is usually achieved through a password function and this is essential for authenticating users (Bonham *et al.*, 2000). There is usually a trade off between security features and usability (Besnard & Arief, 2004). Increased security procedures may result in some students not gaining access to the exam and having to complete a paper based version. This has occurred in the author's institution. This has implications for the validity of the assessment as there is empirical research suggesting that the same test taken on paper and computer are not comparable (Noyes *et al.*, (2004; Pommerich, 2004). This could lead to an additional unacceptable consequence:

- **Situation:** Not all students doing the exam using the same delivery method, some may be completing the exam on paper (Noyes *et al.*, 2004).
- **Consequence:** Exam results between different forms are not comparable

Once in the test environment the students need to interact with the interface in order to answer the questions, therefore, the interface has a crucial role in facilitating the achievement of their goals. There are numerous variables that could have an influence on students' achieving their goals. For example, Clariana and Wallace (2002) identified the way in which the text is displayed on the screen and whether the questions require scrolling as variables that could affect test performance. There is evidence to suggest that the Arial font is perceived to be easier to read than the Times font and that varying the font size effects legibility (Bernard *et al.*, 2003). Within most systems the instructions and questions are text based, therefore, inappropriate fonts and poor colour contrast could affect legibility and thus affect usability. Within summative assessment, ambiguous questions can also have a negative impact on user satisfaction and there is also concern about being penalised for spelling mistakes (Sim *et al.*, 2004). The consequences may be:

- **Situation:** Poor on screen legibility of the questions
- **Consequence:** Loss of time through poor legibility

- **Consequence:** Misunderstanding the question and, therefore, answering it incorrectly

When navigating between pages, students should only be devoting a limited amount of mental resources to navigational activities. Within the test environment the users should be able to determine: where they are, what they can do, where they are going and know where they have been (Dix *et al.*, 2004). Navigational problems are a concern, particularly under time constrained conditions, as this may affect students' ability to answer all the questions. For example, students need to be able to identify which questions they have answered and be able to easily return to previous questions (Sim & Holifield, 2004b). If the navigational structure of the test is unclear, requires undue attention, or leads to test-taker errors, then arguably the scores will be meaningless or wrong (Fulcher, 2003).

- **Situation:** Poor navigation (Fulcher, 2003)
- **Consequence:** Might miss answering a number of questions due to confusion in the navigation.

Within the summative context, finishing the exam should only be allowed in two circumstances, once the time limit has expired or once a student has decided he or she has completed the exam. It is essential that students can not accidentally exit the test as this would be a critical usability problem potentially leading to lost results. Secure browsers have been developed to overcome this issue, such as Questionmark Secure (Kleeman & Osborne, 2002), however, many of the applications are web based and the user could simply close down the browser exiting the test and being unable to re-start or losing data.

3.3.2 Evaluating Usability in CAA

In section 3.2 some of the main UEMs were discussed and the feasibility of using these methods was analysed. From analysing the goals of the user in section 3.3.1 it is apparent that there are numerous variables that could have an impact on usability whilst the user is accessing the application, navigating, answering the questions or exiting the text. It may be possible to analyse certain variables within a CAA application such as the colour, fonts and navigational structures to determine the effect they have on usability. Many of the software applications identified in Section

2.5 allow a certain level of customisation and it is possible to change some of the attributes including colours and fonts. However, for ethical and practical reasons, empirical research looking at single variables such as fonts and colours, could be difficult to design and a real assessment could not be used, this is discussed further in Chapter 4. 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. The evaluations of the CAA applications in this thesis will use survey methods to answer hypotheses in Chapter 1 and will focus on the usability problems associated with the user's goal. O'Hare (2001) and Ricketts and Wilks (2002) have successfully evaluated user satisfaction of CAA applications through the use of survey methods therefore this seemed an appropriate method.

Within the Blueprint for Computer-Assisted Assessment (Bull & McKenna, 2001) there is a questionnaire that has been used at Luton University and other institutions for evaluating CAA, this will be used adapted in order to answer both hypotheses reported in Chapter 1:

- Usability problems exist which could have an impact on students test results thus leading to unacceptable consequences.
- Student's are satisfied with commercial CAA applications.

The rationale and application of this method will be discussed in greater detail in Chapter 4.

3.4 Conclusions

This chapter has provided an outline of some of the main UEMs used within both educational technology and HCI. It has shown that although considerable research exists in the area of usability evaluation there is limited research in the CAA domain. Fulcher (2003) provides useful guidelines and usability concerns in relation to CAA but does not provide evidence of their existence within applications. These guidelines are geared towards software developers and it is, therefore, possible that some of these guidelines may have been overlooked in commercial applications. In addition, academics customising the interface may not be familiar with this research. This lack of research relating to usability and CAA gives support for the fact that

major usability problems may exist in CAA applications and this is explored in the thesis.

The knowledge gained in this literature review on UEM, has been used to refine the research methodologies used in studies in Chapters 5 to 9. It was identified that some of the constructs that are traditionally measured within usability studies may not be appropriate within CAA. Furthermore students have the most to lose through poor usability because it could affect their overall grades. The thesis will examine CAA and usability from the students' perspective.

Section 2.15 showed there is a close link between the assessment practices and technology, this has further been expanded to incorporate usability. It may not be feasible to evaluate the technology in isolation of the assessment as there is a dependency on the questions being produced in order for a student to participate in a test. In Section 3.3.1, a number of potential unacceptable consequences were identified, some expanding and reconfirming those identified in the previous chapter. Based on the literature review, the methods used in the body of the thesis are discussed in the next chapter.

3.4.1 Limitations

The literature review in this chapter relating to usability has been scoped as follows in an attempt to ensure the discussion was appropriate to the CAA domain. Section 3.2 discussed UEMs, presenting an overview of some of the widely researched methods. Other methods exist including Wizard of Oz and Eye Gaze which have not been reviewed on the assumption that these would not inform the thesis and provide the necessary data to answer the hypothesis in Chapter 1. For example the data gathered from a series of eye gaze experiments would be vast, timely to process and may not necessary reveal a significant number of problems or reveal any indication of user satisfaction.

In Section 3.3 the literature surrounding the usability of CAA was discussed and analysed. From the review it is clear that there has been a limited amount of research published within this area. However, it may be possible that there is research in articles related to CAA that are published in domain specific journals such as accounting or medicine.

Key word searching was problematic and time consuming, for example conducting a search using 'computer assessment' would produce a divergent array of articles in excess of one thousand using major indices like Science Direct. Browsing through entire journal contents and papers references were methods used to help alleviate these problems.

3.4.2 Contributions

The literature review looked at UEMs within the context of CAA. Although many methods exist, not all are feasible within the domain under investigation. For practical and ethical reasons the decision was made to use survey methods to elicit usability problems from users of CAA applications and gauge their satisfaction.

In Section 3.3 the users' goals were established and a number of potential variables that may hinder usability within a CAA environment were established.

Chapter 4 Methodology and Research Design

4.1 Introduction

In the previous chapter, the literature revealed a diverse number of UEMs that had been applied to usability testing in the context of educational technology research. This chapter explains the rationale for the research design adopted in this investigation, building on the discussion in Chapter 3 to discuss more general research methods. An analysis of the various methodologies used within the context of HCI and educational technology justifies both the appropriateness of the methodology and its limitations.

4.1.1 Objectives

The purpose of this chapter is to provide an overview of the research methods identified in Chapter 3, describing how these methods have been adopted in this research, discussing their suitability within this domain and their limitations. It also looks at issues relating to ethics, validity and reliability.

4.1.2 Structure

Section 4.2 begins with reiterating the research objectives and Section 4.3 outlines some of the common research methods used within HCI and Educational Technology. The design of the research in this thesis is outlined in Section 4.4 focusing on a mixed method approach incorporating survey tools, followed by sections on participants and ethics, Section 4.5 and 4.6. The ethical implications of experimental design within CAA are discussed, with the emphasis on the potential to affect the test results of the participants.

4.1.3 Contributions

The main contribution is:

1. An outline of the research method used within this thesis

4.2 Selecting Research Methods

As stated in Section 1.1 the objective of the research is “*To determine whether severe usability problems exist that can cause users difficulties and dissatisfaction with unacceptable consequences whilst using existing commercial CAA software applications*”. The research was conducted in parallel with two projects see Figure 13.

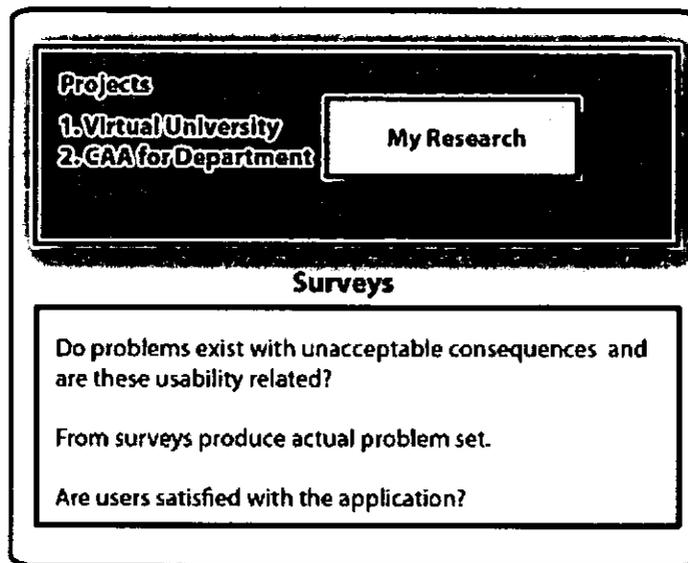


Figure 13 Overview of research structure

The intention is to determine if severe usability problems exist in CAA applications that may hinder test performance, resulting in unacceptable consequences for the user and to construct a list of these problems. To meet the aims of the other projects, for example, establish if CAA is an acceptable assessment strategy for computing students, some of the questions used in the surveys were not directly relevant to the objectives of the work in this thesis and were, therefore, omitted from the analysis reported here.

Another objective was to further the knowledge relating to evaluation of CAA applications by analysing the effectiveness of the methods applied.

In Section 3.2 the various methods for UEM were identified and discussed within the context of CAA. Some of these methods of evaluation have been widely used across a number of diverse subject areas such as psychology and education (Breakwell *et al.*, 2000; Cohen *et al.*, 2001). Survey methods appeared a viable technique for answering the hypotheses reported in Chapter 1.

4.3 Research Methods

The research in this thesis is embedded within two domains: educational technology and HCI. Research in educational technology usually falls within the field of social sciences, whilst HCI is predominately within computer science, but has emerged from a variety of domains including linguistics, psychology and mathematics. A common factor between both the domains is the reliance on an appropriately designed research method. Cresswell (2003) proposed a framework for research design based on three strategies for enquiry:

- Quantitative – experimental design, non experimental designs
- Qualitative – Narratives, Ethnographies
- Mixed Methods – Sequential, Concurrent

The mixed method approach will be further discussed in Section 4.3.7. An alternative to this approach is discussed by Breakwell *et al.* (2000) who suggest that research can differ along a series of four dimensions:

- Type of data elicited
- Techniques of data elicitation
- Types of design for monitoring change
- Treatment of the data as qualitative or quantitative

For the purpose of this thesis the research design strategies for these four dimensions will be examined within the context of usability and CAA.

4.3.1 Types of Data

Types of data can in the context of psychology, refer to phenomena such as emotion, communication patterns or institutional hierarchies. Within this research the phenomena that needs to be elicited relates to the usability of CAA applications. This could relate to user satisfaction and/or problems within the applications.

4.3.2 Techniques of Data Elicitation

Techniques for data elicitation also tend to fall into two categories: direct or indirect (Breakwell *et al.*, 2000). The target within the context of usability research may

differ, depending on the focus of the research it may be the end user or the system under evaluation, therefore, different techniques may be required.

4.3.2.1 Direct

Methods of direct elicitation include self reporting methods such as interviews or questionnaires, along with self-revelation methods through behaviour including role play and performance in tasks. The self reporting method relies on the participants recording the data. In self-revelation the researcher may be primarily responsible for recording the information. Many of these direct methods have been utilised in research studies in both Educational Technology and HCI. Van Veenendaal (1998) used questionnaires to measure usability attributes, interviews were used to determine staff views regarding the introduction of CAA (Hodson *et al.*, 2002) and empirical investigations often analyse users' performance in completing tasks (Read *et al.*, 2001).

4.3.2.2 Indirect

Indirect methods tend to rely on the researcher observing behaviour or examining archival records. Again these methods have been adopted in educational and HCI research. Sim *et al.* (2006) observed children interacting with three educational applications, recording signs of engagement and any usability problems. In this study counter-balancing techniques were used to minimise any learning effect, the participants may have gained skills in one application that carry over to the next thus biasing the results. To overcome this problem a Latin-square approach can reduce this effect. For example, if two products are being evaluated (A and B) by 20 users, the users would be split into two groups of 10 (G1 and G2), G1 would evaluate the product in the order of AB whilst the order for G2 would be BA. This counter-balancing technique is predominately used in within-subject design experiments. The opposite of this approach is the between-subject design, relating this to the example above, G1 would examine A and G2 would examine B; then the results would be analysed.

Analysis of archival records or secondary research can lead to the synthesis of new knowledge. For example Squires and Preece (1999) used established educational theory and Nielsen's heuristics (Nielsen, 1994a) to devise a set of heuristics for

evaluating educational technology and Paddison and Englefield (2003) used accessibility guidelines to inform the development of a set of heuristics.

4.3.3 Types of Design for Monitoring Change

In psychology and HCI, one of the key aspects of research design is the ability to measure change. This may be done by examining users' attitudes to an incremental software development or improvements in performance by using two different input methods. It is suggested that change can be measured in three different classes (Breakwell *et al.*, 2000):

- Longitudinal
- Cross-Sectional
- Sequential

4.3.3.1 Longitudinal

Longitudinal studies involve collecting data from the same sample on at least two different occasions. There is no specification on the interval required between data collection points, it may be a few days or months. Longitudinal studies have been used in educational research to establish the impact technology has had on a cohort of students (Giza & Awalt, 2005). By definition, if a repeated measure design is used, as in Sim *et al.* (2006) then it may be deemed a longitudinal study as the criterion has been fulfilled.

4.3.3.2 Cross-Sectional

Cross-sectional studies involve eliciting information at a single point in time from participants in different situations. Cohen *et al.* (2001) suggest that it produces a snapshot of a population at a given point in time, for example the national census. Bryman (2004) identifies four key elements to cross-sectional research:

- More than one case
- At a single point of time
- Quantitative or quantifiable data
- Patterns of association

From the literature review, conducted in Chapter 3, there is a lack of evidence to suggest that this method has been widely adopted in HCI evaluations, although it has been used in educational technology research for collecting data about a particular cohort (Thompson & Radigan, 2002).

4.3.3.3 Sequential

The sequential studies method will choose samples from a particular condition, this may be age, and then study them at different intervals. This concept has emerged from industrial quality control for inspecting the quality of products (Sapsford, 1999). A small sample from each batch would be tested, if none were faulty the batch would pass, however, if more than the predetermined acceptable level of rejects were found, the batch would be scrapped or further samples would be taken.

4.3.4 Treatment of the Data

Once the data has been elicited two main data types will emerge: qualitative or quantitative data. There are different approaches to usability evaluations that yield different types of data in both quantitative and qualitative formats. The treatment of the data to yield conclusions will vary depending on the data type.

4.3.4.1 Qualitative Data

There is a notion that qualitative data can be construed to emphasise words rather than quantifications in the analysis of data. Bryman (2004) and Cresswell (2003) both suggest that qualitative research is exploratory by nature and is used to explore phenomenon when the variables and theory base are unknown. This definition addresses the qualitative research method and does not focus on just the data type, but the output of the research. The data gathered can be analysed using various methods from data display where relationships are coded and classified to certain criteria and from this process conclusions can be drawn.

4.3.4.2 Quantitative Data

Quantitative data may be in numerical format and is used to test a theory or hypothesis relating to variables which influence the outcome (Cresswell, 2003). The data gathered can usually be classified into one of four data types:

- Interval/Ratio – distance between the categories identical across the range
- Ordinal - categories can be ranked in order
- Nominal – categories cannot be ranked in order
- Dichotomous – data can only have two categories

Depending on the type of data statistical analysis can be performed to test the hypothesis or infer causality. It is important to understand the data type to prevent the wrong analysis being performed as this would potentially invalidate any conclusions presented. It is possible that new theories will emerge from the analysis of the data and this approach is taken in grounded theory research design.

4.3.5 Reliability

Reliability refers to the data collection tool being consistent in its recording of the phenomena and the measure being insensitive to change (Sapsford, 1999). There are different meanings to the term as discussed by (Bryman, 2004):

- Stability – whether the measure is stable over time
- Internal reliability – are the indicators that make up the scale consistent
- Inter observer consistency – where more than one observer is involved there is a possibility that there is a lack of consistency in the decisions

4.3.6 Validity

Validity, means the accuracy with which a set of scores actually measures what it ought to measure (Ebel, 1972; Salvia & Ysseldyke, 1991; Wood, 1960). Several different types of validity have been identified: curricular; construct; predictive; (McAlpine, 2002) logical; content; (Wood, 1960) convergent and discriminant; incremental; face; interpretive; (Walsh & Betz, 1985) and criterion (Moskal & Leydens, 2000). It is not possible to say a tool is valid because it is a continual process of measurement to predict its degree of validity. To overcome the problem of statistical analysis with a small sample size in some subjects, criterion-related validity is the most viable option (Rafilson, 1991).

Chapter 4

Black (1999) identified 15 basic sources of invalidity that are found within educational and social science research which threaten construct, internal, external or statistical validity.

Construct validity refers to the experimental demonstration that the test is measuring the construct it claims to be measuring. For example, if an online assessment required a high level of I.T. skills it would be inappropriate if you were testing the student's ability at maths. Fulcher (2003) argues that poorly designed interfaces within CAA applications may be a threat to construct validity.

Internal validity is concerned with whether it is possible to correctly infer a causal relationship between two or more variables. Threats to the internal validity include experimental procedures, treatment of participants or the researcher making incorrect inferences from the data.

External validity refers to whether the results of the study can be generalized beyond the specific research context. The threats usually occur when the results are generalised beyond the groups analysed. Similar to external validity Bryman (2004) suggests another source of invalidity 'Ecological' which deals with whether social scientific findings are applicable to 'people's everyday, natural social settings. In the context of this research it would be difficult to simulate a summative assessment environment and predict how the students would behave, therefore, the data was captured during their examinations prior to leaving the exam.

Statistical validity can be defined as the degree to which an observed result, such as a correlation between 2 measurements, can be relied upon and not attributed to random error in sampling and measurement. The threat usually arises from the researcher being selective in the data analysis, unreliable measures or incorrect statistical analysis, for example, performing parametric tests on non-parametric data.

Gray and Salzman (1998) reviewed the validity of five studies which compare UEM and identified that these studies were invalid based on one of the following five forms of validity:

- Statistical conclusion validity – was the change to the dependant variable caused by manipulation of the independent variable?
- Internal Validity – was the change caused by a unknown confounding variable?

- Construct Validity – does the research study measure the construct it claims?
- External – are the results generalisable?
- Conclusion Validity – are the claims supported by the data presented?

4.3.7 Mixed Methods

The reliance on a single method may lead to bias. The mixed method approach utilises various methods triangulating data types to draw conclusions. Cresswell (2003) identifies a number of different decisions that need to be addressed within this paradigm:

- What is the implementation sequence of both the quantitative and qualitative data collection?
- What priority will be given to the data collection and analysis?
- At what stage will the data be integrated?
- Will an overall theoretical perspective be used in the study?

By combining the data sets, a better understanding of the problem can be formulated than if either datasets had been used alone (Cresswell, 2007). In this research multiple studies will be performed, the results presented and the results will be integrated. There are many methods for merging the data sets from various studies, for example, the convergence model where the data is analysed separately and merged at the end or multilevel models where different methods are used for different parts of the system and the findings are then merged at the end (Cresswell, 2007).

An overview of the research methods used in the experiments within this thesis is discussed in the next sections with an analysis of their limitations.

4.4 Research Design

This section outlines the research design used in this thesis, discussing the purpose of the study, how it was constrained, ethics and any measures taken to improve the validity and reliability of the findings.

4.4.1 Purpose

The objective of the research was *to determine whether severe usability problems exist that can cause users difficulties and dissatisfaction with unacceptable consequences whilst using existing commercial CAA software applications*. The work focuses on usability from the perspective of the students, therefore, members of staff views were elicited only for the purpose of aggregating the data sets and evaluating the potential consequences of any problem found. The research design is outlined in Figure 2, Chapter 1. Chapters 2 and 3 have identified a number of potential unacceptable consequences. The first stage of this research is to determine if they are real and usability related. If it is found that they are not usability related, then the nature and direction of this research would be refined at this stage.

In Section 3.2 some of the most widely applied UEMs are discussed and the feasibility of these are analysed within the context of this research. Automated and formal methods are excluded, leaving inspection and empirical methods. These methods would fall under the methods for eliciting data discussed in Section 4.3.2. Initially direct methods would be used in the form of self completing questionnaires to avoid any ethical and practical concerns that may arise in using indirect methods, such as observations which may be too intrusive. Through the use of a carefully constructed questionnaire it would be feasible to answer the hypotheses reported in Chapter 1.

4.4.2 Survey Design

The purpose of the first study is to identify whether severe usability problems exist within a single CAA application that would lead to unacceptable consequences. A survey methodology will be adopted based on self-completion questionnaires distributed to the students following CAA tests. This method will be used for data collection in Chapters 5 and 6 of this thesis. Within the literature on CAA and Usability, questionnaires are often used to gather data from participants (Eckersley, 2004). Attitudes can be measured by presenting a list of declarative statements and asking participants to rate them in terms of agreement or disagreement (Black, 1999). The survey tool will be designed to gather both quantitative and qualitative data about their experience of the CAA software. The quantitative data is mainly

gathered to meet the requirements of a CAA project running in conjunction with this research but to also answer the second hypothesis reported in Chapter 1:

- Student's are satisfied with commercial CAA applications.

Bryman (2004) suggests three guidelines for questionnaire design that researchers tend to follow:

- have fewer open questions, as closed ones tend to be easier to answer
- have easy to follow design to minimize the risk that the respondents will fail to follow filter questions or omit a question
- be short to reduce the risk of respondent fatigue,

The issue relating to fatigue is an important consideration as the students will be completing the questionnaire after an exam. Specific questionnaires could have been used for example, SUMI which has been used in usability evaluations (Moore *et al.*, 2001; Van Veenendaal, 1998) but due to the length of the questionnaire, 50 questions, it was considered too long for students to complete following an exam. The response rate may suffer if too many questions are presented or the majority of questions are open ended.

In devising the attitudinal questionnaire for use in Chapter 5, a selection of questions will be used from research conducted at the CAA Centre in Luton (Bull & McKenna, 2001) which unlike the SUMI questionnaire is shorter in length and focused on CAA.

The purpose of the study in Chapter 6 will be to build on the findings of Chapter 5 by examining another CAA application and extend the problem set. In Chapter 6 the survey tool will be expanded to incorporate additional questions based on the findings from Chapter 5.

4.4.2.1 Analysis of Survey Data

The questionnaires used in Chapters 5 and 6 will use a mixture of dichotomous, five point Likert scale ranging from Strongly Agree to Strongly Disagree and open ended questions. The statements will be coded between 0 and 4 where 0 was strongly disagree and 4 was strongly agree. Parahoo (2006) suggests that the five point scale is the most widely adopted. However, it could have been possible to use a seven point spread, Black (1999) argues that this better represents the range of attitudes or

reviews as people often do not want to be viewed as extremists. However, many examples within the literature still use a five point scale (Cohen *et al.*, 2001; Howitt & Cramer, 2003) and as this questionnaire is adopted from the CAA Centre at Luton who used a 5 point scale, this will be adopted.

The survey tools adopted in Chapters 5 will use Cronbach's alpha to test for internal reliability, as a series of Likert scale questions are to be used to determine students' opinions of interface components and their satisfaction with the assessment technique.

The statements from the open ended questions are to be coded by the author with a unique code for each of the usability problems reported by the participants. For example, in Chapter 5 the problems will be coded first with a W indicating it is WebCT® followed by a U for usability and finally a number:

- WU1 – Problem description

The same coding method is to be applied across all studies thus ensuring that problems can be cross referenced. To try and minimise confusion within a chapter the raw data from a study (unmerged) will be coded first with lowercase letter representing the software and once the data is merged a capital letter will be used, for example a w would be used in the first instance followed by a W. However, the initial coding scheme applied was found not to be generic enough as there were multiple studies using Questionmark® and in different contexts. The basic principle was still applied but additional lettering was used for example in Chapter 8 hF is used where h represents heuristics and F the context of use, Formative.

Figure 8 in Section 2.8.1 identified the user tasks associated with CAA and for the purpose of coding in later chapters, a classification system has been created based on this diagram. This is presented below.

Test (T)
Start Test (S)
 Access (S1)
 Login (S2)
 Select Test (S3)
During Test (D)
 Answer Question (D1)
 Understand how to answer (D1.1)
 Construct an answer (D1.2)
 Confirm answer (D1.3)
 Review / edit question (D2)
 Navigate through questions (D3)
 Feedback (D4)
End (E)
 Awareness of finish (E1)
 Check answers (E2)
 Submit answer (E3)
 Feedback (E4)

Each problem reported in a study will be classified to one of the above codes, for example if it is navigational related it will be coded against D3. This will enable the problems to be grouped based on user task and will help in aggregating the data sets from the various studies. This will be achieved by examining the task code along with the description of the problem, merging if necessary, in order to produce a final corpus. Burns (2000) recommends that in order to avoid bias in the analysis and interpretation, the researcher should engage another researcher to critically question the coding. Therefore, the process of coding the problems to each task code for consistency will be performed by the author and a lecturer in HCI.

In addition to the task code each problem will be given a code to determine whether it would lead to unacceptable consequences. The coding used to establish the consequences of problems are based on the following scale:

- Dissatisfied – the user would be dissatisfied but it is unlikely to affect the overall test performance
- Possible – there is a possibility that the problem may affect the user's test performance
- Probable – it would probably affect the user's test performance
- Certain – It would definitely affect the test performance of the user

To aid the classification of problems this scale was devised to establish the consequences of a reported problem. An alternative to this could have been to use Nielsen's severity rating scale (Nielsen & Mack, 1994), however, it was anticipated that this would be too generic and evaluators would find it difficult to identify the boundaries between the scales.

The coding strategy outlined above did not follow any established methods which have been used within HCI research, such as open coding, axial coding or selective coding (Adams *et al.*, 2008), as the objective was to build a corpus of usability problems and these methods have been predominately used to analyse interviews and questionnaire data. However, the approach uses elements from both axial coding and content analysis (Burns, 2000) in coding the usability problems. Content analysis requires a coding scheme that relates to the research question, therefore, it seemed appropriate to devise a scheme based around unacceptable consequences. The coding method established uses some of the attributes of axial coding, by using the task code this identified the context in which the problem would arise and the consequences were identified by applying the consequences scale.

4.4.2.2 Limitations of Survey Methods

The limitations of surveys are well documented in the literature. Issues include: coding errors which can occur when open ended questions are misinterpreted by the researcher (Breakwell *et al.*, 2000); closed questions that create forced responses, ruling out unexpected responses (Burns, 2000); classification or coding of data to the incorrect data type such as nominal instead of ordinal, resulting in the wrong statistical test being performed (Howitt & Cramer, 2003).

The surveys used in this research will also have to address the known confounds of using a single cohort and a single CAA application this will be examined in Chapter 6.

4.4.3 Merging of Data

One of the key aspects of a mixed method research approach is the triangulation of the data from various studies (Cresswell, 2007). The convergence model will be adopted and Figure 14 shows how problems will be added and discarded throughout the course of this research.

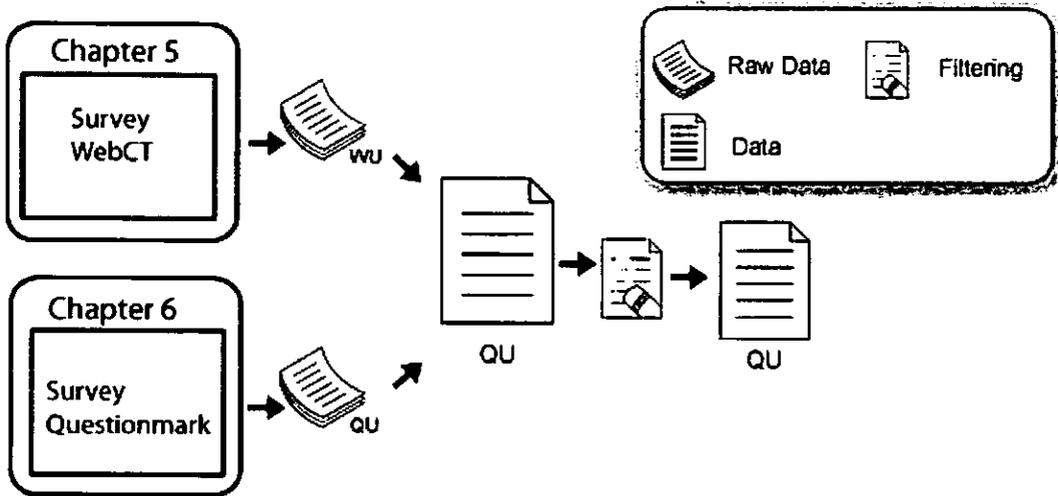


Figure 14 Triangulation of data

The data sets have been coded to ensure that each problem has a unique code and to ensure that it can be traced back to its original study. The survey methods are then merged into an aggregated list of usability problems in the CAA applications.

4.5 Participant Selection

The research was conducted in one higher education institution within the UK. It would be unfeasible to randomly select a number of higher education institutions across the UK because of licensing costs for the software and the potential reluctance of academic staff to adopt this assessment technique. There is considerable literature surrounding CAA that has often focused on a single institution (McLoughlin & Reid 2003; Maclaran & Sangster 2000; Buchan, 2000) however, there have been a number of large scale studies comprising of collaborating institutions (Ashton *et al.*, 2003; Herd & Clark, 2003; Ashton & Bull, 2004). Usability research has also been conducted within single institutions (Woolrych & Cockton, 2000), therefore, this approach appeared viable.

Both staff and students within the Department of Computing, now the School of Computing Engineering and Physical Sciences, were used as subjects for this research. The research was limited to a number of modules within the first year undergraduate and postgraduate programmes. The first year modules used were Programme Design and Implementation and Web Development and incorporated the entire first year cohort. The postgraduate module was Web Development and incorporated the cohort on the MSc in Multimedia.

This research was not trying to differentiate the first year cohort based on demographic data such as age, race or educational background. However, within Computing there is a strong bias in the number of males over females (Carter *et al.*, 2004; HESA, 2002), and this trend is also evident within the School. Unlike other departments such as Humanities, assumptions were made that the students would be I.T. literate and not suffer from computer anxiety as there is evidence that this can affect test performance (Ergun & Namlu, 2004).

The decision was made to use independent subject sampling rather than the same subjects in a longitudinal study over several modules or years. Predominately CAA was being adopted within first year modules within the department and there was scepticism over the suitability of this technique for use on modules which counted towards the students' degree classification. This was due to concern amongst staff of the ability of CAA to test higher cognitive skills. In addition the researcher had no control over the modules that would adopt CAA as part of their assessment strategy within the department. This meant that there would be no guarantee that a longitudinal study could be performed over several years using the same participants therefore an independent sampling method was adopted. One of the objectives of the thesis was to determine whether severe usability problems exist, and independent sampling is widely used in usability studies and therefore was judged to be appropriate.

Another objective was to measure students' satisfaction and there was concern that a longitudinal study would suffer from high attrition rates as a result of retention issues, and even if modules in the 2nd and 3rd year adopted CAA these are specialised to a particular degree unlike the 1st year, therefore the sampling rate would be reduced, for these reasons independent sampling was chosen.

4.6 Ethics

There are numerous statements of ethical practices that research bodies and professional bodies have devised, such as the ACM and British Psychology Society. Diener and Crandall (1978) identify four main areas:

- Whether there is harm to participants;
- Whether there is lack of informed consent;

Chapter 4

- Whether there is an invasion of privacy;
- Whether deception is involved;

Harm, in the context of this research, relates to the effect any of the experiments may have on students' results. Therefore, it would be unethical to design an experiment consisting of two interfaces and hypothesis that one would yield more usability problems than another. It would also be unethical to cause students any additional stress by having tests delivered by a computer than by paper. Within this research there may have been an issue with regards the students having the option to sit a paper based version of the test if they had special educational needs that prevented them using the computer or if they did not have access to the university network.

In relation to informed consent, the completion of the summative exams was compulsory and they did not have the option of doing an alternative paper based exam. However, in the context of traditional paper based tests students do not usually have the option of using a computer. The students had the option to opt out of the research study by not completing the questionnaire. The questionnaire was placed next to the computer at the beginning of the tests and the students were asked if they would complete the questionnaire before leaving. The questionnaires were anonymous and, therefore, it was not possible to establish the identity of the participants. The students were informed of the nature of the research but the amount of detail still remained limited. However, a severe usability problem was established in Chapter 6 and it was decided to inform the next cohort about the problem found in the software prior to the test as it may have effected their results. Also the staff, whose modules were used in this study volunteered to use the various software and consented to the questionnaires being distributed to their students.

Within the research invasion of privacy was not a major issue as no personal details were collected and the methods adopted were not invasive. The anonymity of the students was maintained as no personal details were collected on the questionnaires. The students' test results were not used so issues surrounding confidentiality of the data was also limited.

The students were not deceived in anyway as to the nature of the research and the results. They were informed about the purpose of the study and given the option to opt out.

Ideally once problems have been reported using the survey approach falsification testing should occur to determine whether the problem is real (Woolrych *et al.*, 2004). Falsification testing would require user testing which cannot be associated with genuine summative assessment for 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. Also it would not be possible for the students to take the same exam paper to enable falsification testing to occur and even if using different cohorts there is an expectation that the exam would be altered reflecting any changes in syllabus. Therefore falsification testing was not a viable option.

4.7 Conclusions

This chapter has outlined the research methods used within the body of the thesis along with the limitations of the research. The first hypothesis to be tested is:

- Usability problems exist which could have an impact on students' test results thus leading to unacceptable consequences.

Survey methods will be used to elicit information from students relating to usability and satisfaction of various CAA applications. The coding method that will be utilised for the analysis and classification of problems has been outlined in Section 4.4.2.1 and this will enable the first hypotheses to be tested.

The second hypothesis is:

- Students are satisfied with commercial CAA applications

To answer the second hypothesis a series of Likert questions will be devised based on the Blueprint for Computer-Assisted Assessment (Bull & McKenna, 2001) ensuring that questions are kept to a minimum in an attempt to maximise response rates.

Chapter 5 Usability Pilot Test

5.1 Introduction

The literature review established that there had been considerable work published in both the HCI and educational technology domain regarding usability, however Fulcher (2003) highlighted that there is very little published within the context of CAA. This chapter describes an exploratory study using the assessment tool within WebCT® a commercial Learning Management System (LMS). The objective of the study was to establish whether, usability problems exist that lead to unacceptable consequences within a CAA application, to gauge whether users can identify these problems and measure user satisfaction.

The work in this chapter was published at EDMEDIA and the 7th HCI Educators Conference (Sim & Holifield, 2004a; Sim, Horton *et al.*, 2004). The results also provide a foundation for later work in this thesis.

5.1.1 Objectives

As described in Chapter 1, the main objective is to establish “*If severe usability problems exist that can cause users difficulties and dissatisfaction with unacceptable consequences whilst using existing commercial CAA software applications?*”. This resulted in the two hypotheses being formulated and the objective of this Chapter is to investigate the following:

- Usability problems exist which could have an impact on students’ test results thus leading to unacceptable consequences.
- Students are satisfied with commercial CAA applications.

Other objectives are:

1. *Establish if user can report usability problems within the CAA system using the survey tool.*

An initial questionnaire was used which was adapted from the Blueprint for Computer Assisted Assessment (Bull & McKenna, 2001) and it was anticipated that students would be able successfully complete this and identify usability problems. However if inter-rater consistency is low an alternative approach may be required.

2. *To start building a corpus of usability problems.*

If usability problems are found, then these will form the basis of the corpus of usability problems.

3. *To determine the direction of the research strategy.*

Chapter 4 identified this stage as a stop/go process, that is if no problems are discovered with unacceptable consequences or the problems are not usability related then the research strategy would have to be refined as it would not be worth pursuing.

4. *To identify whether the context of assessment, either formative or summative, affected satisfaction.*

Often the interface within a CAA environment is altered slightly depending on context, usually through increased security procedures and the level of feedback that is provided. Within a summative setting there is the possibility of increased anxiety for the students and this may influence their satisfaction of the system.

5.1.2 Scope

This study was devised to establish if severe usability problems exist in a single LMS (WebCT®) that would lead to unacceptable consequences. A convenience sample was used from a single department at a university, it is acknowledged that the findings may not be generalised to other subject disciplines whose assessment practises differ considerably to computing. The study was also constrained by the difference in sample size between the two groups and the limited number of questions.

5.1.3 Contributions

The main contributions in this chapter are:

1. Within the CAA environment there were a number of potentially severe usability problems found that have unacceptable consequences, identified in Section 5.4 but inter-rater consistency is low.
2. Section 5.3 shows within the limitations of this research study students appeared satisfied with CAA as an assessment method. The study also

showed there to be no difference in satisfaction between the students who used the software for formative or summative assessment.

5.1.4 Structure

The structure of the remainder of this chapter is as follows: Section 5.2 reports the experimental design and the results are presented in Sections 5.3 and 5.4. The conclusions are presented in Section 5.5 with a summary of the findings, the identification of a number of usability problems, a discussion of limitations, and suggestions for further research.

5.2 Study Design

The design was between-subjects single factor with two conditions: Formative and Summative assessment. The summative assessment accounted for 10% of the student's overall grade for the module.

5.2.1 Participants

The sample consisted of 101 undergraduate students on a first year web development module and 23 postgraduate students on a web development module. One staff member was responsible for both modules and agreed to embed CAA into his/her teaching strategy. Therefore it was a convenience sample, of both genders and a diverse age range. Both groups completed the same test, however, the undergraduates undertook a summative test whilst the postgraduates' test was formative. None of the users had any prior experience of using the software for assessment purposes.

5.2.2 Apparatus

All tests were conducted in computer laboratories within the university and all the PCs had the same specifications. This was essential, as differences in equipment, such as monitor resolution, is known to influence test result (Bridgeman, Lennon, & Jackenthal, 2002).

5.2.3 Questionnaire Design

As discussed in Chapter 4, a questionnaire was designed based on evaluation resources within the Blueprint for Computer-Assisted Assessment (Bull &

McKenna, 2001). The same five point Likert scale was used (Strongly Disagree=0, Disagree=1, Neutral=2, Agree=3 and Strongly Agree=4). Only one question was directly used from the questionnaire (The test was easy to use) therefore the reliability of this instrument would need to be established.

Five Likert style questions were designed in total, three relating to the assessment method and two to usability. The statements relating to the assessment were devised to act as an indicator of the students' level of satisfaction with the method. The three statements relating to the assessment method were:

- This type of testing on a regular basis would be beneficial to my studies
- I would find this assessment acceptable as replacement for part of the final exam
- I found this format of assessment less stressful than a paper based exam

There were only two usability statements as this was an exploratory study to establish any issues for further investigation. These statements were:

- The test was easy to use
- The navigation was clear

Although the statements were rather generic and did not focus on specific interface attributes, the purpose at this early stage in the research was to establish if usability problems existed within the software that would have unacceptable consequences and test the survey instrument. Two additional open ended questions were incorporated to enable students to comment on specific issues about the assessment process, these were:

- Did you have any additional problems when using the test?
- Do you have any comments about using this test?

It was anticipated that these two questions would enable students to report any usability problems they may encounter, thus providing the data to establish if usability problems occur with unacceptable consequences. The questionnaire was distributed post-test containing a mixture of Likert, dictotomous and open-ended questions in order to minimise any interference with the test. A Cronbach Alpha reliability test was conducted for each measure, see Table 1.

	Cronbach's Alpha
Perception of CAA 3 questions	0.6037
Usability 2 questions	0.8743

Table 1 Reliability of Questionnaire

The total reliability coefficient for the combined data sets and questions was 0.76.

5.2.4 Exam Design

The majority of questions on the CAA tests were testing the lower cognitive levels (knowledge, comprehension and application) as defined by Bloom's Taxonomy (Bloom, 1956). For example, one of the questions related to the ability of students to recall HTML syntax thus testing their knowledge. The majority of tests conducted by the students were under summative conditions and the questions were internally peer reviewed by subject experts within the department to ensure they were appropriate for the level of study. The researcher had no input into the design of these tests. The test comprised of thirty one questions relating to web development using three question styles: Multiple Choice, Multiple Response and Text Entry.

To access the software both groups followed the same procedure. The students were required to first log onto the university network, using their username and password, then log into WebCT® which required a further username and password (the same as the university account) and finally, once within WebCT®, they needed to navigate to the assessment tool and enter another password to start the exam, this was chosen by the module leader. Questions were presented using single question delivery to minimise scrolling, as this is a factor that may influence test results (Ricketts & S. Wilks, 2002). The main navigation was on the right hand side with all question numbers visible see Figure 15.

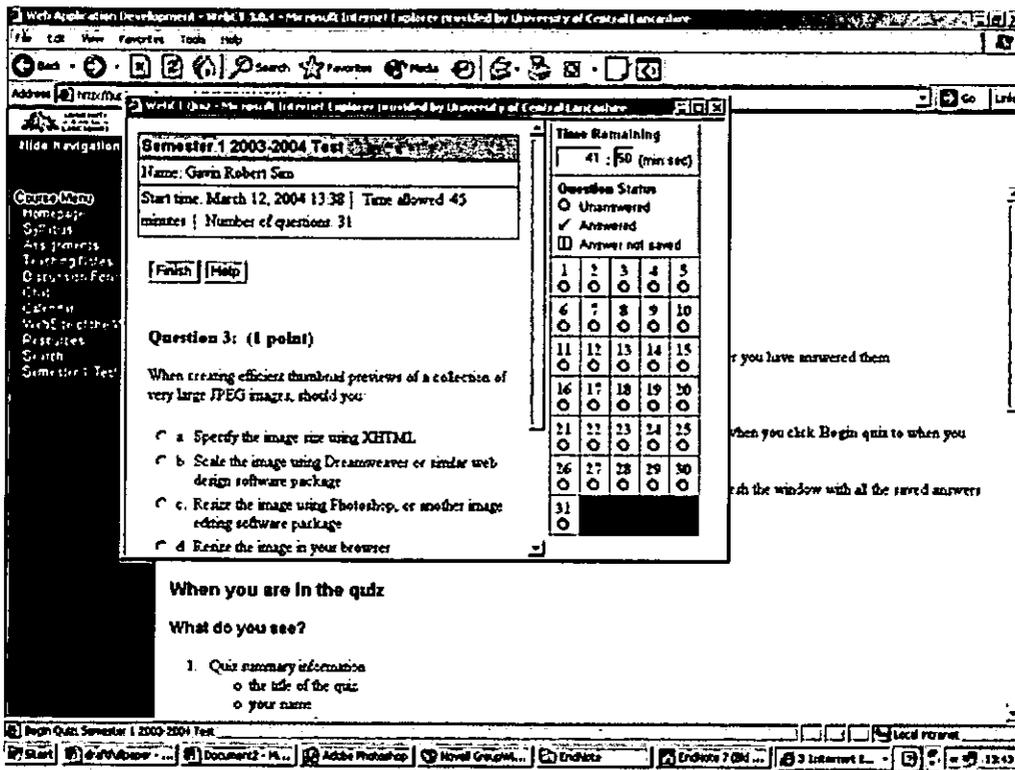


Figure 15 The assessment interface used within WebCT®

In this presentation of the software, to navigate between questions, students had two options: either go directly to a question by selecting it from the right hand navigation panel or press the 'Next Question' button to move through the exam in a linear motion (the next button appears at the bottom of the page). Once a question had been answered the students had to press the 'Save Answer' button, a tick then appeared next to the number to indicate that it had been answered. Finally, when the student had finished answering the questions two navigational steps were required to terminate the exam, but three steps were required if some questions were unanswered or the exam timed out (the time limit had been reached).

5.2.5 Procedure

Both groups received the same test with the questions being delivered in the same order. The duration of the test was 45 minutes and all the undergraduate students completed the test within the time limit. As it was a large cohort multiple computer rooms were used with 25 students in each room and a member of staff invigilating. The postgraduate students had the opportunity to do the test in their class and repeat it in their own time as it was designed for formative assessment.

Once the students had completed the exam they were then asked to complete the questionnaire, this was optional, however, the majority of students completed the questionnaire.

5.2.6 Analysis

The survey gathered both quantitative and qualitative data. The response rate for the undergraduate students was 64% and the postgraduate was 89%. Each of the Likert questions was given a score between 0-4, where 0 was strongly disagree and 4 was strongly agree.

The qualitative data reported by the students was a result of the two open ended questions only and was interpreted by the researcher. This is a subjective process and could potentially lead to bias or misinterpretation however there were only a small number of statements reported by the students, therefore, the task was not complex. For example, *The having to save your answer was annoying next should save it as well for you* and *I dislike having to click save to enter an answer on the test, there was also another button next to the save button* both of these statements were interpreted to be discussing the same problem (having to use the save button). When a usability problem had been identified it was coded using a WU code, the first letter (W) representing the software (WebCT®) and the last (U) indicating it is a usability problem as discussed in Chapter 4.

After the first coding exercise was complete (WU added), the author and a lecturer in HCI also coded the data based on the consequences of the problem and the task the user would be performing when the problem occurred. This coding strategy is a thematic analysis approach, as themes were identified and used to code the data. Fereday and Muir-Cochrane (2006), state that thematic analysis is a form of pattern recognition within the data, where emerging themes becomes categories for the analysis. The themes have emerged through analysis of the literature surrounding the domain in Chapters 2 and 3 in this study. The coding used to establish the consequences of problems was based on the following scale:

- Dissatisfied – the user would be unsatisfied but it is unlikely to affect their overall test performance
- Possible – there is a possibility that the problem may affect the users test performance

- Probable – it would probably affect the users test performance
- Certain – It would definitely affect the test performance of the user

As the research was concerned with problems with unacceptable consequences this scale is used as an alternative to Nielsen's severity rating scales (Nielsen, 1994b):

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

It was anticipated that Nielsen's severity rating scale may not be effective in the context of CAA. For example a problem could occur after the student has submitted their answers and this could be classified as a usability catastrophe, however it may not affect the test performance of the user. Therefore it was judged necessary to devise an alternative scale which would concentrate on the consequences of the problem to the end user. Unlike Nielsen's severity ratings where numbers are used to represent the item, in the coding scheme used in this study first few letters of the item was used. This was felt necessary, in order to prevent any confusion with Nielsen's scale which is used in later studies.

Each problem was classified to one of the consequences codes, for example, Certain, based on academic judgement as to whether, if the problem occurred, a student would have sufficient grounds for appeal. This coding would be used to identify the problem and merge (by task step and severity) the results in subsequent chapters and the same method has been used throughout the thesis.

5.3 Quantitative Results

The first task of the user was to access the test. Users are required to enter a user name and password and this is essential for authenticating the user and recording the results (Bonham *et al.*, 2000). The students were asked 'Did you have any difficulty accessing the test?' as all the students were familiar with accessing the university

network this question directly related to the WebCT® environment and the results are displayed in Table 2.

Answer	Undergraduate	Postgraduate
No	45 (69.2%)	13 (65%)
Yes	20 (30.8%)	7 (35%)

Table 2 Results to the question ‘Did you have any difficulty accessing the test?’

In both groups over 30% of students had difficulty gaining access to the test within the WebCT® environment. Given the high percentage of students having difficulty accessing the test, for novice users this could be deemed a usability problem as it hinders the students accessing and starting the test on time. The problem was coded with a W for WebCT®, U for usability and then a number as a unique identifier.

- WU1 – Problems accessing the test – Poss – S

The problem was judged to be a *possible* unacceptable consequence as it may disrupt the test and cause additional stress to the student, the *S* indicates that the problem arose whilst starting the test.

The mean scores for the two specific usability questions are displayed in Table 3.

Question	Undergraduate		Postgraduate	
	Mean	Standard Deviation	Mean	Standard Deviation
The test was easy to use	3.18	0.73	2.85	1.22
The navigation was clear	3.03	0.77	2.85	1.18

Table 3 Mean scores for usability questions

The results would suggest that the student perceived the application to be easy to use and navigate. The undergraduate students had a mean score between Agree and Strongly Agree, whilst the postgraduates scores were slightly lower between neutral and Agree on the 5 point Likert scale. A Mann-Whitney U test was performed and there was no significant difference between the two groups for either *The test was easy to use* $U=581$, $p=0.442$ and *The navigation was clear* $U=637$, $p=0.884$. The level of usability does not seem to be affected by the degree level and/or assessment method based on these results. The results would also suggest that the students perceive the software to be usable.

The mean scores and standard deviations for the three questions relating to the students satisfaction of CAA are shown in Table 4.

Question	Undergraduate		Postgraduate	
	Mean	Standard Deviation	Mean	Standard Deviation
This type of testing on a regular basis would be beneficial to my studies	2.83	0.69	2.90	1.07
I would find this assessment acceptable as replacement for part of the final exam	2.95	1.01	2.35	1.46
I found this format of assessment less stressful than a paper based exam	2.91	0.94	2.35	1.27

Table 4 Mean scores for student satisfaction with WebCT

The mean scores for all three questions were between neutral and agree on the Likert scale, suggesting a positive experience. A Mann-Whitney U test was performed to determine whether there was any significant difference between the students' level of satisfaction with the assessment method. No significant difference was found between the students completing the test for summative or formative purposes for the three questions. In response to *This type of testing on a regular basis would be beneficial to my studies* the Mann-Whitney results were $U=590$, $p=0.394$; *I would find this assessment acceptable as replacement for part of the final exam* $U=523$, $p=0.14$; *I found this format of assessment less stressful than a paper based exam* $U=489$, $p=0.069$. The results suggest that the students may find this an acceptable assessment technique reporting a high level of satisfaction with the method, overall, their responses to the questions fall between neutral and agree.

5.4 Qualitative Results

An analysis of the qualitative data from the questionnaires revealed a total of 6 usability problems. The qualitative results focused on the users' tasks identified in Chapter 4 of the thesis. All the problems identified are reported and matched to the users' tasks.

5.4.1 Logging on

Section 5.3 reported a high percentage of students having difficulty accessing the test (WU1 Problems accessing the test), this could have been attributed to them having had no exposure to the assessment software. This problem was confirmed by the qualitative data as two undergraduate students commented on accessing the test stating *It took over 15 minutes for everyone to log in* and *Access to the test was lengthy as no one in the class was registered on the test*. This problem may have

been attributed to an administrative error by the module leader in not releasing the test to all the students. This could have possibly caused unacceptable consequences for some students due to being distracted or an increase in their stress level. Therefore, usability problems within administration may also affect CAA users. It is recommended that students have access to practise tests to understand how the process works (Daly & Waldron, 2002) and alleviate anxiety caused by the introduction of new assessment methods (Zakrzewski & Steven, 2003). This may also help prevent any problems with students not being registered on the course or not having access to the test. The problem of students accessing the test may not be persistent, as having prior exposure seemed to reduce the number of students encountering difficulty within the WebCT® environment (Sim, Horton *et al.*, 2004).

5.4.2 During the Test

When navigating the test environment one user from the undergraduate group stipulated *I didn't realise that when you didn't save your answer that it would completely forget what you had written for each answer*, this was coded as:

- WU2- If you do not press save you will lose your answer when you leave the screen – Prob – D1.3.

Upon further investigation it was found that if a question is answered without being saved, when attempting the next question, the following message appears, *warning: the current question has not been saved since the last edit. Proceed to the new question?* It does not mention that if you proceed your previous answer would be erased.

Relating to this issue another four users from the undergraduate group and one from the postgraduate group commented on the issue of *Saving of the answer should be automatic*, implying they did not like this process and this was coded as:

- WU3 - That they did not like having to save their answer after every question – Dissat – D1.3.

There is no obvious reason why this process could not be automated when the user navigates between questions thus alleviating the problem, this option is available within Questionmark® an alternative CAA application.

Another user from the undergraduate group commented *It would be nice to be able to change answers as sometimes you can change your mind after saving*. This was coded as:

- WU4 - The inability to change your answer once you press save. – Poss – D2.

When returning to a question already answered there is no indication that it is feasible to alter the previous entry. It is possible to make another selection, resaving the new answer but there is no option to deselect the radio buttons used in multiple choice questions. This is a severe problem within summative assessment, particularly if negative marking is used as the user may prefer to leave the question unanswered.

5.4.3 Ending the Test

One student from the postgraduate group reported *I did one question and then I am not sure how in an attempt to go to the next question I exited the test and was not able to return back to the test*. This could be two problems; one relating to navigation and the other about exiting the test, however, it was coded to the later. The problem was coded as:

- WU5 – It is possible to accidentally exit the test. – Cert – E3.

This would be a critical problem if it was a summative assessment as students may lose their results and this may increase test anxiety for the student. As the test here was for formative purposes the student could access the test again but it may affect their perception of the software if they were to use it for summative assessment at a later date.

Further investigation identified two possible ways this could occur. The user would have to press the *Finish button*, which would generate the following message; *Some questions have not been answered: Do you want to proceed?* The latter part of the message could be deemed ambiguous as to whether it means proceed with the test or proceed to end the test. If you proceed another message appears which states *Submit quiz for grading?* If cancel is pressed at either of these two points, it takes you back to the test interface where you can continue. This is the only way a user could not regain access to the test. The language used in the feedback messages is of

particular importance within a university setting as the users' first language may not necessarily be English.

Alternatively, after the user has successfully logged in, the test is generated in a small pop up window and the entire interface cannot be seen see Figure 15. If the user resizes the window he/she could accidentally exit the test by pressing the close button within the browser. Other applications and browsers, such as Opera 7, prompt the user to confirm that he/she wants to exit the application. However, even with Opera 7 this feature is not available when using a pop up window. At the time of the study there was a problem with browser incompatibility within WebCT® as it is only compatible with Internet Explorer and this has been acknowledged as a problem in other research (Pain & Le Heron, 2003). This issue has since been resolved in later versions of WebCT®, for example it is now compatible with the Safari browser on the Mac.

Five users from the postgraduate group commented on the use of text entry boxes expressing a concern over spelling mistakes, for example, *Free text answers were often marked 'wrong' due to punctuation/spelling mistakes*, coded:

- WU6 - Answers were marked incorrect due to spelling mistakes. – Cert – D1.1

The undergraduate students did not report this problem as they did not receive immediate feedback as the lecturer had opted to moderate the test results to take into account spelling mistakes. If text entry boxes are used without moderation then students with poor spelling would certainly be disadvantaged and may have grounds for appeal, this would be a certain unacceptable consequence of using CAA if the answer would have been marked correct in a different test mode.

5.5 Conclusions

This initial pilot study had two primary objectives and four minor objectives specified in Section 5.1.1. The first primary objective was to determine if:

- Usability problems exist which could have an impact on students' test results thus leading to unacceptable consequences.

The qualitative data reported in Section 5.4 revealed only a small number of usability problems, 6 in total, and two of these would certainly lead to unacceptable

Chapter 5

consequences. Therefore within WebCT® usability problems exist that could have an impact on students test results and the hypothesis is true.

The second primary objective was to establish if:

- Students are satisfied with commercial CAA applications.

The quantitative data in Section 5.3 would suggest that the students at both undergraduate and postgraduate level are satisfied with CAA as an assessment method. The students mean responses to the series of Likert questions all fell between neutral and agree indicating a reasonable level of satisfaction.

The first minor objective was *To identify whether the context of assessment, either formative or summative, affected satisfaction.* Overall based on the quantitative data, there was no significant difference between the two groups with regards to usability and their satisfaction of using WebCT® for assessment purposes. Both groups would appear to find it an acceptable technique and had little difficulty in using the software. Therefore, it may be acceptable to gauge students' satisfaction by evaluating CAA in just one context. However, despite an overall level of satisfaction the qualitative data revealed a number of issues with the software and two of these problems would certainly lead to unacceptable consequences see Section 5.5.1.

The next objective was to *Establish if users can report usability problems.* Students were able to report usability problems within a CAA environment but only a small number provided qualitative data. This may be because the Likert scale is relatively easier to complete than providing details about specific issues. The questionnaire was a reliable measure but only a limited number of questions were used with regards to usability, therefore, it may need to be modified if a more thorough insight is required.

A total of 6 usability problems were reported by the students and this will be used to form the initial corpus, therefore, the objective *To start building a corpus of usability problem* has been met. Some of the problems could be alleviated, for example, the problems of accessing the test, students should be given a practise test to ensure they are familiar with the test environment. For summative assessment, this may put extra pressure on the invigilators to ensure the exam starts on time,

especially if they are dealing with a large cohort of students and a number of them are experiencing difficulties.

When a user answers a question he/she entry should be automatically saved and no data should be lost when navigating between screens. Feedback messages within the test need to be explicit and matched to the users' required actions. This could prevent users from accidentally logging out of the software. It may be necessary to use secure browsers to prevent students accidentally terminating the exam in summative assessment, as the severity of this occurring may be significantly higher than for formative assessment.

The final objective *To determine the direction of the research strategy* has also been established. Students were able to report usability problems in WebCT® (see Table 5) below and some of these would have unacceptable consequences, therefore, the next stage is to establish if problems exist in other applications.

5.5.1 Usability Problems Identified in WebCT®

Table 5 provides a summary of the reported usability problems from WebCT® with the consequences code attached and task codes described in Chapter 4.

Code	Reported Usability Problem	Consequence	User Task
WU1	Problems accessing the test	Poss	S
WU2	If you do not press save you will loose your answer when you leave the screen	Prob	D1.3
WU3	That they did not like having to save their answer after every question.	Dissat	D1.3
WU4	The inability to change your answer once you press save.	Poss	D2
WU5	It is possible to accidentally exit the test.	Cert	E3
WU6	Answers were marked incorrect due to spelling mistakes.	Cert	D1.1

Table 5 Usability problems found in WebCT®

5.5.2 Methodological Limitations

The results relating to satisfaction reflect the views of computing students and as such generalisation of the results is rather limited due to the sample used, and the fact that a single subject within computing was evaluated. In addition only one CAA environment was analysed and there were only three questions styles therefore further research is still required. The aggregation of the data was performed by the

author, which could have led to misinterpretation. However, only a few statements were provided, so the process was not complex and thus the classification was judged to be accurate. The students had little difficulty in completing the Likert style questions within the survey tool but only a small number of problems were reported by the students and there was little consistency in the data. Within CAA survey tools might not be the most appropriate method for eliciting usability problems due to low inter-rater consistency.

5.5.3 Research Questions

Having established that severe usability problems exist that can lead to unacceptable consequences in a single CAA application the next stage of the research is:

- Do these problems exist in other CAA software environments?
- Are there additional severe usability problems inherent in other CAA systems that would lead to unacceptable consequences?
- If using a number of surveys how can the data be effectively combined and the usability problems prioritised?
- In using the survey approach is the yield per student still low with respect to reported usability problems?

Chapter 6 2nd Pilot Usability Evaluation

6.1 Introduction

This chapter describes a study examining reported usability problems within a second commercial CAA environment. The main objective of the study is to establish whether severe usability problems existed within this CAA environment that would lead to unacceptable consequences. A secondary objective was to compare the problems found in this environment with those reported in the study in the previous chapter. Ricketts and Wilks (2002) looked at different cohorts' attitudes towards CAA but these did not concentrate on usability. The results also provide a foundation for later work. Some of the work in this chapter was published at the 8th International Computer Assisted Assessment Conference (Sim & Holifield, 2004b).

6.1.1 Objectives

Having established in Chapter 5, that students can identify and report usability problems, this next study aims to identify whether similar problems exist in another CAA environment, Questionmark Perception®. This would provide additional evidence to answer the hypothesis presented in Chapter 1:

- Usability problems exist which could have an impact on students' test results thus leading to unacceptable consequences

In addition the following objectives are examined:

1. *To establish the extent and severity of usability problems within the Questionmark® environment.*

In the study reported in Chapter 5 there were a small number of problems found within the WebCT® testing environment. These problems may also exist in other environments, along with additional issues that could cause unacceptable consequences.

2. *To produce a list of known usability problems within this CAA environment expanding the corpus.*

This study aims to produce a list of known problems within the Questionmark® environment. This corpus can be used as a benchmark for known problems within the software.

3. *To establish if using surveys the yield per student is still low with respect to reported usability problems and determine the direction of further research*

Chapter 5 established that problems exist with unacceptable consequences and if additional problems are found within this application but inter group consistency is low, then the next stage of the research will be to perform heuristic evaluations as discussed in Chapter 4.

6.1.2 Scope

This study is devised to establish if usability problems exist in another CAA environment and to determine whether these have unacceptable consequences. Questionmark Perception® was selected as this is widely adopted within Higher Education (Cosemans *et al.*, 2002; Pretorius, 2004). A default interface template was selected with minor modifications to the layout, positioning the navigation to the left hand side of the page as opposed to the bottom. This was done in order to prevent vertical scrolling. The study was conducted under summative conditions and, therefore, the design was constrained to avoid any possible ethical issues in relation to students' grades. For example, some severe usability problems were identified by the first cohort Group A (QU2 and QU7 in Section 6.4) and the second cohort were notified of these prior to starting the exam. As a consequence of this action the reported results might have been affected as the students are unlikely to report these problems or encounter them.

6.1.3 Contributions

The main contributions in this chapter are:

1. Section 6.5 shows usability problems are not application specific.
2. Within this CAA environment there were a number of potentially severe usability problems found by each group that would have unacceptable consequences.

3. Inter group consistency is low and open ended questions only revealed a small number of problems within the environment, therefore surveys are not effective at eliciting usability problems in CAA.

6.1.4 Structure

The structure of the remainder of this chapter is as follows: Section 6.2 reports the experimental design and the results are presented in Sections 6.3 and 6.4. The discussion of the results is presented in Section 6.5 and conclusions are reported in Section 6.6 with a summary of the findings, identification of a number of usability problems, limitations, and further research.

6.2 Study Design

Chapter 1 highlighted the fact that some parts of the thesis research were conducted in conjunction with other projects, in this instance the study was also looking at the adoption of CAA within the department. The survey tool used quantitative methods to ascertain students' satisfaction with Questionmark Perception® software, in the context of summative assessment, to address the requirements of this additional project. Student satisfaction was evaluated in Chapter 5, and the results suggested they were satisfied with the technique, therefore it was decided not to further investigate this construct within the thesis, but focus on usability. Only the qualitative data gathered from the questionnaires was used to answer the research questions specified in Section 6.1.

6.2.1 Participants

The sample consisted of two cohorts of 1st year undergraduate students studying the Programming Design and Implementation (PDI) module. Group A were the 2003-2004 cohort comprising of 101 students and group B were the 2004-2005 cohort consisting of 116 students. The other modules were all from the 2004-05 cohort Group C were the Web Development module comprising 108 students, Group D the VB module with 19 students and finally Group E were from the Network Design and Implementation with 66 students. The samples consisted of mixed genders, diverse ethnicity and a varied age range.

6.2.2 Apparatus

In all instances all the students in each group were required to sit the exam at the same time. Within the university there is no single room available for sitting online exams, therefore, the students used several computer labs to complete the test. However, for group A the specification of the equipment varied depending on the room, with some students viewing the test on 17" or 19" monitors. However, all the questions and options were visible on both monitors so no additional scrolling was required. This was not an issue for the others as the PC's and monitors all had the same specification.

The save as you go feature within Questionmark® was enabled so that students were not required to save after each question. When this feature was disabled students have reported that it is an inconvenience (Sim & Holifield, 2004b).

6.2.3 CAA Question Design

The lecturer responsible for the module designed the questions to be used within the CAA application, determined the duration of the test and scoring algorithm applied. This varied for each of the modules as there was no formal procedure established. For this reason there was a difference in the amount of time each group had to complete the test, number of questions and scoring algorithm applied see Table 6.

	Duration	Questions	Question Styles
Group A	1.5hs	25	MCQ
Group B	2hrs	25	MCQ
Group C	2hrs	35	MCQ (11) MR (1) Essay (8) Text Entry (15)
Group D	2hrs	50	MCQ
Group E	2hrs	29	MCQ (25) Essay (4)

Table 6 Overview of test design for each group

The essay questions required the lecturer to mark the students' answers, no automated marking of these were possible within the software. The increase in time for the second cohort (group B) was not a result of findings from the first test but a request from the lecturer. It was observed that no student actually required the full

amount of time and Bull and McKenna (2001) suggest that students should not receive more than 40 questions per hour, therefore, it is unlikely that any time constraint would influence the students' performance.

6.2.4 Questionnaire Design

The questionnaire design built on the findings from Chapter 5 with respect to the questions relating to the interface and the tasks associated with the users (Section 2.5.1). The questionnaire was again distributed post-test containing a mixture of Likert scale, dichotomous and open-ended questions. The questionnaire aimed to analyse four areas:

- Interface
- Navigation
- Answering Questions
- Exiting the Test

As stated in Section 6.2 the Likert questions were used to address the requirements of another project and have been published at the 8th International CAA Conference (Sim & Holifield, 2004b) and therefore will not be discussed in the results section below. Similar to the first survey tool two open ended questions were used to capture specific issues with regards to the CAA application, these questions were:

- Is there anything you don't like about doing the exam on the computer?
- Is there anything you particularly like about doing the exam on the computer?

The wording of the questions was altered from the original survey tool in attempt to gather both positive and negative experiences of using the software and improve on the low response. Although the research is mainly focused on identifying the severe usability, an understanding of the positive experiences may help inform future research.

The response rate was high with 89 students (88%) completing the questionnaire for group A, for group B it was 82 (70%), group C 83 (76%), group D 19 (100%) and group E 39 (55%).

6.2.5 Procedure

Questionmark Perception version 3.4® was used to deliver the tests to the students see Figure 16. All students had had prior experience of using the software prior to using it in a summative exam conditions.

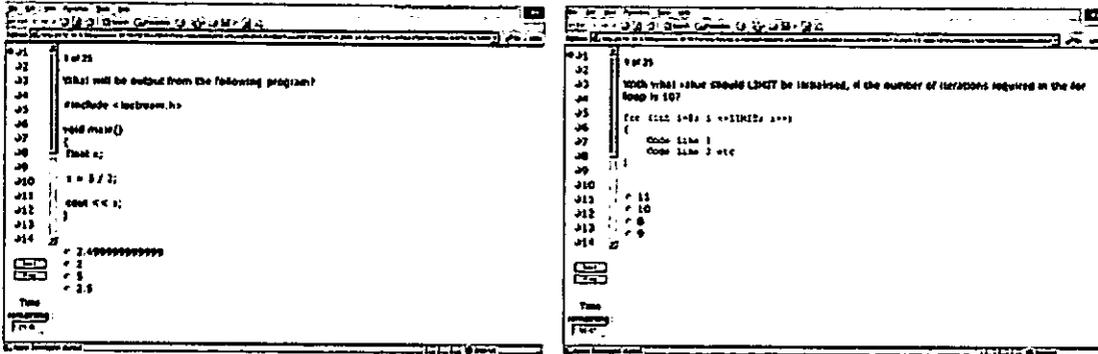


Figure 16 Left Group A interface and Right Group B interface

Each student was informed which computer room he/she would need to attend several days before the scheduled exam. The student was required to log onto the university network and then log into Questionmark® using a different username and password which had been emailed to the student previously. The questionnaires were distributed to the students at the beginning before they started the test. They were asked to complete this at the end, however, it was reiterated that this was optional and were informed that they did not need to complete the questionnaire. Once everyone in the room had accessed Questionmark® the invigilator wrote the password on the white board which enabled the students to gain access to the test. Once the student had finished the test he/she was allowed to leave the exam room and return the questionnaire to the invigilator or leave it on the desk.

6.2.6 Analysis

The student responses to the open ended questions were coded using the same thematic analysis process as described in Chapter 5, however an additional step was incorporated. As before, each problem was given a unique code, then the user task and consequences were assigned, following this the researcher and lecturer in HCI reduced the problem sets by merging duplicate problems. In some instances the consequences of problems varied when problems were merged and a final judgement was made regarding the consequences by examining the problem set.

6.3 Qualitative Results

6.3.1 Group A Results

The students were asked the general question: *Is there anything you don't like about doing the exam on the computer?* From the 89 students who completed the questionnaire only 10 responded with any answer, from which the problems identified below were derived:

qU1 - I did not see the scrollbar leading to more questions

qU2 - Using radio buttons you can't deselect an answer (this was reported x4)

qU3 - It is harder to read from the screen than from paper

qU4 - Typed my in password several times before being let in

qU5 - It's easier to accidentally click the wrong answer

qU6 - Don't trust the computer think something will go wrong

qU7 - If you have a space at the end of your name whilst you attempt to login it does not recognise the username or password.

6.3.2 Group B Results

The same question was asked to group B and they reported the following problems:

qU8 - When you clicked finish you get a blank screen, needs to say thanks (reported x4)

qU9 - There should have been a don't know button

qU10 - There is no way of removing an answer once selected

qU11 - Accidentally finished the test

qU12 - It is hard to see all the code at the same time

qU13 - Never sure if it was processed properly

As in the study reported in Chapter 5, there was a small response rate to the open ended question 6 in relation to usability, with 9 users reporting usability problems from 82 completed questionnaires.

6.3.3 Group C Results

From the 83 students who completed the questionnaire 16 problems were reported:

qU14 - Cannot untick questions if they are negative marking and you decide you don't want to answer x3

qU15 - Hard to see all the code at the same time

qU16 - Didn't accept user ID and password

qU17 - Negative marking it is easy to click wrong answer

qU18 - Never sure if it was processed properly

qU19 - May accidentally press a wrong button and the exam will be over

qU20 - Noise of keys at start of exam x3

qU21 - Larger text boxes for written questions

qU22 - Dyslexic found P and p confusing

qU23 - The computer turned itself off I lost my test and had to retake it

qU24 - Some of the questions are ambiguous

qU25 - Could have put how many marks there was for each question

6.3.4 Group D Results

There were 6 problems identified from the 19 students who completed the VB exam these were:

qU26 - A gap between answers would be an idea as your eyes seem to mix them up – mainly the ones with several lines of code x 3

qU27 - Scrolling down for answers as I wanted to look at the question at the same time

qU28 - Staring at the screen for two hours is painful

qU29 - Can't annotate and understand questions as easy as when on paper

6.3.5 Group E Results

The final group reported 9 problems from the 39 questionnaires completed these are:

qU30 - After finishing there is no way to change an answer if the right answer occurs to one in a later time.

qU31 - You don't have the option to cross out answers you know are wrong

qU32 - The font used made it hard to read.

qU33 - Unless you bring in scribble paper it's hard to work things out

qU34 - The prev, next and flag buttons were too small and too close together

qU35 - Navigation was very bad lower left with small writing

qU36 - Colour of text/background was harsh.

qU37 - Don't like staring at screens

qU38 - Checking the answers at the end of the exam is impossible

6.3.6 Merged Problem Sets

Using the same coding method reported in Chapter 5 the data sets from the 5 groups were merged. The results are displayed in Table 7 below:

Code	Reported Usability Problem	Merged	Consequence	User Task
QU1	I did not see scrollbar leading to more questions		Prob	D
QU2	Using radio buttons can't deselect an answer	qU9, qU10, qU14, qU30, qU31	Prob	D2
QU3	It is harder to read off the screen than off paper	qU12, qU15, qU26, qU27, qU32	Poss	DE1
QU4	Typed my password and copied it several times before being let in	qU7, qU16	Dissat	S2
QU5	It's easier to accidentally click the wrong answer	qU17	Poss	D1
QU6	Don't trust the computer think something will go wrong	qU13	Dissat	T
QU7	Clicked finish and you get a blank screen, needs to say thanks		Dissat	E3
QU8	Accidentally finished the test	qU19	Cert	E3
QU9	It gets wearing on your eyes	qU28	Poss	D
QU10	Noise of Keys at start of exam x3		Dissat	D
QU11	Large text boxes for written questions		Poss	D1.2
QU12	Dyslexic found P and p confusing		Poss	D
QU13	The computer turned itself off I lost my test and had to retake it		Cert	D
QU14	Some of the questions are ambiguous		Poss	D1.1
QU15	Could have put how many marks there was for each question		Poss	D1.1
QU16	Can't annotate and understand questions as easy as when on paper	qU33	Poss	D1.2
QU17	The prev, next and flag buttons were too small and too close together	qU35	Dissat	D3
QU18	Colour of text/background was harsh.	qU37	Dissat	T
QU19	Checking the answers at the end of the exam is impossible		Poss	E2

Table 7 Reported Usability Problems for Questionmark

When a problem was merged the problem was recoded for example qU2, qU9, qU10, qU14, qU30 and qU31 were judged to be the same and as a consequence was recoded to QU2. A capital letter was used to distinguish between the unmerged and merged data sets. The new codes are used for the remainder of the discussion in this thesis. There are a number of usability problems identified in this study that would lead to unacceptable consequences, for example QU2 would affect students' marks if negative marking is used and they accidentally clicked an answer.

6.3.7 Problems Reported across WebCT® and Questionmark®

Of the six problems identified in WebCT® three are also reported in Questionmark® which suggests that some of the problems are not unique to an individual application. The matched problems are:

- WU1 and QU4 – Problems accessing the test
- WU4 and QU2 – Can't deselect a radio button
- WU5 and QU8 – Accidentally finishing the test

In this study although text entry style questions were used the students did not receive feedback immediately unlike the formative tests in WebCT®, therefore WU6 could be discarded from the comparison resulting in an overlap of 60% as 3 of the 5 reported problems were identified. However it is likely that other usability problems exist in WebCT® and the survey method did not reveal all the problems.

6.4 Discussion

Both cohorts were able to report usability problems found within the Questionmark® interface. The inter group consistency was low, of the 19 problems identified, 9 were not reported by another group. There was only 1 instance where a problem was reported by all groups QU3 *Issues with reading of the screen*. All groups identified unique problems therefore it is difficult to ascertain the total number of usability problems in the interface. The use of survey tools does not appear to be effective for eliciting usability problems within the CAA domain due to the low inter group consistency. To overcome this problem it may be necessary to use additional cohorts to further expand the corpus of problems but this would make

the method less efficient. Therefore an alternative approach is required for evaluating the usability of CAA applications.

If additional studies are required then clearly merging and prioritising the data sets from multiple evaluations becomes significantly more problematic. This does not seem unique to this study as Law and Hvannberg (2008) suggest that within HCI the practice of usability problem consolidation is largely open, unstructured and unchecked, therefore a more systematic approach is required.

One of the objectives was to establish whether similar usability problems occurred in two software applications. Because there were a number of major differences to the interfaces and question styles used, there was only a small overlap. Three of the problems identified in the WebCT® test (Chapter 5) also occurred within Questionmark®. In both interfaces users encountered difficulties gaining access to the test WU1/QU4, deselecting answered questions WU4/QU2 and accidentally exiting the test WU5/QU8. There may have been more similarities if the students received immediate feedback for the text entry style questions.

The data gathered in this study is merged with the data from Chapter 5 to form an initial corpus of problems see Appendix C. Although the aim was to find usability problems that have unacceptable consequences, problems coded with dissatisfied at this stage were not discarded in order to attempt to:

- Maximise false negatives
- Minimise false positives
- Minimise errors in coding

If similar problems are reported in other studies then it provides more evidence that the problem is real. When the data is merged with other studies it will also enable the classification to the consequences scale to be reviewed, therefore helping minimise the possibility of eliminating problems which have been classified incorrectly.

Some of the problems may be easily overcome for example it is recommended that using Lightweight Direct Access Protocol (LDAP) may reduce the problem with accessing the test, as authentication could occur by taking the password from the network login (Sim & Holifield, 2004b). To overcome the problem of deselecting radio buttons an additional option should be included that simply says *Do not know*

and this should be scored as zero to ensure that students do not lose marks unfairly if negative marking is adopted.

6.5 Conclusions

This initial study set out to establish *the extent and severity of usability problems within this CAA environment*. The results showed that the users identified 19 problems within the Questionmark® environment and of these, 13 were judged to have unacceptable consequences. These findings, along with the results from Chapter 5, provide substantial evidence to support the hypothesis:

- Usability problems exist which could have an impact on students' test results thus leading to unacceptable consequences

Therefore the initial objectives of the thesis outlined in Chapter 1 have been fulfilled.

The other objective was to *To produce a list of known usability problems within this CAA environment expanding the corpus*. This has also been achieved and the corpus has been expanded to incorporate the problems from Questionmark® and WebCT®. Three of the problems were identified in the previous chapter, WU1 gaining access to the test, WU4 deselecting radio buttons and WU5 accidentally exiting the test, therefore these were not added to the corpus.

The final objective was *To establish if using the surveys the yield per student is still low with respect to reported usability problems and determine the direction of further research*. The questionnaire was administered to 5 groups and the total response was 312, however only 19 problems were reported, producing a yield per respondent of 0.06 (19/312). In addition the inter group consistency is low therefore the use of survey methods might not be a suitable method within CAA.

The objective of the thesis was *“To determine whether severe usability problems exist that can cause users difficulties and dissatisfaction with unacceptable consequences whilst using existing commercial CAA software applications”*. From the data gathered from the surveys in Chapters 5 and 6 it is clear that severe usability problems exist that can have unacceptable consequences therefore this objective has been achieved.

6.5.1 Methodological Limitations

This study was constrained by a number of factors, for example only a limited number of question styles were used in each of the tests and, for ethical reasons, serious errors identified by the first group were shared with the second.

With such a low response rate to the open ended questions and low inter group consistency, other evaluation techniques may be more effective in identifying usability problems within a CAA environment. In the open ended questions there were a total of 19 unique problems identified from the 312 users who completed the questionnaires. Of these 19 problems only 13 may result in unacceptable consequences for the end user, with 2 being classified as certain, 2 probably and 9 possible. The remaining 6 problems would lead to the user being dissatisfied but would not affect the test results or grant them grounds for appeal. Therefore it is necessary to continue the investigation into usability and CAA but adopt a more appropriate method than surveys. Heuristic evaluations may be a suitable solution.

It is evident from Chapters 5 and 6 that usability problems exist in CAA applications that may lead to unacceptable consequences but despite a high response rate only a small number of problems were identified. Although there is some overlap between the problems identified in both Questionmark® and WebCT®, it is not possible to ascertain whether the usability problems that would have unacceptable consequences occur in other applications, therefore further research is still required.

6.5.2 Research Questions

This study's objectives, as outlined in Section 6.1.1, have been met. Within Questionmark® the questionnaire was administered on five occasions with a total response of 312, producing a yield per respondent of 0.06 (19/312) which is judged to be low. The two main hypotheses from Chapter 1 have been satisfied through the research conducted in Chapters 5-6:

- Usability problems do exist which could have an impact on students' test results thus leading to unacceptable consequences.
- Students are satisfied with commercial CAA applications.

As survey methods have been found to be ineffective in revealing usability problems within CAA applications an alternative approach will be investigated. A heuristic

Chapter 6

evaluation will be performed as this seems to be the only viable alternative for ethical and practical reasons identified in Chapters 3 and 4. The combined problem set from the users studies see Appendix C, will be used for establishing the effectiveness of the Nielsen's heuristic set within the CAA domain (Hartson *et al.*, 2003; Sears, 1997)

The objectives for the research will now focus on answering the following questions:

- Would heuristics evaluation techniques reveal the same problems?
- Would heuristic evaluations reveal different problems?
- Are heuristics more effective than surveys at identifying severe usability problems which would lead to unacceptable consequences?
- How can problem sets from multiple evaluations be merged and prioritised?

Chapter 7 Heuristic Evaluations

7.1 Introduction

In Chapters 5 and 6 surveys were found to be an ineffective and inefficient method for identifying usability problems within CAA applications, and inter group consistency was low, therefore another approach was required. The use of inspection based methods appears to be a viable alternative to overcome the problems of user testing discussed in Chapter 3. Therefore this chapter will focus on heuristic evaluations and in particular Nielsen's heuristics (Nielsen, 1992; Nielsen & Molich, 1990) as these are the most widely cited and applied. In certain domains, Nielsen's heuristics may be regarded as dated, but the latter remains the best option for CAA, where you 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 could be essential for purchasing decisions, as well as for instructor training within CAA and software developers.

7.1.1 Objectives

The purpose of this chapter is to provide an introduction to heuristic evaluations, identify some of the issues associated with the method, and define the new research direction within this thesis.

7.1.2 Scope

Since the early 1990's there has been a vast amount of literature published relating to heuristic evaluations. The review predominately used the ACM digital library as the majority of the literature has been published within ACM conference proceedings. In addition other publications were also consulted such as *Interacting with Computing* and the *International Journal of Human Computer Interaction*. These publications represent a significant proportion of key literature within this domain.

7.2 Heuristic Evaluations

The use of heuristics for evaluation purposes within the area of HCI is well documented within the literature and Nielsen's heuristics set are one of the most widely cited and applied (Nielsen, 1992; Nielsen & Molich, 1990). These heuristics have evolved over time and the most recent version is (Nielsen & Mack, 1994):

1. Visibility of system status
 2. Maximise match between the system and the real world
 3. User control and freedom
 4. Consistency and standards
 5. Error prevention
 6. Recognition rather than recall
 7. Flexibility and efficiency of use
-
8. Aesthetics and minimalist design
 9. Help users recognise, diagnose and recover from errors
 10. Help and documentation

The 10 heuristics above were synthesised from 7 sets of heuristics and guidelines, by determining how well they explained an existing problem set using a factor analysis approach (Nielsen, 1994a). This approach was used to determine if a few factors could account for most of the variance in the problem set. The 7 initial heuristics were found to only account for 30% of the variance and it was discovered that 53 factors would be required to account for 90% of coverage of the problem set, which is too much for a practical heuristic evaluation. Nielsen then tried to select the heuristics that offered the widest explanatory coverage of the problem and another set that provided the best explanation of the serious usability problems. It was found that the initial 7 heuristics were almost all found in the top 10 list, apart from error prevention which was still retained. However, 2 problems offering the widest coverage of minor usability problems were added (8 and 9) forming the basis of the heuristic set, the final heuristic, 10, was added at a later date. It is evident that

there is some subjectivity in the development of the heuristics and it is not clear what overall coverage these heuristics offer of the problem set. If used within the context of CAA it would be important for them to identify the problems that would cause unacceptable consequences.

In a classic heuristic evaluation as described by (Nielsen, 1992), several evaluators independently identify usability problems and then their individual lists of problems are aggregated to form a single list of known usability problems within the system under investigation. At this point, or whilst the problems are still individual, severity ratings are attached that indicate the potential impact of the problem. The severity ratings that are used by (Nielsen, 1994b) 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

It is reported that between 3-5 evaluators will reveal about 75% of the overall usability problems (Nielsen & Landauer, 1993; Nielsen & Molich, 1990). This claim is based on the model by Nielsen Landauer (1993) in Figure 17.

$$problemsfound(i) = N(1 - (1 - \lambda)^i)$$

Figure 17 Nielsen and Landauer Formula

Where $problemsfound(i)$ represents the number of unique usability problems found by aggregating the results from each evaluator, N is the number of usability problems found in the interface and λ is the proportion of usability problems found by a single evaluator.

However, other research into heuristic evaluation using 5 evaluators found only 35% of the problems (Spool & Schroeder, 2001) and it has been suggested that high

levels of variance in the evaluators (some finding a high proportion of problems and some only a few) may undermine the sample required. Another limitation of this approach is that the evaluators often have to imagine or try to simulate novice users' mental model of the knowledge level of performance (Fu *et al.*, 2002). Errors may occur for novice users because of inappropriate actions based on their mental model of the system. The discoverability of the problems is also not taken into account by this formula, some problems are simply harder to find than others when using HE (Woolrych & Cockton, 2001).

Despite these limitations Nielsen's heuristics have been used for evaluating a wide variety of domains including hypermedia browsers (Connell & Hammond, 1999), edutainment applications (Embi & Hussain, 2005) and to improve the hardware of musical products (Fernandes & Holmes, 2002). Therefore, these may be more suitable for evaluating CAA applications compared to the other methods discussed in Chapter 3. In recent years there has been a rise in the development of domain specific heuristics. Such heuristics have been developed for specific technologies such as educational software (Evans & Sabry, 2003; Squires & Preece, 1999), as well as for 'features' aligned to usability like accessibility (Paddison & Englefield, 2004) and game playability (Desurvire, Caplan, & Toth, 2004). Niensens' heuristic set may not necessarily be effective for evaluating CAA, but an evaluation method using domain specific heuristics could improve the coverage of problems that may have unacceptable consequences. Therefore the next sections will analyse how domain specific heuristics have been synthesised as it is anticipated that Nielsen's may be ineffective within the CAA domain.

7.3 New Heuristics

A rationale for the creation of new domain specific heuristics, centres upon the potential ineffectiveness of Nielsen's heuristics within the domain. In recent years there has been a rise in the development of domain specific heuristics. As assessment is part of the educational experience, the use of existing educational derived heuristics was considered. However, these are geared towards the learning process, whereas summative assessment is usually performed at the end of a course of study, so there was little match between how users would interact with a learning and assessment application. For example Evans and Sabry (2003) in their e-learning heuristic set, defined the following heuristic *Engaging learner frequency*:

learner content interactions should occur frequently, this would not be applicable within the context of CAA as the main interaction would be answering the questions and the navigation process. One of the purposes of the thesis was to *devise a set of CAA heuristics to enable educational technologists to evaluate the appropriateness of a CAA application*, therefore, a literature review was performed in relation to the synthesis of heuristics. There is a two stage approach to deriving a new set of heuristics, the development stage and the validation stage. In the first stage, the heuristic set is created, in the second, the set is tested for fitness for purpose.

7.4 Developing Heuristics

There have been a number of studies into the development of heuristics (Korhonen & Koivisto, 2006; Mankoff *et al.*, 2003) however, there is no consensus as to the most effective approach. Paddison and Englefield (2004) suggest that there are two main methods for developing heuristics; one being the examination of literature, the other the analysis of data from prior studies. Nielsen (1994a) used the analysis of data from prior studies, the approach used factor analysis and an explanatory coverage process to devise a set of 9 heuristics from a list of 101. Paddison and Englefield (2004) did not especially clarify the meaning of analysing the data from prior studies and this could be interpreted as conducting primary research or carrying out a meta-analysis of other peoples' results. More clarity is found by Ling and Salvendy (2005) who identified three methods for developing heuristics, highlighting previous research (Literature), modification of existing (Nielsen's) heuristics and from evaluation results (Primary Research). As these criteria are more explicit, they are used in the discussion that follows.

7.4.1 Developing Heuristics Based on Literature

The e-learning heuristics derived by Evans and Sabry (2003) and Squires & Preece (1999) were synthesised based on an analysis of the literature applicable to the domain, with the researchers extracting key attributes to create the heuristics. This approach was also adopted by Baker *et al.* (2002) in the creation of groupware heuristics and by Paddison and Englefield (2004) in the creation of accessibility heuristics.

7.4.2 Modification of Nielsen's Heuristics

Some heuristics have been created using Nielsen's original set of heuristics as a starting point. In these studies, the original set has been modified by domain experts to synthesise new heuristics, this approach has been used by Mankoff *et al.*, (2003) for evaluating ambient displays and Korhonen and Koivisto (2006) for playability of mobile games. Bertini *et al.*, (2006) developed a set of heuristics for mobile computing based on primary research using usability researchers to independently perform a literature review, derive a heuristic set and empirically evaluate the set.

7.4.3 Primary Research

There are a number of methods that may aid in the creation of a set of heuristics and avoid repetition. Kurosu *et al.* (1999) developed the structured heuristic evaluation method where they divided the heuristics into related subcategories. Expanding on this method Sommervell and McCrickard (2005), proposed a creation process focusing on the system class. This approach identified the critical parameters within the system focusing on the users' tasks.

Table 8 (Section 7.5.4) summarises the methods used to derive heuristics. It is evident from this table that the main development processes tend to focus on previous research and modification of existing heuristics.

7.5 Validating Heuristics

The raw count of the number of usability problems identified is not an appropriate indicator of the effectiveness of a set of heuristics (Gray & Salzman, 1998) as it does not deal with false positives or false negatives. To validate heuristics, certain criteria are used including:

- thoroughness (Sommervell & McCrickard, 2005)
- correctness, coverage and terminology (Paddison & Englefield, 2004)

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 the heuristics adequately represent the domain being evaluated. Effectiveness is based on the ability of the new heuristics to capture all the significant problems within the domain and ease of use is

concerned with the application of the heuristics by the evaluators. Hartson *et al.*, (2003) report that effectiveness is a combination of thoroughness and validity. Formulas have been devised to calculate these criteria, see Figure 18.

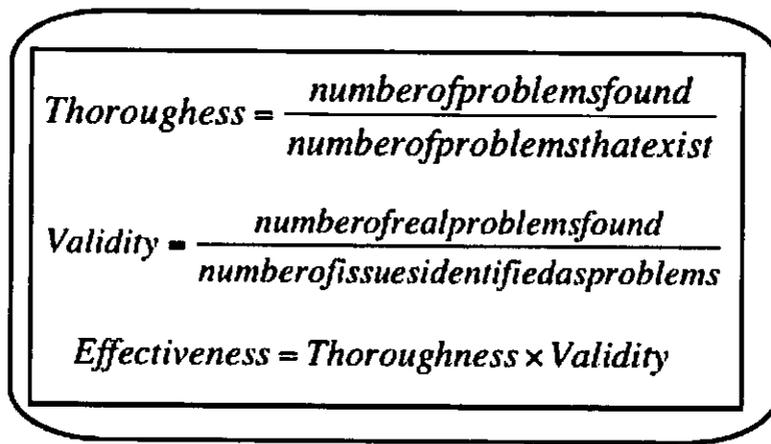

$$\text{Thoroughness} = \frac{\text{number of problems found}}{\text{number of problems that exist}}$$
$$\text{Validity} = \frac{\text{number of real problems found}}{\text{number of issues identified as problems}}$$
$$\text{Effectiveness} = \text{Thoroughness} \times \text{Validity}$$

Figure 18 Hartson et al. formulas

One of the constraints with using these formulae is establishing the number of problems that exist. Different evaluation methods tend to identify different usability problems so the total number may never accurately be captured. It is not possible to have closure on the problem set, there can be undiscovered problems. Therefore, the thoroughness score is an upper bounds, newly discovered problems will increase the denominator (Cockton *et al.*, 2007). Whilst the validity will always be lower bounds, additional problems that match predictions will increase the numerator.

In addition to this formula, Paddison and Englefield (2004) further suggest a good heuristic should be concise, memorable, expressive and easy to relate to underlying knowledge and principles.

Validation of the heuristics can be achieved using a research based approach in which evaluators perform an evaluation using the heuristics. A direct comparison can be made between the performance of the new heuristic set against Nielsen's original set, or a direct comparison can be made against findings from user studies. These methods, and other issues pertaining to validation, are expanded on in the following sections.

7.5.1 No Validation of Heuristics

In some instances heuristics have been created but are not formally validated (Squires & Preece, 1999). Without validation there is no guarantee that the heuristics will be adopted or accepted within the domain; for example, since this

initial unvalidated set of e-learning heuristics were developed, a further two sets have been developed (Evans & Sabry, 2003; Reeves *et al.*, 2002), with neither of these studies citing the earlier set of e-learning heuristics. Albion (1999) cited the learning heuristics proposed by Squires and Preece (1999), yet opted to use the set devised by Nielsen and Mack (1994) and another set by Quinn (1996) to evaluate their multimedia application.

7.5.2 Validating Heuristics by using them

A common method to validate heuristics is to use them in a study and report how well the heuristics performed. This is not always very insightful, for example, in the e-learning heuristics (Reeves *et al.*, 2002) the validation merely indicated that a number of important usability problems were found within the system and proceeded to describe them. It did not give a clear indication of how many problems were identified or the severity of these problems. Many of the problems may have been left undiscovered or false positives could have been identified (Woolrych & Cockton, 2000); these two factors may affect the validity of the heuristics along with the reporting format used (Cockton *et al.*, 2004). Therefore, it is impossible to claim that the e-learning heuristics are valid without further research.

Using a single evaluation it is possible to establish coverage limitations as evident in the playability heuristics for mobile games (Korhonen & Koivisto, 2006), which revealed 16 problems that could not be classified to a heuristic, thus requiring further modification of the set. If additional games were evaluated using these heuristics, additional problems may be identified which cannot be classified to the set and further modifications may be necessary.

7.5.3 Comparison of new Heuristics with Nielsen's Heuristics

Baker *et al.* (2002) compared the performance of evaluators using their shared workspace groupware heuristics to the original data sets from Nielsen's experiments (77 inspectors of the Mantel System and 34 of the Saving Systems) and found by overlaying the results, that their evaluators' performance was similar. It was anticipated that the performance of the evaluators using the groupware heuristics would improve in the field as the evaluators who participated in the studies had no

incentive and were not highly motivated, therefore, the performance might improve using a different set of evaluators.

7.5.4 Comparison Between Heuristics and User Studies

In comparing the results from usability studies with those from heuristic evaluations, unique problems are identified in both methods. A comparative study (Desurvire, Kondziela, & Atwood, 1992) used both heuristic evaluation and user studies and the results revealed that the expert evaluators performing the heuristic evaluation only revealed 44% of the problems reported in user studies. Additional problems were identified that did not appear in the user studies and these were judged to be potential problems, modification of the evaluation method could have revealed the problems to be true positives (real problems). In another study by Jeffries *et al.* (1992) they discovered that heuristics discovered more problems than user studies, and in a study by Karat, (1992), user studies revealed more problems than heuristics. This could be a result of a number of factors such as experience of the evaluators, their motivation or false negatives (usability problems being discarded). Although many of these studies reported here are over a decade old, in a more recent comparative study using domain specific heuristics, it was found that the heuristics outperformed the user studies (Desurvire et al., 2004). Fu *et al.*, (2002) suggested that this variation could be related to the complexity of the interfaces and user tasks; claiming that if significant domain knowledge is required then user studies would outperform heuristics. It is apparent that relying on a single evaluation method does not reveal every problem within a system and, therefore, the validation of domain specific heuristics is problematic based on comparison techniques. This is further supported by Gray and Salzman (1998) who reviewed the validity of 5 comparative studies including: Desurvire *et al.*, (1992); Jeffries *et al.*, (1992) and Nielsen (1992) and found validity issues with each study. For example, in the study by Nielsen (1992), it is claimed it suffered from conclusion invalidity in that the results contradict the claims or the claims were not investigated. In the study it stated that “*usability specialists were much better than those without usability expertise by finding usability problems with heuristic evaluation*”, however, it was not clear what effect the Heuristic Evaluations had on the evaluator ability to find usability problems.

Another problem in using a direct comparison of results from a heuristic evaluation and other evaluation methods is the evaluator effect (Hertzum & Jacobsen, 2001). There is variability between the performance of evaluators when performing an evaluation and even when using double experts as recommended by Nielsen (1994a) there still is variability in the number of problems identified (Sim *et al.*, 2006b). Also there is no clear definition or guidance in the literature as to what constitutes an expert in relation to heuristic evaluations. When adopting a between-subject design there is a risk that these differences between evaluators may bias the results. Gray and Salzman (1998) identified this issue in Jeffries and Desurvire (1992) as they used few evaluators it was not clear whether the effect was caused by chance or whether the evaluators performing the heuristic evaluation performed better than average. Fu *et al.* (2002) claim heuristic evaluations are more effective at predicting usability problems that more advanced users will experience. Experts may have difficulties in predicting the behaviour of novice users within a system, therefore, missing a number of problems. In contrast Woolrych and Cockton (2001) suggest that heuristic evaluations appear to work best for identifying superficial and obvious problems. Therefore, a direct comparison between two heuristics sets may be problematic for the reasons identified.

As evidenced in Table 8, many of the domain specific heuristics developed to date appear not to be thoroughly validated.

Reference	Domain	Developed based on	Validating Heuristics	Validation Criteria
(Squires & Preece, 1999)	E-learning	Modification of Nielsen's heuristics and Previous research	No validation	No Criteria
(Baker <i>et al.</i> , 2002)	Groupware	Previous research	Two different groups evaluated two systems using new heuristics	Effectiveness Coverage
(Reeves <i>et al.</i> , 2002)	E-learning	Modification of Nielsen's heuristics	Used set to perform single evaluation	No Criteria
(Sommervell <i>et al.</i> , 2003)	Large Screen Information Exhibits	Previous research	No validation	No Criteria
(Evans & Sabry, 2003)	E-Learning	Previous research	Performed three evaluations and compared results with Nielsen's set	Correctness Effectiveness
(Mankoff <i>et al.</i> , 2003)	Ambient Displays	Modification of Nielsen's heuristics	Direct Comparison with Nielsen's set	Correctness Coverage Effectiveness
(Paddison & Englefield, 2004)	Accessibility	Previous research	Single evaluation and survey of evaluators	Correctness Coverage Effectiveness
(Desurvire <i>et al.</i> , 2004)	Playability of Games	Previous research	Results of evaluation compared to user studies	Correctness Coverage Effectiveness
(Bertini <i>et al.</i> , 2006)	Mobile Computing	Previous research and evaluation results	Direct comparison with Nielsen's set	Correctness Coverage Effectiveness
(Korhonen & Koivisto, 2006)	Playability heuristics for Mobile Games	Modification of Nielsen's heuristics and Previous research	Used set to perform single evaluation	Coverage Effectiveness

Table 8 Development and validation of domain specific heuristics

If domain specific heuristics are to be devised an understanding of usability within the context of CAA is required to help better inform the design.

7.6 Heuristic Evaluations Research Design

Based on the literature it is anticipated that Nielsen's heuristics will be ineffective within the CAA domain as they are generic and, therefore, domain specific heuristics will be required. This assumption was then used to deduce the following hypothesis:

- Nielsen's heuristics are ineffective within the CAA domain.

There is also uncertainty whether the severity ratings proposed by Nielsen will be adequate. Hertzum (2006) analysed the reliability of severity rating and found them to be low. A more specific scale may be necessary for CAA to improve the reliability, such as the one proposed in Chapter 4. Establishing the reliability of the scale is beyond the scope of work carried out in this thesis.

The purpose of the heuristic evaluation reported in Chapters 8 will be to determine the effectiveness of Nielsen's heuristic set. If they are found to be ineffective the research will focus on the synthesis of domain specific heuristics for CAA. It is anticipated that additional studies would be required to expand the corpus of usability problems to maximise coverage. This approach is used as an alternative and compliment to the survey method in an attempt to yield wider coverage of CAA applications, thus extending the problem set. From a practical point of view heuristic evaluations would be easier to schedule, as these could be performed at any time throughout the year, instead of a reliance on academics incorporating CAA into their modules assessment strategy.

7.6.1 Analysis of Heuristics Data

The data provided from the heuristic evaluations, will be in the form of qualitative and quantitative data. The problems reported provide qualitative evidence of potential usability problems within the application, along with the heuristic it violated and a severity rating using Nielsen's scale. A decision was made to use this scale as it may have been confusing for the evaluators to use a new severity scale that they were unfamiliar with, therefore, the unacceptable consequences scale

outlined in Chapter 4 would be used retrospectively for analysis and filtering of the problems.

Once problems are identified analysis is usually performed to determine if the problem is kept or discarded, five possible prediction outcomes have been identified (Woolrych *et al.*, 2004), these are:

- True Positive – discovery of a real problem
- True Negative – a problem that is correctly eliminated
- False Positive – a problem that is incorrectly retained
- False Negative – a problem that is incorrectly eliminated
- Missed Problems – a problem that is failed to be discovered

Only the first two categories are desirable and falsification testing is usually required to classify the problems, however, this is not feasible for ethical reasons as user testing is required as discussed in Chapters 3 - 4. However, by performing a series of heuristic evaluations and using a mixed methodology approach it was anticipated that few problems would be missed, false negatives would be minimised and true negatives maximised. The filtering and merging of problems from the various evaluations are discussed in Section 7.7.2.

The validity of usability inspection methods (UIMs) can be measured based on the formula proposed by Hartson *et al.*, (2003) discussed in Section 7.5. This formula will be used to establish the effectiveness of Nielsen's heuristics for evaluating CAA in Chapter 8, using the data set from the surveys in Chapter 5-6 as the actual problem set (APS). Even if the heuristic set are found to be ineffective, they will be used to expand the corpus as this method appears to be the best option for CAA, it is expected that they will still reveal plausible usability problems and there is no real viable alternative.

In Chapter 8, two researchers and two HCI lecturers will perform an open card sort of the raw data in order to aggregate this into a single list of usability problems for the application and to remove any duplicates. The heuristic evaluations in Chapters 8 will examine whether there is a low inter-observer consistency based on the number of unique problems identified.

7.6.2 Limitations of Heuristic Evaluations

In conducting a heuristic evaluation with the purpose of expanding the corpus care needs to be taken to address the known factors that may affect corpus quality these are:

- Evaluator Effect (Hertzum & Jacobsen, 2001; Jacobsen & John, 1998)
- Different question styles within CAA applications
- Different CAA applications
- Whether the evaluation would be formative or summative in context
- Data capture method (Cockton *et al.*, 2004; Woolrych & Cockton, 2002)
- Number of Evaluators (Woolrych & Cockton, 2001)

It is also known that a heuristic evaluation is likely to yield a different problem set than user studies (Desurvire *et al.*, 1992; Jeffries & Desurvire, 1992) and ideally falsification testing would be performed to remove any false positives ensuring the quality of the corpus (Woolrych *et al.*, 2004).

7.7 An Evidence Based Design Approach to Corpus Building

If Nielsen's heuristics are found to be ineffective then an evidence based design approach to the synthesis of heuristics will be adopted. It is clear from the literature that the reliance on a single method for developing heuristics may result in some important aspects being overlooked or yielding biased results based on the evaluator's experience. A mixed method approach to developing heuristics may address the shortcomings of creating and validating based on a single method. By using a combination of surveys, heuristics and literature the following factors will be addressed to ensure the quality of the corpus:

- Evaluator Effect
- Different question styles
- A range of CAA applications
- Context summative or formative

- Cohorts

An evidence based design approach for developing heuristics is proposed, the corpus will be developed over a number of studies and any remaining confounds will be addressed through a literature review.

7.7.1 What is Acceptable Evidence?

Determining what constitutes acceptable evidence is a key challenge for evidence based design methods. A meta-analysis approach could be adopted where sources of evidence include guidelines, journal papers or using grounded theory based on primary research, but in each instance careful attention has to be paid to the credibility and validity of the data to ensure the quality of the corpus. In an evidence based design approach for the development of heuristics, the evidence gathered from these various sources would be used to develop a set of heuristics that would address the validation criteria of coverage established in Section 7.5. The efficacy of the use of an evidence based design approach to develop heuristics would be tested by showing how this criterion has been satisfied.

7.7.2 The Research Strategy - Summary

Based on the methods used by others in the study of the literature in developing heuristics in Section 7.4, three main sources of evidence were identified as being appropriate for an evidence based design approach for the synthesis of heuristics. These were: existing heuristics e.g. Nielsens' heuristics, primary research for the development of the corpus, secondary research from the literature to deal with the remaining factors. This is represented in Figure 19 which is a modified version of the process outline in figure 14.

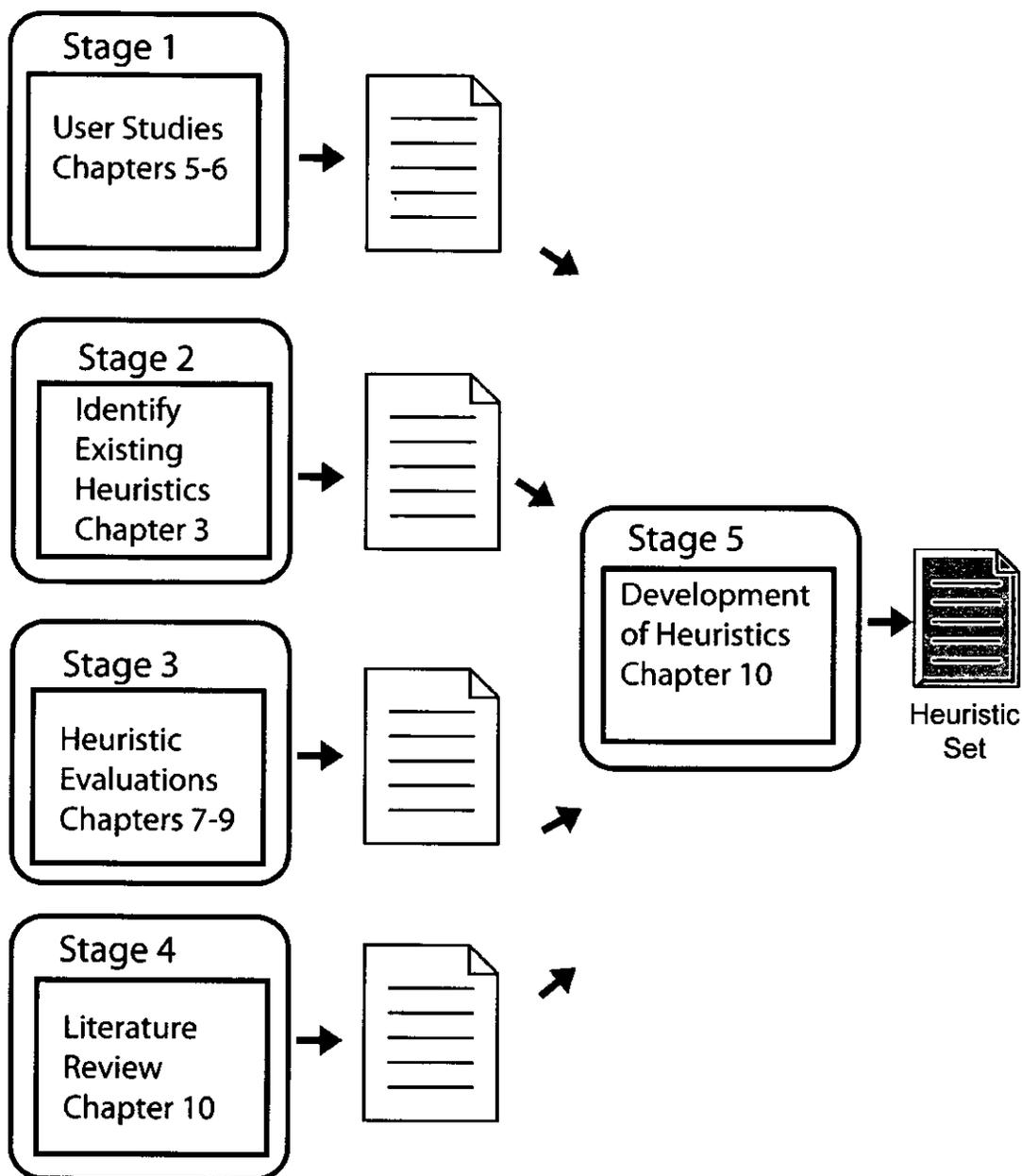


Figure 19 Sources of evidence that feed the development of new heuristics

For the purpose of this research a sequential process is adopted drawing on evidence from the different sources to ensure appropriate coverage of the new heuristics set, other criteria such correctness and effectiveness will not be evaluated at this stage. Within this research the initial starting point is the analysis of existing heuristics, followed by primary research and literature review. It is possible to start the development of the new heuristics earlier.

7.7.2.1 Stage 1 – Pilot Studies

The user studies are designed to establish if usability problems that can lead to unacceptable consequences exist within a number of CAA applications. The data

sets can be used to identify any shortcomings of the existing heuristics identified in stage 2 and applied in stage 3, by validating their effectiveness within the domain. Desurvire *et al.*, (2004) applied this process of validation by comparing data from heuristics and user studies in the creation of the playability of games heuristics.

7.7.2.2 Stage 2 – Existing Heuristics

The second stage, Existing Heuristics, determines whether new heuristics are necessary by investigating existing sets. This process involves an analysis of the literature relating to both heuristics evaluations and the domain under investigation (Chapter 7). The output from this stage is a list of heuristic sets that may be suitable for evaluating the domain see, Table 8 in Section 7.5.4. Ling and Salvendy (2005) suggested that it is naïve to develop domain specific heuristics without consulting Nielsen's original heuristic set. Therefore, existing heuristics would be evaluated by domain experts to establish their suitability.

7.7.2.3 Stage 3 – Heuristic Evaluations

The heuristics identified in the second stage are used to evaluate the domain. There are two purposes to this stage, to identify any shortcomings with the existing heuristics by examining their effectiveness, and secondly to provide a data set that will expand the corpus. A series of heuristic evaluations will be performed to expand the corpus to deal with the known confounds of software types, question styles, context of use and evaluator effect.

7.7.2.4 Stage 4 – Audit from the Literature

A literature review is then used to provide evidence from other applications to support coverage of the heuristics. Using digital resources, conference proceedings and journals, a review of the domain is performed to elicit reported problems within the domain. This is a time consuming process as problems may not necessarily be reported in usability studies, however, by having completed stage 3 an understanding of the vocabulary and problems within the domain would help refine the search. The aim is to produce a data set of usability problems derived from the literature, thus expanding the corpus and addressing factors that were not adequately dealt with in the previous stages these being: question styles, subject domains and software.

7.7.2.5 Stage 5 – Synthesis of Heuristics

In the final stage the data sets are filtered removing any problems that may not lead to unacceptable consequences, they are then merged by mapping each problem to their associated task. This will produce a list of user tasks and their associated usability problems. From this, themes will be established and these themes will form the basis of the heuristic set. Throughout this whole iterative process the heuristics will evolve in relation to the number and the definitions based on the evidence gathered. This process will ensure that a heuristic set has been devised that maximises coverage of the domain.

7.7.3 Merging data

At the end of Chapter 6 the survey methods were merged into an aggregated list. This was done in order for the data to be used to establish the effectiveness of Nielsen's heuristic set in Chapter 8 using the formula outlined by Hartson *et al.*, (2003). The results from the heuristics will be merged within each study and no problems will be discarded until the end. The data sets from the heuristics and surveys will be compared to reinforce usability problems, deal with the known factors (these are discussed in Section 7.7) that could affect corpus quality and try and reduce the potential of false positives. The data sets from each study will be filtered removing any problems which were classified as dissatisfied, as these would not lead to unacceptable consequences. The filtered problem sets would then be merged to their associated task and from this a set of heuristics will be devised, this is further discussed in Chapter 11.

7.8 Conclusions

Although Nielsen's heuristic set are the most widely cited and applied within certain domains their suitability has been questioned, resulting in the development of domain specific heuristics. However, from the analysis of the literature surrounding the development and validation of heuristics it is evident that no single methodology has been adopted for the creation of heuristics and in most instances heuristics have been created without any validation.

Based on the literature review, the assumption was made that Nielsen's heuristics would be ineffective within the CAA domain and a new hypothesis was established:

- Nielsen's heuristics are ineffective within the CAA domain.

If the hypothesis is proven then a corpus building strategy will be adopted using the evidence based design approach outlined in Section 7.7. It is important at this stage to address the known factors that may affect the quality of the corpus and to ensure maximum coverage of the domain.

A difficulty that still remains is how to effectively merge and filter the problems from multiple evaluations. The corpus will need to be merged, aggregated and the problems prioritised, ensuring that the problems with unacceptable consequences remain. From the literature there does not appear to be an established method for aggregating usability problems from multiple evaluations. This issue will need to be overcome in order to implement the evidence based design approach.

Chapter 8 Pilot Study of Nielsen's Heuristics

8.1 Introduction

This chapter describes a pilot study examining the use of heuristics to evaluate a commercial CAA environment. The objective of this study is to determine the effectiveness of the Nielsen's heuristics using the formulae proposed by Hartson *et al.*, (2003). In addition, other objectives are to establish whether heuristic evaluations can reveal severe problems within a CAA application that would cause unacceptable consequences and to discover if the usability problems that are reported compare well with those reported in the studies in Chapter 6.

8.1.1 Objectives

As described above, the objectives are to establish within a CAA environment, whether heuristic evaluation uncovers usability problems that have unacceptable consequences and to discover if similar problems are discovered, using this method, to those reported by users. In fulfilling these objectives the following hypothesis can be answered as reported in Chapter 7:

- Nielsen's heuristics are ineffective within the CAA domain

Other objectives were:

1. *To establish whether Nielsens' heuristics (Nielsen, 1994a; Nielsen & Mack, 1994) are appropriate for evaluating CAA environments.* There are a number of different heuristics available for performing usability evaluations but arguably the most widely adopted are Nielsen's heuristics. However, these were not developed specifically for CAA environments and may not be appropriate to the domain, therefore the effectiveness will be evaluated.
2. *To identify the impact the evaluator effect will have on the development of the corpus.*

In Chapters 5 and 6 the inter group consistency was very low, this needs

to be further examined as it may have an impact on corpus quality as false positives may be retained.

3. *To establish the extent to which context influences the identification of problems.*

CAA applications can be used under various assessment methods including formative or summative assessment. It may be that different problems are identified within different context of use.

4. *To identify further research / issues relating to heuristics and CAA.*
As discussed in the literature review usability of CAA is a relatively new field with little published work so this study aimed to provide grounding for the subsequent work relating to heuristics.

8.1.2 Scope

This study was devised to establish if heuristic evaluations can uncover similar problems to those reported in surveys. Unfortunately, due to time constraints it was not feasible for the evaluators to aggregate their results into a single list of problems or agree final severity ratings. Therefore a card sorting exercise was performed to merge the data and this is reported in Section 8.2.5.

8.1.3 Contributions

There have been a number of comparative studies examining the effectiveness of heuristics compared to user testing (Jeffries & Desurvire, 1992; Lavery *et al.*, 1997; Woolrych & Cockton, 2001) however, these are not within the domain of CAA. The main contributions in this chapter are:

1. A list of usability problems within the CAA application which would have unacceptable consequences; reported in Section 8.3.
2. Nielsen's heuristics are shown to be relatively ineffective within the CAA domain, Section 8.3.7.

8.1.4 Structure

The structure of the remainder of this chapter is as follows: Section 8.2 reports the study design and the results are presented in Section 8.3. The conclusions are

reported in Section 8.4, with a summary of the findings, identification of a number of usability problems, limitations, and further research.

8.2 Experimental Design

The design was between-subjects single factor with two conditions: Formative and Summative assessment. There are several heuristics that can be used for heuristic evaluations. For this study the decision was made to use Nielsen's heuristics as they are the most generic and widely applied. The heuristics used are:

1. Visibility of system status
2. Maximise match between the system and the real world
3. User control and freedom
4. Consistency and standards
5. Error prevention
6. Recognition rather than recall
7. Flexibility and efficiency of use
8. Aesthetics and minimalist design
9. Help users recognise, diagnose and recover from errors
10. Help and documentation

8.2.1 Participants

The sample consisted of 11 HCI practitioners of both genders and a diverse age range. The candidates were given a questionnaire to establish their prior experience of heuristic evaluations in order to allocate them to one of two groups. The groups were balanced based on the evaluators' experience of performing a heuristic evaluation. Group A consisted of 5 evaluators whilst Group B had 6. Both groups completed the same test but evaluated the application within different contexts, group A had summative test conditions whilst group B had formative. None of the users had any prior experience of using the software for assessment purposes but were experienced in respect to assessment practices.

8.2.2 Apparatus

Questionmark® for Windows® was used to deliver the test, but unlike the previous studies in Chapter 6, this was a standalone application and did not rely upon Internet access. The application was loaded onto the evaluators' laptops which varied in both specification and manufacturer. The application was designed to be portable and operate under Windows® operating systems, so the evaluators experience resembled real life as not all users would have the same machine.

8.2.3 CAA Questions Design

In order to provide a reasonable user test it was necessary to provide a 'test' environment for the evaluators. To do this, seventeen questions were designed by the researcher based around general knowledge. These questions were based on three question styles that were known to be used for assessment purposes within computing; Multiple Choice, Text Entry, and Essay (Sim & Holifield, 2004a). The test consisted of 10 MCQ, 5 text entry and 2 essays with the same questions presented in both tests contexts (Summative and Formative).

8.2.4 Procedure

The evaluators conducted the experiment in one morning. All evaluators received a copy of Nielsen's Heuristics, an explanation of Nielsen's severity ratings and information relating to the context of use. They were not presented with the consequences scale as this would be used to analyse the problems post-hoc in a similar manner to Chapters 5 and 6. In addition, they were given a form on which to record the usability problems found. This form simply required them to report the problem, heuristic it violated and attach a severity rating, an example of a completed sheet is shown in Figure 20.

Summative

A

Problem	Heuristic No.	Severity Rating
No clear indicator of whether system is loading successfully indicator of which questions answered	1	3
Horrible aesthetics layout, background, spacing	8	4
Time remaining not clear		
Cannot edit a question	7/3/5	4
Allows non numeric data in data fields entry Only checks if return is hit not if hyperlinked to other questions	5/6	4
Flag - what is it		
No help	10	-
Scroll bars when not necessary Question 7	2	
Have to check a box for each question		

depends if you want to look professional or not both ways.

depends if the sys needs it.

Figure 20 Completed Heuristic Evaluation Report Sheet

The evaluators were allowed to categorise a usability problem as a violation of multiple heuristics.

To access the software, both groups were required to follow the same procedure. They were required to open the application and enter a username and group.

Questions were presented using single question delivery to minimise scrolling and the main navigation was on the left hand side, Figure 21.

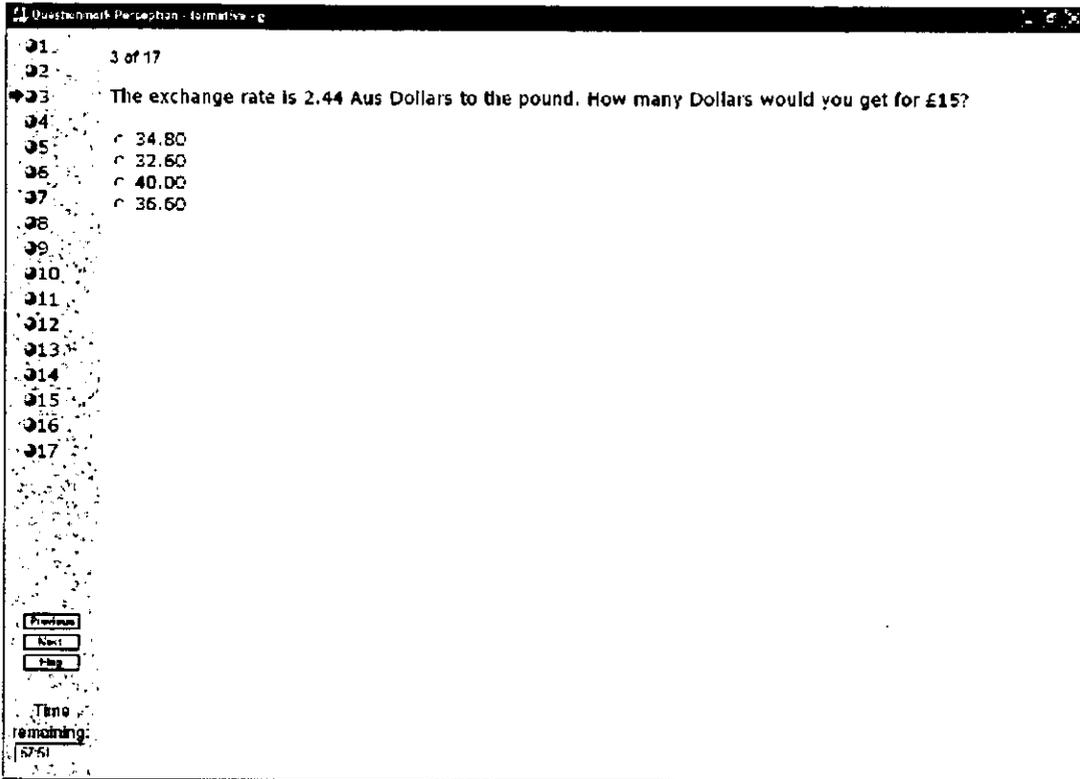


Figure 21: The assessment interface used within Questionmark® for Windows®

The proceed button was only revealed once all questions had been viewed, pressing this would save the answers to an Access® database which was configured to work with the application. Once all the evaluators had completed the evaluation the reporting sheets were collected by the researcher. Unfortunately due to time constraints it was not possible for the evaluators to come together and collectively merge the individual problem sets.

8.2.5 Analysis

A week after the heuristics evaluation was performed the heuristic sheets for formative and summative assessment were analysed separately using card sorting. An open card sort was used to aggregate the list, 2 research student and 2 lecturers in HCI completed the task recording formative problems followed by the summative with a short break (30 minutes) in between. Each of the statements recorded by the evaluators was examined to establish whether or not it was a unique problem (one that no other person recorded). If a problem was recorded by more than one evaluator, the different versions of this were aggregated into a single problem. For

each problem, the overall severity rating was calculated based on the mean scores (there was variability in the severity attached), rounded to the nearest whole number. This practice of using independent evaluators to aggregate the problems has been used in other usability studies (Ardito *et al.*, 2006).

To determine whether a problem identified in the heuristic evaluation was also reported in the user studies, an additional card sorting exercise was performed. The reported problems from the user studies in Chapter 6 were in the context of summative assessment, therefore, only the results from the summative heuristic evaluation were compared. The researcher and a lecturer in HCI participated using the same post-hoc coding method as described in Chapter 4. Each of the recorded problems from the heuristic evaluation was analysed with the reported problems from the user studies to establish if the same problems were revealed using both methods. If both the researcher and lecturer agreed that the problems were the same they were judged a match. In some instances there was disagreement, to resolve this there was a discussion about the problem and agreement was reached.

8.3 Results

8.3.1 Number of Problems Found

On average evaluators in Group A, who examined the software in the context of summative assessment, identified more usability problems than Group B, see Table 9.

	N	Mean problems per evaluator	Standard Deviation
Formative	6	7.33	3.78
Summative	5	10	5.20

Table 9 Mean value of problems identified based on context

Within the context of summative assessment, initially, the evaluators recorded a total of 50 problems; these were aggregated to 41 problems with only 4 problems being identified by two or more evaluators. For example three evaluators commented on the flag button stating *Flag – what is it, Flag button not clear and It is not clear what the flag does.*

For formative assessment there was a total of 44 recorded problems, these were then aggregated to leave 33 problems; here there were 9 problems identified by two or

more evaluators. For example two evaluators commented on the proceed button, stating *Proceed is unclear* and *Button says proceed and action is submit*. In this study the inter-rater consistency was slightly higher than the summative context as less unique problems were identified. The results from the evaluations are presented in Table 10.

	Formative	Summative
Raw Problems	44	50
Card Sorted	33	41
Final Card Sort	23	24

Table 10 Number of problems found within each of the heuristic evaluations

Both groups identified more usability problems compared to the aggregated problems revealed by the user studies in Chapter 6. This would suggest that heuristics may be an effective method for use within CAA as 312 users revealed only 19 problems in the context of summative assessment whilst the 5 evaluators performing the heuristic evaluation identified 41. Therefore the yield per respondent in the heuristic evaluation is $41/5 = 8.2$ compared to $19/312 = 0.06$ for the surveys. However this data does not take into account inaccurate predictions leading to false positives or the severity of the problems identified, therefore the figure for the heuristics might be lower.

8.3.2 Evaluator Effects

There is variability in the performance of the evaluators who conducted the heuristic evaluation. As discussed in Chapter 7, to determine the number of evaluators required Nielsen and Landauer (1993) claim that a typical value of λ to be 31%, this is the percentage of known usability problems an expert evaluator is likely to find. The results in Table 11 revealed a lower lambda value for each group, with the mean value being 0.24 for the summative group and 0.22 for the formative. Although two evaluators from group A found over 31% of problems, three were under 20% and for group B only one evaluator had a lambda value over 31%.

Summative			Formative		
Evaluator	Problems	Lambda	Evaluator	Problems	Lambda
1	7	0.17	1	3	0.09
2	14	0.34	2	4	0.12
3	17	0.41	3	10	0.30
4	7	0.17	4	8	0.24
5	5	0.12	5	13	0.39
			6	6	0.18

Table 11 Total number of problems found by each evaluator and their lambda value calculated on the total aggregated problems

The low lambda value and small overlap between evaluators is similar to results reported by Coyle *et al.*, (2007) who analysed the data from a number of heuristic evaluations and revealed an overlap of only 14%. The study reported in this chapter used HCI experts as evaluators and there was great variability between the experts in the number of problems found and there was only a small overlap between problems found in both studies. In both groups many of the predicted problems were unique, in that no other evaluator identified the same issue. This is a similar problem to the one noted in using survey methods in Chapters 5 and 6 with low inter group consistency. Barnum (2003) analysed the literature and data from usability studies and highlighted the fact that comparisons between studies of the same application with different groups often reveal different problems. Therefore, based on the data from the heuristic evaluations of Questionmark®, there is still the problem of low inter group consistency, and it may be necessary to add additional evaluators to expand the corpus. By increasing the number of evaluators, thus creating a large corpus of usability problems, one of the unresolved issues is how to effectively merge the data sets from various studies and minimise any bias from the evaluators. However, by increasing the number of evaluators this will have added cost to the evaluation process if used as part of a software development life cycle and the aggregation of the individual problem sets will become more complex and time consuming. In addition a mechanism would be required to prioritise the most severe problems from the corpus, without this it may just be left to evaluator judgement and this is not a reliable method.

8.3.3 Attaching Problems found to Heuristics

The evaluators could record a single problem against a number of heuristics. Based on the raw data (un-aggregated list) for summative assessment, 30 problems were

recorded to a single heuristic, 7 to two and 8 to three. While for formative 36 were recorded to a single heuristic, 5 to two and 4 to four. Altogether there were a total 7 reported problems which were not be classified to a heuristic, this accounted for 7.45% of all reported problems. Table 12 displays the number of problems classified to each of the ten heuristics.

Heuristic No.	Summative Problems	Formative Problems
1	12	8
2	6	5
3	5	7
4	7	5
5	11	10
6	3	1
7	5	2
8	7	3
9	6	3
10	6	7
Unclassified	5	2

Table 12 Number of problems classified to each heuristic

In this study both groups classified the highest proportion of problems to Heuristics 5 and 1, with 6 having the least number of violations. It may be that within the context of *CAA Support recognition rather than recall* is not a suitable heuristic.

8.3.4 Severity Ratings

Each evaluator attached severity ratings to the problem once they had been identified and the majority of problems were classified. Table 13 shows the number of problems classified to each of the severity ratings using the raw data.

Context	NC	0	1	2	3	4
Summative (Group A)	7	0	1	8	19	15
Formative (Group B)	1	0	3	19	12	9

Table 13 Number of problems attached to each severity rating

There were a total of 8 problems that were not classified by the evaluators with a severity rating. It is interesting to note that within the summative context the majority of the problems (68%) were rated as major usability problems (3) or

Chapter 8

usability catastrophe (4). This was in contrast to the formative context where the majority were classified as minor or major usability problems.

The first card sorting exercise reduced the problem set to 41 and the mean severity rating score for the problems was calculated. Using the aggregated lists, the effect this had on the severity rating is shown in Table 14.

Context	NC	0	1	2	3	4
Summative (Group A)	2	0	0	7	18	14
Formative (Group B)	1	0	1	14	12	5

Table 14 Mean severity ratings after the card sorting exercise

The summative evaluation revealed 14 problems with a catastrophe severity rating of 4, these were:

hS9 - Even if the student doesn't answer the system proceeds

hS13 - When students don't answer, the colour of the indicator changes to grey but the student is not informed, nothing is said about this

hS17 - Only at the end the student is asked if they want to submit answers.

hS19 - Horrible aesthetics, layout, background, spacing

hS21 - Cannot edit a question

hS22 - Allows non numeric data in data fields entry

hS26 - Why score 7 out of 52

hS29 - Why have some questions not been scored?

hS30 - At end when back into test confusion about quiz

hS33 - Proceed used as a button label

hS35 - Home button takes you irreversibly out of results

hS38 - Q7 Hitting enter closed the window

hS39 - Restarting lost all data

hS41 - No feedback provided after proceed button clicked. Were the answers submitted or not

The evaluators who examined the software within the context of formative assessment only revealed 5 problems with a severity rating of 4, these were:

hF7 - I hit the enter and the program exited and I lost my data

hF13 - It didn't complain if I did not enter an answer

hF15 - The program freezes after I said that I wanted to save i.e. I can't proceed exit etc..

hF22 - Provide finish button

hF31 - Nothing apparently happens when I press proceed

It seems that the severity of the problem could be affected by the context of use. The higher proportion of severe problems may be attributed to the greater consequences of poor usability within summative assessment, as this may affect the students' marks.

Although the statements are not identically phrased it appears that all the severe problems identified within formative assessment have also been identified within summative. Therefore, when evaluating the usability of a CAA environment it may be possible just to analyse the application within one context of use in which case it may be more appropriate to use summative as this is the context in which users have the most to lose.

8.3.5 Further Aggregation of Data Sets

Using the same method applied in Chapter 6, the data was further aggregated incorporating the user task and the consequences. In the summative context the raw data revealed 50 problems, in the first card sorting exercise this was reduced to 41 and the final aggregation reduced it further to 24 problems see Appendix D. Whilst in the formative context the initial 44 problems were reduced to 33 and the final process left 23 problems.

In the process of merging these data sets the accuracy of the first card sorting exercise was questioned and as a result a number of problems were re-classified or discarded. For example one of the evaluators in the formative context reported on their evaluation sheet:

- Prevent errors

- Provide none to answers
- Provide undo facilities
- Provide shortcuts
- Provide better design to support readability

These five statements were discarded as they were judged to be either a reiteration of one of Nielsen’s heuristics or an attempt to devise a heuristic, they are not usability problems. The problem hF30 originally comprised the three statements; *further north of 4 choices is bad grammar; First World War; and Q13/16 is not a question.* The issue relating to grammar was separated out as this was judged to be a separate issue compared to the other two; which were related to the question being ambiguous and the statement relating to the first world war could not be interpreted.

In the summative context hS17 was originally one statement *Only at the end the student is asked if they want to submit answers. Also the unanswered questions are not identified.* These were felt to be two separate issues, one relating to submitting your answers and the other relating to feedback, so therefore were separated.

8.3.6 Problems Identified in both User Studies and Heuristics

After additional aggregation there were 24 problems identified using heuristics in the context of summative assessment and a card sorting exercise revealed that only 6 of these problems were identified in the user studies in Chapter 6.

Code	Problem	User Code
HS04	No feedback provided after proceed button clicked. Were the answers submitted or not	QU7
HS08	No consistency in formatting questions – some take all the space of the screen some just take some	QU11
HS09	Buttons - Previous, Next and Flag are far from where user attention is	QU17
HS11	Lists not visually easy to identify and read	QU3
HS18	Cannot edit a question	QU2
HS24	Hitting enter closed window	QU8

Table 15 Problems identified in both user and heuristic evaluations

Figure 22 shows the overlap between the user studies and heuristics evaluation.

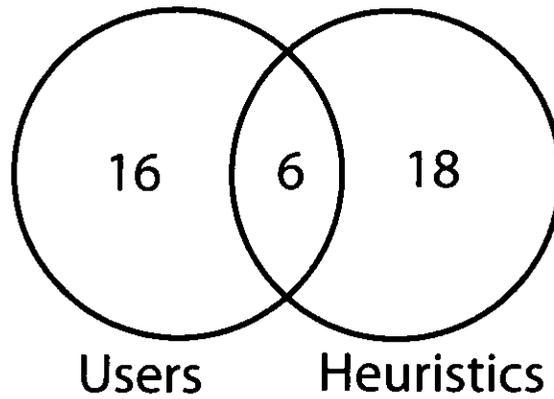


Figure 22 Overlap between two studies

There are 18 problems that were only reported in the heuristic evaluations and the heuristics seemed to miss 16 of the problems identified in the user studies, for example:

- Time remaining not clear
- I can change my answer on feedback – does this change my mark

8.3.7 Analysis of the Effectiveness of Nielsen's Heuristics

Despite the fact that no falsification testing was feasible as user testing is required, using the formula for establishing effectiveness (Hartson et al., 2003) the data from the user studies was used to calculate the effectiveness of Nielsen's heuristics.

Thoroughness = number of problems found / number of problems that exist

$$0.315=6/19$$

There were 6 problems identified in both the user studies and heuristics therefore this figure was used for the number of problems found. As the survey data is in lieu of falsification testing it is difficult to determine whether the unpredicted 18 (24-6) problems from the heuristic evaluation are real.

Validity = number of real problems found / number of issues identified as problems

$$0.25=6/24$$

Effectiveness = Thoroughness x Validity

$$0.079=0.27*0.25$$

The data would suggest that Nielsen's heuristics are relatively ineffective within the CAA domain. Here thoroughness is maximum and validity is minimum, if falsification testing was performed this would reduce thoroughness but increase validity. Problems would be discarded and therefore the number of issues identified as problems would be reduced.

In this study the evaluators were all academics in the area of HCI and, therefore, would be familiar with the domain and may be able to predict problems that may hinder real users and understand the consequences when making severity judgements. The very low effectiveness score may also be attributed to the limited amount of data that was captured in the user studies or although the interfaces were identical, the heuristic evaluation used a standalone version compared to the web based version used in the user tests.

8.3.8 Problems with Unacceptable Consequences

The previous section suggested that the heuristics may have been ineffective but did not examine the types of problems identified in relation to unacceptable consequences or severity. Table 16 shows the number of problems based on the unacceptable consequences scale outlined in Chapter 4.

Study	Dissatisfied	Possible	Probable	Certain
User	6	7	3	3
Heuristic	13	9	1	1

Table 16 Problems with consequences attached

It is evident from the data above that heuristics tend to find different types of problems than the user studies. Over half the problems in the heuristic evaluation were judged not to have unacceptable consequences. For the purpose of this research, the effectiveness was measured again with the removal of all problems rated as dissatisfied. The aim was to establish if problems exist that would lead to unacceptable consequence therefore a direct comparison with this subset of data is performed, the 13 problems from the user study and 11 from the heuristic evaluation. Figure 23 shows the number of problems found in the user studies (Questionmark only) and the heuristics with the dissatisfied problems filtered out.

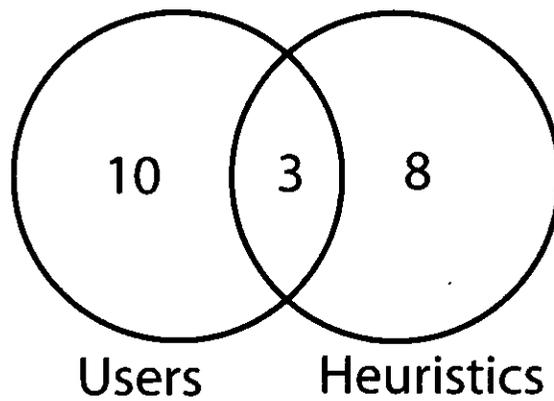


Figure 23 Overlap between studies based on unacceptable consequences

The effectiveness of the heuristics is recalculated based on the problems which may have unacceptable consequences.

- Thoroughness $0.23 = 3/13$
- Validity $0.27 = 3/11$
- Effectiveness $0.0621 = 0.23 * 0.27$

The apparent effectiveness of the heuristic evaluation has decreased slightly when examining the ability of the evaluators to identify problems with unacceptable consequences using the heuristic set. The data from this study would suggest that Nielsen's heuristics are ineffective for evaluating CAA as they fail to identify many of the problems with unacceptable consequences identified in the user studies. Two of the problems which were classified as certain (QU22 and QU13) failed to be picked up in the heuristic evaluation, however it is unlikely that QU13 *The computer turned itself off I lost my test and had to retake it* would have been picked as the probability of this event occurring is low. Overall in lieu of falsification testing on the data sets, Nielsen's heuristics did not appear to be effective at predicting the problems identified in the surveys.

8.4 Conclusions

The formula proposed by Hartson *et al.* (2003) was applied to calculate the effectiveness of Nielsen's heuristics. The effectiveness of the heuristic evaluation was analysed in comparison to the user studies using both the aggregated data and the data with problems classified with dissatisfied removed. Both data sets revealed

a low effectiveness score suggesting that Nielsen's heuristics are ineffective within the CAA domain. In chapter 7 the following hypothesis was deduced:

- Nielsen's heuristics are ineffective within the CAA domain

The results from this study have shown the hypothesis to be true. Therefore domain specific heuristics are required and may be more useful in extracting the problems which would cause unacceptable consequences to the end user.

Another objective of this study was *To establish whether Nielsen's heuristics (Nielsen, 1994a; Nielsen & Mack, 1994) are appropriate for evaluating CAA environments.* The results reported here show that heuristic evaluations can be used to identify usability problems in a CAA environment but with several limitations. The heuristic evaluations identified more problems than the user studies in Chapters 5 and 6. However, despite using HCI experts, many of these problems in both contexts were unique, in the fact they were not identified by another evaluator. A limitation of the study is that no falsification testing occurred so many of the predicted problems may be false alarms. In this study unmatched predictions are treated as false alarms and falsification testing could show they are true predictions. A small percentage of problems identified in this study were also reported by actual users. All of Nielsen's heuristics had problems classified to them there was no redundancy in the heuristic set. However the raw data revealed 7 problems that could not be classified to a heuristic adding further support for their ineffectiveness.

Another objective was *To identify the impact the evaluator effect will have on the development of the corpus.* There was variability between the evaluators in the percentage of problems they predicted, ranging from one evaluator only identifying 9% of reported problems compared to 39% by another evaluator. This study used two groups of 5 and 6 evaluators and if this number was reduced it is likely that many of the problems identified may have not been predicted, hindering the effectiveness of the method and the quality of the problem set.

Context may also affect the severity of a problem, within summative assessment the majority of problems identified were rated between 3 and 4, compared to formative that were rated between 2 and 3. However, all the severe problems identified within formative assessment were also identified within summative, therefore, to uncover problems within the application only one context may need to be evaluated but the

severity may be inaccurate. This result would suggest that context could influence the results and fulfils the final aim which was *To establish the extent to which context influences the identification of problems.*

8.4.1 Methodological Limitations

There were a number of limitations to this study including: Time limited yield which is evident in other studies (Pinelle *et al.*, 2008); Issues with aggregation and impact ratings; Effectiveness computation against survey data, not via falsification testing. Despite the limitations of heuristics the yield per evaluator was 4 (24/6) for summative and 4.6 (23/5) for the formative context. This is considerably higher than the 0.06 (19/312) yield for the survey method and therefore heuristics appear to still be the best available option within the context of CAA.

8.4.2 Research Questions

The objectives of this study have been met and the effectiveness of Nielsen's heuristics can be questioned within the context of CAA but heuristics are still currently the best available option for evaluating CAA. From the analysis of the data, it is apparent that there is very little overlap with the problems reported in the user study and the majority of problems reported in the heuristic evaluations would not lead to unacceptable consequences. Both survey and heuristic methods have problems with rater variation and inconsistency which pose a challenge of how to merge and prioritise problem sets from multiple evaluations. Therefore the objectives of the research will be:

- To extend the corpus of usability problems which have unacceptable consequences. This corpus can be used to synthesise domain specific heuristics for the CAA domain.
- To develop a mechanism for effective combination and prioritisation of results from different usability studies.
- By using Nielsen's heuristics to extract the usability problems, the study will address the following:
 - Does context influence the severity rating?
 - Do experts perform better than novice evaluators?

The results will support further extension of the corpus to cover a range of problem types in order to use the evidence based design approach to synthesise domain specific heuristics for CAA.

Chapter 9 Additional Pilot Study using Heuristics to Expand the Corpus

9.1 Introduction

This study was devised to start expanding the corpus of usability problems, again using Questionmark Perception® software for both formative and summative assessment. In the previous chapter, despite the fact that heuristics were able to identify a number of usability problems the effectiveness of Nielsen's heuristics was questioned, however the method was still judged by the author to be the most suitable for expanding the corpus. Following on from the work in Chapter 8, another objective of the study was to establish if usability problems would only be evident in one of the two contexts. Other objectives were to see if context affects the severity judgement, and if the provision of additional information would assist the evaluators in identifying problems and attaching severity ratings. Chapter 8 used expert evaluators only and there was great variability in their performance, this study aimed to further investigate the evaluator effect as this is a known factor which can hinder the effectiveness of the evaluation, thus affecting corpus quality.

9.1.1 Objectives

As stipulated above, the main objective was to start expanding the corpus of usability problems within a single CAA application. Other objectives were:

- 1. To establish if the severity rating of a problem would vary with context.*

It was envisaged that the severity of problems identified may relate to context. For example, if the test accidentally terminated within a summative context this could be more severe than if it occurred within a formative context.

- 2. To examine the evaluator effect and establish if novice evaluators with domain knowledge can perform a heuristic evaluation of a CAA environment.*

The results from Chapter 8 revealed variability in evaluator performance based on the lambda value and this study aimed to identify if novice users with little experience of heuristic evaluations can identify usability problems in CAA.

3. *To establish whether the provision of additional information about context aided the evaluators in identifying problems and attaching severity ratings.*

The evaluator effect is often widely cited in the literature and by providing the evaluators with additional information about the context of use may help address this issue.

9.1.2 Scope

The study was designed to expand the corpus and build on the findings from the previous chapter. In Chapter 8 it was shown that the majority of problems identified in the summative context were rated as major usability problems or usability catastrophes. This was in contrast to the formative context where the majority were minor and major usability problems. This study was designed to establish if some usability problems were only present in one of the two contexts and further investigate the severity rating of these problems. The study was constrained by the number of experts in HCI who could participate.

9.1.3 Contributions

The main contributions are:

1. A list of context specific usability problems for Questionmark® that further expands the corpus of problems from the previous chapters.
2. Section 9.3.7 shows that in line with other studies, experts are more effective at performing heuristic evaluations than novices, but there was still great variability amongst the evaluators.
3. Additional information provided about context did not appear to assist the evaluators. However no analysis was performed on the appropriateness of the information or the suitability of the format, Section 9.3.6.

9.1.4 Structure

The remainder of this chapter is structured in the following way: Section 9.2 reports the study design and the results are presented in Sections 9.3. In section 9.4 the conclusions are presented with a summary of the findings, identification of a number of usability problems, limitations and further research.

9.2 Study Design

Despite their limitations the decision was made to continue to use Nielsen's heuristics as used in the previous chapter. Using these heuristics, evaluators were still capable of identifying usability problem and this was felt to be the most appropriate method for expanding the corpus.

The evaluators in groups A and B were asked to carry out the evaluations without being given any additional information about context of use, groups C and D received additional information relating to the context. Each evaluator did two evaluations, one evaluation considered the use in a formative assessment (F), the other in a summative assessment (S). To reduce learning effects, the order in which the evaluators applied the heuristics was varied as shown in Table 17.

Group	N	First Evaluation	Second Evaluation
A No Info	2	F	S
B No Info	2	S	F
C Info Provided	2	F	S
D Info Provided	2	S	F

Table 17 The order each of the groups applied the heuristics.

The contextual information provided summarised some of the key issues associated with CAA in formative and summative assessment. For example the summative information was:

- Students are given the opportunity to use the software before completing the test.
- The marks for the test count towards their grades.
- Exam conditions are enforced.
- Students will only be able to login in once, if they exit they can't log in again.

- Once all the students are logged on the exam will start by the invigilator issuing a monitor password.
- Risk is an issue in both the context of cheating and ensuring the data is stored safely.
- At least one of the invigilators has experience of using Questionmark® so can deal with any queries or problems encountered by the students.
- Taking exams usually causes a few people to suffer from high levels of anxiety and this is known to affect test performance.
- Using computers is also another factor that can cause a few people to suffer from high levels of anxiety.
- There are various techniques students use in answering Multiple Choice questions and often negative marking is incorporated to prevent students guessing (which will artificially increase their grades).

The information was partly gathered from the literature and the invigilation process used for CAA within the researcher's own institution. The rationale was that it would provide the evaluator with an overview of some of the issues the users may face such as increased anxiety. Similar information was provided in the context of formative assessment see Appendix F.

9.2.1 Participants

Eight evaluators were recruited to the study. Four of the evaluators were lecturers in HCI and were thus considered to be experts in HCI as well as being familiar with the assessment domain (Double Experts). The other four evaluators were research assistants from within the Faculty of Design and Technology and had no prior knowledge of heuristic evaluations or of computer assisted assessment (this was asked informally) but may have contextual knowledge. The evaluators were split into four groups A, B, C and D where each group consisted of a lecturer and research assistant. This would enable two groups too receive contextual information and evaluate the application in both contexts.

9.2.2 Apparatus

Questionmark® Perception version 3.4 was used to deliver the test, and the same test interface was selected that had been used in the previous studies. There were slight modifications to the interface based on context, for example, the summative assessment interface incorporated a time remaining feature that was not necessary for formative assessment, Figure 24.

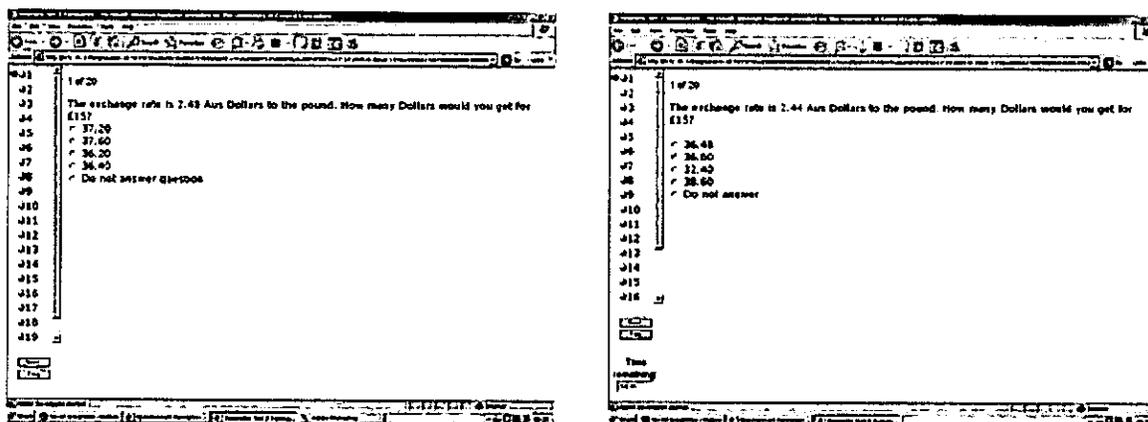


Figure 24 From left to right the interfaces used for formative and summative assessment

A single laboratory within the Department of Computing was used to perform the study to ensure that the specifications of the machines were the same.

9.2.3 CAA Question Design

As in previous studies, in order to provide a reasonable user test it was necessary to provide a 'test' environment for the evaluators. To do this, several questions were designed by the researcher. Three question styles were used there were 14 MCQ, 5 text entry and 1 Essay.

To guard against learning effects and boredom, two sets of 20 matched questions were created (Question set 1 and Question set 2) comprising of questions on Maths, Logic, General Knowledge and Instructions. The order of the questions was shuffled as shown in Table 18. It was felt that by having two different question sets it would reduce complacency, familiarity and would limit the likelihood of evaluators simply recalling problems encountered in the first study.

Groups	Order they saw the questions
Group A	Set 1 Set 2
Group B	Set 1 Set 2
Group C	Set 2 Set 1
Group D	Set 2 Set 1

Table 18 Shows the order the evaluators saw the question sets

9.2.4 Procedure

All the evaluators were given the same brief overview of heuristic evaluation by the researcher and taken through Nielsen’s heuristics and the use of severity ratings prior to completing the first evaluation exercise. This briefing session lasted about 20 minutes, following this they were informed of the task, this was based on the process the students would go through in completing an online test (Sim, Horton *et al.*, 2004).

1. The evaluators will be emailed a user name, password and the URL for the Questionmark® server
2. They will then be required to login
3. They will have to complete a 20 question test answering each of the questions using different input methods and navigating between questions.
4. Once complete - finish the test
5. If formative, examine the feedback and exit (exit only if summative)

The evaluators then went to one of the computer labs within the Department of Computing to perform the evaluation.

Whilst completing the tasks, the evaluators were required to record any usability problems encountered on a form provided. The form was the same design as used in Chapter 8, see Figure 18. Once evaluators completed the task they then matched each problem to an appropriate heuristic and suggested a severity rating. The evaluators were allowed to categorise a usability problem as a violation of multiple heuristics, this method is seen in other studies (Zhang *et al.*, 2003). The researcher collected in the completed forms.

Three days later, the evaluators conducted the second evaluation, which was identical in structure to the first (except there was no introductory talk). After

completing both evaluations the results of the individual heuristic evaluations were then aggregated by the researcher into two single lists of problems, one for the summative interface and one for the formative interface.

Each of these two aggregated lists was sent individually to each evaluator to attach severity ratings approximately one week after they had completed the initial evaluation, see Appendix F. Some of the evaluators completed the attached form immediately and returned it to the researcher but in some instance the researcher had to chase the form from the evaluators, eventually all the forms were returned.

9.2.5 Analysis

The analysis of the data was performed in two stages the first by just the author and the second stage followed the same procedure as outlined in Chapter 4. In stage 1 each of the statements recorded by the evaluators was examined to establish whether it was a unique problem (one that no other person recorded). If a problem was recorded by more than one evaluator this was aggregated into a single problem. The aggregated list was returned to the evaluator to attach severity rating to the problems and the mean severity rating for each problem was again calculated and rounded to the nearest whole number to match the severity rating. There were 8 problems in formative and 12 in summative context that at least 1 of the evaluators could not interpret and therefore did not attach a severity rating to.

In stage 2 an additional analysis and aggregation of the data was performed using the same procedure as the previous chapters, including the task step code and consequences scale to ensure that the presentation of the data was consistent. The revised list of problems was then analysed to establish which problems appeared in both contexts and the data was finally merged into a single list of usability problems associated with Questionmark perception®.

9.3 Results

9.3.1 Number of Usability Problems Discovered

Within the context of formative assessment, initially, the evaluators recorded a total of 56 problems; these were aggregated to 46 problems as 8 problems had been identified by more than one evaluator, some by three evaluators. For example, two

evaluators stated that it was *not obvious when the finish button was shown* and two were unsure *what the flag and unflag buttons did*.

For summative assessment there was a total of 48 recorded problems, these were then aggregated to leave 41 problems, with 5 being identified by more than one evaluator. For example, three evaluators reported that *there should be more spacing between the answers in multiple choice style questions* and four evaluators expressed concern over *being penalised for spelling in text entry style questions*. Table 19 shows the number of problems that remained after each stage of the analysis of the data.

	Formative	Summative
Raw	56	48
Stage 1	46	41
Stage 2	34	28

Table 19 Number of usability problems reported after each analysis stage

9.3.2 Evaluator Effect

Table 20 shows the mean scores for the number of usability problems found in each of the two evaluation sessions. For both novice and experts but in particular for the novice evaluators, the mean score was lower in the second evaluation, one evaluator recorded 12 problems in the first evaluation and only 4 in the second see Table 21. This may have been due to a decline in motivation to participate in the evaluation, the fact they were evaluating a similar interface or they might not have reported problems previously mentioned. It may have been more appropriate to use a between subject design, as in the earlier chapters, to minimise these issues.

Category	Evaluation 1		Evaluation 2	
	Mean	Standard Deviation	Mean	Standard Deviation
Experts	9.0	3.74	8.5	4.80
Novice	6.0	4.08	2.0	1.41

Table 20 The average number of usability problems found based on evaluator experience

Chapter 8 revealed that there was great variability between the performances of the evaluators and similar results are reported in this study see Table 21.

Group	Evaluator Type	Summative Interface	Lambda Value	Formative Interface	Lambda Value
A	Expert	3	.07	5	.11
A	Novice	4	.09	12	.26
B	Expert	15	.35	12	.26
B	Novice	4	.09	1	.02
C	Expert	7	.16	8	.17
C	Novice	1	.02	3	.07
D	Expert	9	.21	13	.28
D	Novice	5	.12	2	.04

Table 21 Total number of problems found by each evaluator with their lambda value calculated on the total aggregated problems

The data revealed a rather low lambda value for the aggregated results for the summative evaluation (0.13) and it was equally low for the formative (0.15). If the experiment had only been conducted with experts then the lambda value would still have been lower than the claimed typical value of 0.31, in this instance, the summative being 0.19 and formative 0.21.

Within the context of summative evaluation, there was great variability between the evaluators. For example, the expert in Group B identified the most problems finding 35% of the reported problems, in contrast, the expert in Group C revealed only 16%. From the novices, there was slightly less variation, the most reported problems from a single evaluator was 26% (Group A) and the least was 2% (Groups B & C). This is far fewer than Nielsen's claim that 5 novice evaluators (no usability experience) would find at least 51% of known problems (Nielsen, 1992). In a study by Slavkovic and Cross (1999) using novice evaluators their results indicated a lower average at only 23%, which they attributed to the complexity of the interface they were studying. Within CAA you would expect the interface to be intuitive in order to prevent errors that could affect test performance and threaten the validity of the test, therefore complexity is unlikely to be the cause of the low score in this study.

If all the evaluators were as good as the best evaluator (expert from Group B) in identifying problems then using four evaluators would reveal 80% of the problems in the summative interface and 73% in the formative. If they had been as poor as the worst, then four evaluators would have found 8% in both the summative and formative interface. This ultimately reveals a weakness in heuristic evaluations, as there is no known method of establishing the effectiveness of the evaluators before they participate in the study. In this study the evaluators were categorised as either

an expert or novice by the author. The evaluator effect in this study and in previous chapters is a concern which may impact on the quality of the corpus. Additional studies are thus required to expand the corpus to diminish the evaluator effect. The potential necessity to conduct multiple evaluations in order to reveal the majority of usability problems, again raise concerns over the effective aggregation of data.

Unlike the studies reported in other chapters this study used a within-subject design. Since the same evaluators were used for both contexts a comparison was performed to determine the effectiveness of the evaluators in finding usability problems. There was a significant Pearson correlation ($r=0.74$, $p=0.036$) between the number of usability problems found by the evaluators in each of the two evaluations suggesting their ability to find problems is consistent between evaluations. The results from this study is most likely attributed to the fact the evaluators were analysing a similar interface within the same domain and prior knowledge may have influenced their judgement of problems or the evaluators behaviour and approach is similar. If the second evaluation was performed after a much greater time frame then the results may have been different. Nielsen and Molich (1990) suggest that there is little consistency in the ability of evaluators to find usability problems and this is supported in this study.

9.3.3 Problem Classification for Formative Assessment

Despite the limitations of the heuristic set each problem identified by an evaluator was required to be related to one or more of the heuristics. It was quite feasible for an evaluator to classify a single problem to more than one heuristic, thus the total number with respect to heuristics violated is greater than the aggregated number of problems. Using the merged data after the stage 1 of the analysis, for formative assessment there was a high proportion of violations against heuristics 1 (ensure visibility of system status), 3 (maximise user control and freedom) and 5 (prevent errors). There were also 7 problems which the evaluators couldn't classify to an appropriate heuristic. These are:

- hF5 - No possibility for the user to change font size
- hF6 - Too much browser information
- hF7 - All navigation keys visible and working, although they must not be used

- hF8 - Status bar information “Applet save applet started” unnecessary and confusing (kiosk mode more appropriate)
- hF28 - I answered q7 but when I returned to it later the answer had gone... but button still blue
- hF40 - This version showed all the buttons so that was a surprise (resolution)
- hF41- Q5 Garbon spelt wrong

Of the 7 problems which could not be classified, 4 of were reported by a novice and 3 by an expert. This provides additional evidence to support the need for domain specific heuristics and to continue expanding the corpus. This could be achieved by keeping Nielsen’s heuristics that are judged appropriate and including additional heuristics, this is further discussed in Chapter 11.

	Nielsen’s Heuristics									
	1	2	3	4	5	6	7	8	9	10
No Problems	7	4	7	5	12	0	3	4	6	2

Table 22 Number of formative problems classified to heuristics

Similar to the results in Chapter 8, heuristic 6 was the least used, suggesting that *Support recognition rather than recall* may not be an appropriate heuristic within the context of CAA. The terminology may need to be modified or a new set of context specific heuristics applicable to the CAA domain may need to be devised. The modification of the terminology has been adopted in other domains such as ambient displays (Mankoff *et al.*, 2003).

9.3.4 Problem Classification for Summative Assessment

Similar to formative assessment there was a high proportion of violations against heuristic 1, see Table 23, but in this study neither 3 nor 5 recorded a high level of reported violations. In this instance there were no reported violations against heuristic 7 support ‘*Support flexibility and efficiency of use*’ and once again heuristic 6 was low, with only one reported violation. In this instance there were 6 problems which the evaluators couldn’t classify to an appropriate heuristic these are:

- hS24 - If I picked I don’t want to answer the question I expected it to indicate I had chosen a non answer

- hS29 - Lost exam answers message came up ‘page has expired’
- hS30 - No font size selection
- hS39 - Tend to read all answers this time rather than just selecting when you first possible correct answer seen
- hS40 - Expectation of getting immediate response (results) even though I know it’s a test simply because it’s on a computer
- hS41 - Learning curve having used it in a formative mode, know what to expect so issues last time are not an issue this time

In this instance 4 of the 6 problems that were not classified were reported by the experts. Different evaluators in both studies identified problems which they could not classify to a heuristic suggesting that Nielsen’s heuristics do not offer adequate coverage of the domain.

	Nielsen’s Heuristics									
	1	2	3	4	5	6	7	8	9	10
No Problems	12	7	3	6	6	1	0	4	4	1

Table 23 Number of summative problems classified to heuristics

9.3.5 Problems Identified in Both Contexts

Of the 46 problems identified in formative assessment only 18 of these were identified in summative assessment see Figure 25. For example, the fact that *the navigation panel does not automatically scroll to reveal the next question* was identified in both contexts. Therefore, the heuristic evaluation would appear to have revealed 28 problems that were unique to formative assessment and 25 unique to summative.

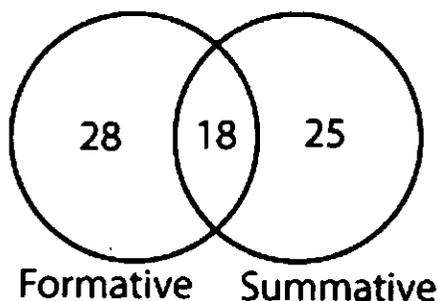


Figure 25 Problems found in both contexts

Some of the problems will be context specific, for example, summative assessment usually has a time limit, therefore, the interface incorporated a clock. However, upon examining the statements it is clear that a number of the problems would also persist in both contexts. One evaluator reported hF26 – *allows user to close down window and lose all work*, this was only identified in the context of formative assessment but could also occur in summative.

This highlights the need to evaluate interfaces in different contexts to reveal a wide range of possible problems. By just relying on one evaluation a number of problems, such as *In one of the fill questions I typed 7 but it didn't go in as I hadn't clicked in the box – if I hadn't looked at question would not have been recorded*, may not have been identified. Woolrych and Cockton (2000) suggest that heuristic evaluations appear to work best for identifying superficial and almost obvious problems and this appears to be the case for this study. For example for the user to experience the problem identified during the formative evaluation relating to the browser window closing he/she would need to perform an unanticipated action. Further evaluations of the interface are probably still required to further expand the corpus.

9.3.6 Inclusion of Information about Assessment

The evaluators in groups C and D were provided with additional information about the context of use. In retrospect it was felt inappropriate to perform any statistical comparison between the two groups as they were both likely to have contextual knowledge therefore and variation is unlikely due this additional information.

Group	Summative	Lambda Summative	Formative	Lambda Formative
Context (C & D)	21	0.12	26	0.14
No Context (A & B)	25	0.14	30	0.16

Table 24 Number of problems found by each group based on context

In both cases the group who received no additional information identified more problems, Table 24, however this does not take into account that many of these problems may be false positives.

9.3.7 Severity Ratings

Each evaluator independently attached severity ratings to the aggregated list of problems see Table 25.

Context	Severity					
	n/a	0	1	2	3	4
Formative	0	0	11	31	4	0
Summative	0	0	15	21	5	0

Table 25 Problems classified to each of the severity ratings

For formative the major usability problems were (with severity 3):

- hF26 - Allows user to close down window and lose work
- hF27 - Using back button in browser exits test rather than returning to question 1
- hF28 - A question was answered and when it was returned to it later it was blank, but it still indicated it had been answered
- hF32 - Can't deselect a radio button question

The summative assessment evaluation major usability problems were (with severity 3):

- hS15 - No option to quit
- hS16 - When all questions attempted finish appears. It exits without confirmation and doesn't check whether and flags are still set
- hS23 - A user thought they put in the correct answer but got an error message, and could no find a solution so had to quit
- hS29 - Lost exam answers message came up Page Expired
- hS32 - There were browser navigation problems in the fact that if you press the back button the exam is terminated

There appeared to be a great deal of variance between the ratings attached to the same usability problem by the evaluators. This is supported by Hertzum and Jacobsen (2001) who suggest that evaluators differ substantially in their classification of the severity of problems. Within the context of formative assessment there were a total of 46 problems identified and in 10 instances at least

one evaluator classified the problem as 0 (not a usability problem at all) whilst another evaluator had classified it as 3 (Major usability problem. Important to fix should be given high priority). For the full problem set see appendix G, Table 26 below shows the problem code and the rating by each evaluator.

Code	N1	N2	N3	N4	Mean	E1	E2	E3	E4	Mean
hF6	3	1	2	1	1.75	3	0	0	1	1
hF9	3	0	1	1	1.25	3	2	3	1	2.25
hF19	3	3	1	2	2.25	3	2	3	0	2
hF21	3	0	2	1	1.5	3	1	2	2	2
hF33	2	1	1	2	1.5	3	1	3	0	1.75
hF36	0	1	3	1	1.25	2	1	2	1	1.5
hF37	0	3	2	2	1.75	3	1	3	2	2.25
hF38	0	2	1	3	1.5	3	1	3	2	2.25
hF39	0	0	1	3	1	3	2	3	2	2.5
hF42	1	1	1	1	1	3	0	2	?	1.67

Table 26 Formative problems where an evaluator rated the problem 0 and another rated it 3

There was disagreement between both sets of evaluators (novices and experts) in their classifications for example, experts classifying problems as 0 and another expert rating it as 3, novices rating problems as 3 with experts rating it as 0. For example, two evaluators (both experts) rated hF6 - *too much browser information* as a 0 whilst two rated it a 3 (novice and expert).

A similar pattern emerged within summative assessment and in this instance there were a total of 41 problems and in 5 cases an evaluator classified the problem as 0 whilst another evaluator had given it 4 (Usability catastrophe. Imperative to fix this before product can be released). An example of this was the rating of hS5 - *Not clear why I would select do not answer question rather than guessing, I don't recall being told the rules for marking*. There was a further 9 instances where at least two evaluators disagreed between a 0 and 3 classification, see Table 27.

Code	N1	N2	N3	N4	Mean	E1	E2	E3	E4	Mean
hS1	2	0	2	3	1.75	4	3	2	?	3
hS5	1	3	2	2	2	4	0	1	2	1.75
hS21	0	3	1	1	1.25	3	3	4	2	3
hS23	3	3	2	3	2.75	4	0	?	?	2
hS32	4	2	1	2	2.25	4	0		?	2
hS2	0	2	1	2	1.25	3	0	0	0	0.75
hS3	0	3	2	2	1.75	2	1	2	2	1.75
hS6	0	0		?	0	3	0	1	1	1.75
hS7	0	0	1	3	1	3	1	2	1	1.75
hS11	1	1	0	2	1	2	1	1	3	1.75
hS13	0	2	2	2	1.5	3	0	0	2	1.25
hS26	1	3	2	2	2	3	0	0	0	0.75
hS38	1	2	2	?	1.67	3	2	1	0	1.5
hS40	0	0	1	1	0.5	3	1	?	1	1.67

Table 27 Summative problems where an evaluator rated the problem 0 and another rated it 3 or more

In both studies evaluators E1 and N1 appeared to rate the problems either more severe or less severe than the other evaluators. For example in the summative context 4 of the 5 problems with a severity rating of 4 were due to E1 classification and problem hS32 both E1 and N1 classified the problem with a severity rating of 4. It may be that the evaluators had difficulty interpreting others comments or it may be that the severity ratings are too generic and they have difficulty distinguishing between the boundaries within the scales.

As discussed in Section 9.2.5, of the 46 problems identified in the formative assessment evaluation there were only 40 problems that all 8 evaluators rated. A Kendall's coefficient of concordance between the eight evaluators was performed on these 40 ratings, $W=0.264$, which is statistically significantly $p<.001$. This indicates that the agreement is not purely by chance. Despite this, calculating the coefficient of determination indicates that there is a 7% consistency among the 8 evaluators. Therefore, it is very difficult to determine accurately the severity of a specific problem, even taking the mean score, if a number of evaluators scored it 3 and others scored it 0, the problem may not be classified to the appropriate rating.

Kendall's coefficient of concordance was also conducted on the summative interface for the 8 evaluators. Of the 41 problems, there were 12, which at least 1 evaluator, couldn't interpret and did not attach a severity rating to, therefore, these were omitted from the analysis. For the 29 remaining problems, $W=0.288$, which is

statistically significant $p < 0.001$, again highlighting that agreement was not by chance. Hertzum (2006) analysed the reliability of severity ratings and suggested that the reliability of the evaluators' severity assessment is so low that it is risky to use a single evaluators' judgement therefore using context specific severity ratings may help evaluators classify problems more reliably.

9.3.8 Severity Reliability over Time

In the first part of the study described in the procedure, Section 9.2.4, the evaluators individually identified problems attached a severity rating and attributed it to a heuristic. Not all evaluators attached a severity rating to every problem reported in Table 20. After a brief period of time they received the aggregated list and were asked to attach severity ratings to the entire problem set. Having established that there was low reliability between evaluators, an analysis was performed to establish the stability of evaluators in rating the problems over time. Table 28 shows the number of problems with a severity rating attached, and whether the evaluator classified it to the same rating when presented with the aggregated data set. In the table 'Match' means that the problem reported by the evaluator on the individual list, was given the same severity rating by them on the aggregated list. The evaluators did not have access to their original data from the evaluation.

Evaluator	Formative		Summative	
	Problems	% Match	Problems	% Match
E1	7	6 (85%)	2	2 (100%)
E2	0	n/a	2	0 (0%)
E3	8	2 (25%)	13	4 (31%)
E4	8	2 (25%)	7	1 (14%)
N1	2	2 (100%)	5	1 (20%)
N2	9	4 (44%)	4	4 (100%)
N3	3	3 (100%)	1	1 (100%)
N4	1	1 (100%)	2	2 (100%)

Table 28 Problems with consistent severity ratings over time

E2 did not attach severity ratings to any of the problems reported therefore there is no percentage for the formative context. The novices seem to be more consistent in classifying the problems to the same severity rating however this may be because they reported on average fewer problems.

There were 3 problems in both the formative and summative context that the evaluator's initial classification differed from their secondary classification

(aggregated data set) by greater than 1. For example in the summative context one evaluator reported *Q20 I put in <7 but got an error message "<7not in " no solution I had to quit*, this was initially classified with a severity rating of 3 and the second classification was a 0. It could have been that the summary changed and the evaluator no longer recognised the problem. If this data is used to prioritise fixes as part of a development life cycle then resources may be allocated to superficial problems due to inaccurate classification. However this problem would probably lead to unacceptable consequences for the student and may affect their overall grade if it is marked incorrect. Therefore it would appear that reliance on using Nielsen's severity rating scale within the context of CAA is not advisable. A new scale may need to be developed along with the heuristics.

9.3.9 Difference in Severity Ratings Based on Context of Use

Section 9.3.5 revealed that there were 18 problems which were identified in both contexts and the severity ratings were the same in 61% of instances. 28% of the problems were rated more severe in formative assessment and 11% were rated more severe in summative assessment. An example of this related to question 20 (in Question set 1) in that it could not be answered. In formative assessment this would not be a major issue for the student, however, in summative assessment they are likely to lose marks so the problem would become more severe. A Wilcoxon test was performed to establish whether there was a significant difference between the severity ratings $Z = -0.632$ which is not significant $p=0.527$, indicating no difference in ratings based on context. This may be because only a small number of the problems are likely to change because of the context of use or due to the variability between evaluators rating problems.

9.3.10 Further Aggregation of Data Sets

Using the same method applied in Chapter 8, the data was further aggregated on the basis of the user task and consequences scale. The summative context's raw data revealed 48 problems, in the first card sorting exercise this was reduced to 41 and the final aggregation reduced it further to 28 problems see Appendix I. During this task 3 problems were discarded these were:

- hS2 - Clues to questions, connects in relevant refreshers on algebra for example
- hs26 - Q7 far too open ended for a test of this nature
- hS41 - Learning curve having used it in a formative mode, know what to expect so issues last time are not an issue this time

Whilst in the formative context the initial 56 problems were reduced to 46 and the final process left 34 problems see Appendix I. One of the problems HF29 - *This version showed all the buttons so that was a surprise (resolution)* could not be classified using the consequences scale as it was a positive statement it was not a problem therefore it was coded with *Good*.

9.3.10.1 Problems in Both Contexts

In section 9.3.5 a total of 18 problems were identified in both contexts based on the initial data set. Further aggregation of the data shown in Figure 26, revealed that a total of 16 problems were identified in both contexts see Appendix J.

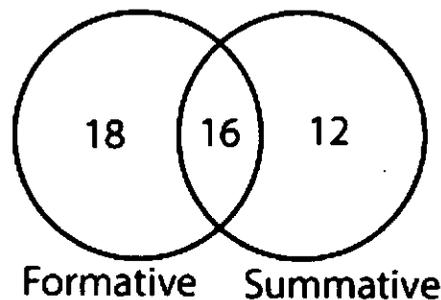


Figure 26 Number of problems in both contexts

The final stage was to aggregate all the problems into a single data set of problems within the Questionmark® application. There were a total of 44 problems identified in all contexts that could be used to expand the corpus.

9.4 Conclusions

The main objective of this chapter was to start to expand the corpus and this study reported a total of 44 usability problems to enable a growing diverse corpus for the development of the heuristics. In expanding the corpus the problem still remains of an effective method for aggregating the data from the multiple studies.

Another objective was *To establish if the severity rating of a problem would vary with context*. It was reported in Section 9.3.9 that there was no significant difference between the severity ratings based on context. However, the problems that were identified as important to fix differed across two contexts and some of these, for instance, hS15 - *No option to quit* and hS32 - *Back button exits test* were certainly contextually important. Based on the findings of Chapter 8, it was anticipated that some of the problems would only appear in one of the two contexts and this proved to be the case, as shown in Figure 24. However, there were problems that were only identified in one context for instance, hF26 - *allows users to close down window and lose work* that could have related to both.

There were similar results to those presented in Chapter 8 regarding inter-rater consistency. There was a lot of variability between the effectiveness of the evaluators in identifying the number of problems within the application as demonstrated through the Lambda value. Even if only the experts had been used to perform the evaluation, the mean Lambda value would have only been 0.21 for formative and 0.19 for summative which is lower than the 0.31 stated as typical by Nielsen and Landauer (1993). An objective of this study was *To examine the evaluator effect and establish if novice evaluators with domain knowledge can perform a heuristic evaluation of a CAA environment*. Choosing to use two different groups of evaluators (experts and novices) demonstrated that, in line with other studies, expert users were better at finding problems. The low inter-rater reliability and high variability in severity ratings was disappointing. Despite their performance being less effective than experts, as anticipated, novice evaluators were able to identify usability problems within the CAA application.

The final objective was *To establish whether the provision of additional information about context aided the evaluators in identifying problems and attaching severity ratings*. With retrospect all the evaluators would probably have contextual knowledge therefore no statistical comparison was performed. However, the provision of additional information about the context of use did not appear to assist the evaluator in identifying usability problems and attaching severity ratings. This may be attributed to the fact that all the evaluators had experience of assessment within higher education, or it could be that the additional information provided was

in some way deficient, or it could be that there were too few evaluators to determine an effect.

Similar to Chapter 8, the suitability of using Nielsen's heuristics alone for evaluating CAA is questioned. There were some instances were:

- No problems were classified to heuristics, *Recognition rather than recall* and *Flexibility and efficiency of use*
- Problems were identified that could not be classified to a heuristic

This gives supporting evidence to suggest that domain specific heuristics are required for CAA. The usability problems that were identified as significant are presented in Appendix J. The data set has been recoded to remove duplications in both contexts.

9.4.1 Methodological Limitations

Although the evaluators were all from the same institution, from a logistical perspective it was difficult to co-ordinate due to peoples work commitments. To aid this process the author aggregated the results of the two evaluations and returned the aggregated lists to the evaluators in order for them to attach severity ratings. However, it would have been ideal if each of the groups aggregated the problems to prevent any misjudgement by the author. For example unrated problems may have been due to the aggregation process; the evaluator no longer recognised or understood their problem. It may have been better to list the merged problems so the evaluators could recognise theirs.

Another issue that is still prevalent is the inconsistency amongst evaluators at identifying problems and attaching severity ratings. It would appear that multiple evaluations are necessary to reveal the majority of problems within a CAA application. The aggregation of data from studies with a larger number of evaluators would require a mechanism to aggregate the data sets whilst minimising bias in the process.

9.4.2 Further Research

This study used one commercial CAA application to grow the corpus whilst examining the relationship between novices and experts. It was found that novices

Chapter 9

could identify problems but on average fewer problems were reported than experts therefore the next study proposes:

- To expand the corpus using Nielsen's heuristics by evaluating three CAA applications.
- To use a large student cohort to perform the evaluations to mitigate against the effect caused by novice evaluators.
- To build on the results of Chapter 8 to determine whether the problems identified are application specific.
- To develop a mechanism for merging the problem sets from multiple evaluations.

Chapter 10 Expanding the Corpus and Developing an Aggregation Instrument

10.1 Introduction

Expanding on the earlier studies, this chapter describes a study that was devised to collect usability problems from three commercial CAA applications: WebCT®, Questionmark Perception® and TRIADS®. All of these have been used within higher education for both formative and summative assessment (O'Hare & Mackenzie, 2004; Pretorius, 2004; Sim, Holifield *et al.*, 2004). This is the first time TRIADS® had been investigated in this research. The objective of the study was to expand the problem corpus and in Chapter 6 it was revealed that some problems may only be prevalent in a single application therefore the number of applications evaluated was expanded. The problems identified will be added to a corpus of usability problems and a method will be devised to aggregate the problem sets from different usability studies. Initial findings from this study have been published at EDMEDIA (Sim *et al.*, 2007).

10.1.1 Objectives

As stated above, the main objective was to expand the problem corpus. Other objectives were:

1. *To develop a mechanism for filtering and aggregating usability problems from multiple evaluations.*

It is anticipated that this evaluation will reveal a large number of usability problems and therefore a suitable method of filtering and aggregating problems will be investigated.

2. *To establish what effect increasing the number of evaluators may have on the number of new problems found.*

In chapter 9 the results showed that many of the usability problems identified were unique problem, by increasing the number of evaluators it is anticipated that more severe unique problem will be identified, thus expanding the corpus.

10.1.2 Scope

As stated, this study was designed to further expand the corpus by revisiting WebCT® and Questionmark®, in addition expanding the coverage of CAA application by using TRIADS®. The study was constrained by using HCI students to perform the evaluations, but as indicated in Chapter 9 these would have some contextual knowledge and are therefore not complete novices.

10.1.3 Contributions

The main contributions are:

1. A Damage Index formula for merging and rating the severity of usability problems is presented in Section 10.2.5.
2. Evidence that, in line with other studies (Hertzum & Jacobsen, 2001) increasing the number of evaluators increased the number of unique usability problem.
3. The findings from Chapters 8 and 9 showed there was great variability in the number of problems identified by the evaluators, similar results are reported here, Section 10.3.6.
4. Although a large number of problems were identified, many were applications specific reported in Section 10.4.1.3.
5. A list of specific usability problems for Questionmark®, TRIADS® and WebCT® which can be used to expand the corpus.

10.1.4 Structure

This chapter is structured in the following way: The study design is reported in Section 10.2, and the results are presented in Sections 10.3 - 10.4. In Section 10.5 the conclusions are presented with a summary of the findings, identification of a number of usability problems, limitations and further research.

10.2 Study Design

The study was devised to investigate the usability of three CAA environments; Questionmark Perception, TRIADS and WebCT, Figure 27. Questionmark is widely

used within higher education for both formative and summative assessment (Sim, Holifield *et al.*, 2004), WebCT is an example of a learning management system that has also been used for assessment purposes (Alexander *et al.*, 2003; Pretorius, 2004) and TRIADS is a university developed system that claims to offer more flexibility than some commercial systems, this has been used within a number of institutions within the UK (Evans *et al.*, 2004; McLaughlin *et al.*, 2004b). All three applications use the internet to deliver the assessment to the user, however TRIADS is reliant on the Authorware plugin being installed.

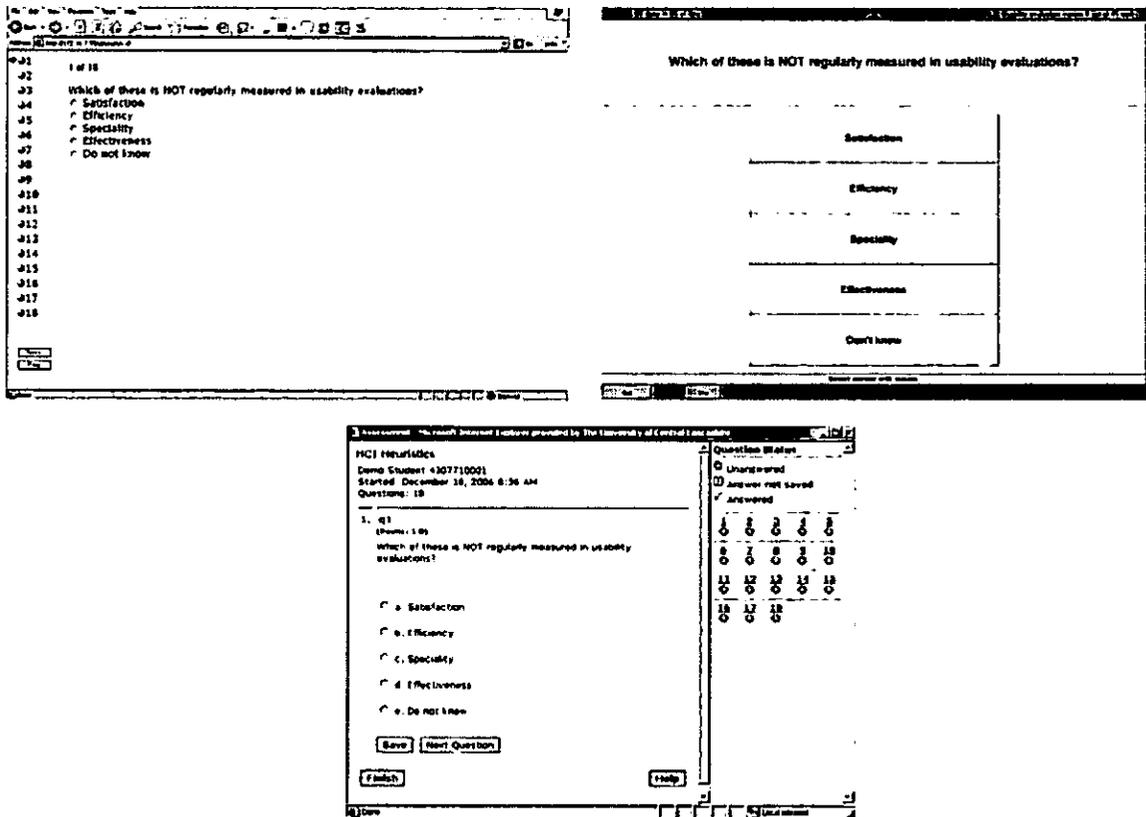


Figure 27 From left to right Questionmark , TRIADS and WebCT.

10.2.1 Evaluators

The study involved 98 students from the Computing Department of the University of Central Lancashire. They participated in the evaluation as part of their coursework for a second year undergraduate HCI module. All participants had prior knowledge of usability and had been trained in heuristic evaluation in the lecture and tutorial prior to participating in the study. This training included a one hour

lecture on heuristic evaluation followed by them performing an evaluation based on Nielsen's heuristics. The training took place the week before the evaluation.

Other studies have examined the relationship between expertise and successful heuristic evaluations (Nielsen, 1992). The evaluators in this study could be perceived to be domain experts in the assessment process as they have had prior experience of it within higher education, having successfully completed their first year. This domain knowledge is also acknowledged in Chapter 9.

10.2.2 Design

As stated earlier, there are several sets of heuristics that can be used for heuristic evaluations. For this study the decision was made to continue using Nielsen's heuristics, as this was still considered the most suitable heuristic set, despite its limitations and the students were familiar with the set. The three software applications were evaluated using a between-subjects design as there was not enough time for each student to evaluate every application. Also in Chapter 9 the number of problems revealed in the second evaluation was lower and it was anticipated by the time they evaluated the 3rd application they would be demotivated. The evaluations took place in several identical computer laboratories at the host institution. Evaluators were randomly divided into the three groups upon entering the computer laboratory, each participant then performed an evaluation on either Questionmark®, TRIADS® or the assessment tool within WebCT®.

10.2.3 CAA Question Design

In order to provide a realistic user test for CAA it was necessary to provide a 'test' for the evaluators. To do this, 18 questions were designed by the module leader for HCI, relating to the first part of the syllabus. These questions were presented in three different styles that were known to be used for assessment purposes within computing and were available in each of the three software applications. The styles that were used were Multiple Choice, Multiple Response, and Text Entry. The first 6 questions were presented as multiple choice, the next 6 were multiple response and the final 6 were text entry. In each of the three different CAA applications, the same questions were presented, the same styles were used and the same feedback was displayed. In addition, the scoring algorithms for the questions were kept consistent across the three applications.

10.2.4 Procedure

The evaluations were performed over a three week period. In the first week, all the evaluators were given a brief overview of heuristic evaluations and taken through Nielsen's heuristics and the use of severity ratings in the lecture. In the second week the evaluators went to one of the computer laboratories within the Department of Computing to perform the heuristic evaluation. Two different rooms were used over the course of three days, but each room had the same equipment ensuring minimal technical variability (e.g. monitor resolution or bandwidth). Upon entering the room the individual students were assigned to one of the three test conditions Questionmark®, WebCT® or TRIADS®. The students sat at computers ready to start the test and, using a script to ensure that each group was treated the same, they were informed of the task and given instructions on how to complete the evaluation. The task was based on the process the students would go through in completing an online test (Sim, Horton *et al.*, 2004).

1. They will be required to login
2. They will have to complete an 18 question test
3. Once complete - finish the test
4. Examine the feedback and exit

Although the evaluators came to the task as a group, they each conducted the evaluation individually. The evaluators were provided with a copy of the Nielsen's Heuristics and a description of each of the five severity ratings. Alongside this they were given a booklet, the students were required to complete a front sheet where they indicated their name and experience, this was used to ensure the sheet could be returned to them the following week see Appendix K.

In addition, they were given a form on which to record the usability problems found. The form required the evaluators to state what the problem was, which heuristic(s) was violated, where each heuristic was violated, and how the problem was found see Figure 28.

Problem found (write a single problem in the space)	Heuristic(s) violated	Where it was violated		How was it found	Severity Rating
		Task	Location		
<p><u>Example</u> Could not get into the room door was locked.</p>	3	<input checked="" type="checkbox"/> Accessing the test <input type="checkbox"/> Navigating within the test <input type="checkbox"/> Answering the question <input type="checkbox"/> Finishing the test	CM26 DO NOT PUT THE ROOM REACTS IT TO THE SOFTWARE!	<input type="checkbox"/> Scanning for problems <input type="checkbox"/> Systematically searching for problems <input type="checkbox"/> Trying to force errors <input checked="" type="checkbox"/> Following users task	4
<p>No note to tell the user to click 'save' before clicking 'Next Question'.</p>	10.	<input type="checkbox"/> Accessing the test <input type="checkbox"/> Navigating within the test <input checked="" type="checkbox"/> Answering the question <input type="checkbox"/> Finishing the test	Page 1.	<input checked="" type="checkbox"/> Scanning for problems <input type="checkbox"/> Systematically searching for problems <input type="checkbox"/> Trying to force errors <input type="checkbox"/> Following users task	2.
<p>When clicking 'Next Question' the answer to the question (error) should be Yes (No not Ok & Cancel. Very Misleading</p>	5.	<input type="checkbox"/> Accessing the test <input type="checkbox"/> Navigating within the test <input checked="" type="checkbox"/> Answering the question <input type="checkbox"/> Finishing the test	Page 2.	<input type="checkbox"/> Scanning for problems <input checked="" type="checkbox"/> Systematically searching for problems <input type="checkbox"/> Trying to force errors <input type="checkbox"/> Following users task	3.

Figure 28 Reporting form used in the heuristic evaluation

The form was based on a design described in (Cockton *et al.*, 2004) and it also required the evaluators to record the severity of the problem. The evaluators were allowed to categorise a usability problem as a violation of multiple heuristics, this method is seen in other studies (Zhang *et al.*, 2003) and had been used in previous chapters. The evaluation took between 40 minutes and an hour, as the evaluators went through the application they recorded any problems they encountered on the evaluation sheets. Once the students had completed the evaluation they handed their completed forms to the researcher and left the room.

One week later, the evaluators participated in the final stage of the study. At this point they were required to aggregate their results into a single list of problems. This was conducted in their tutorial class and upon entering the room each evaluator received his or her original form back (which had been used to document usability problems in the first stage), and was then randomly assigned to a group of between two and six evaluators based on the software earlier examined. In most instances the numbers were between 3 and 5 but in some cases this was not possible because of the number of students within the tutorial group. Thus, all the students that had evaluated TRIADS® aggregated their problems with other students who had evaluated TRIADS®. For each of the three CAA environments there were eight groups, the groups varied in size of between 2 and 6 students. Once within a group the evaluators were required to aggregate their list with the other evaluators and were required to document the problems, indicating the frequency of each and agree a severity rating. This process would yield a single complete list of usability problems and severity ratings for each of the twenty four (three applications with eight groups each) groups, for an example see Figure 29.

Heuristic Evaluation Part 2 - WebCT

G

	Problem	Frequency	Severity Rating
Q1.	No note to tell the user to click 'Save' before clicking 'Next Question'	3.	2.
Q2	When clicking 'Next Question' 'are you sure' should be Yes/No not okay/cancel.	3.	3
Q3	When accidentally clicking <input checked="" type="checkbox"/> it restarts at Question 1 not where you left off.	1	2.
Q3	Top left corner, question no. is shown in 2 different formats	1	2.
Q7-12.	On multiple choice questions no instructions to let you it's multiple choice.	3	3.
Q12	Instructions beginning at end of a sentence (looks like 1 sentence)	1	4.
Q13	Textbox has a 1 by it.	1	2.
Q13/14.	Can type gibberish into textbox.	2.	3.
Q14	If leaving textbox empty no error message given.	1	3.
Q15	there is a close bracket but no open one.	1	1

Figure 29 Example of a groups aggregated problem set

10.2.5 A Method for Aggregating the Data

The analysis of the data was quite complex. At the beginning, the data was made up of 98 paper forms, each containing a list of problems and severity ratings for one of the three applications. This raw data was then reduced in the following way:

- Stage 1: Student Aggregation: In the second week of the study, the students clustered into small groups based on the software they had evaluated and

aggregated the problems they had found. This resulted in 8 x 3 aggregated lists of usability problems, severity ratings and frequencies of discovery (the percentage of students in the group that had found the problem) see Figure 27. At the problems were coded with a letter to represent the group and a number to represent the problem.

- Stage 2: Duplication Treatment: Looking at a single software application at a time, two researchers compared the eight related lists and duplicated problems were removed. This resulted in 3 lists of usability problems, severity ratings and group frequencies (related to the duplication across groups). The problems were then re-coded at this stage with a letter to represent the software, for example t for TRIADS® and HU to identify the fact it was from a heuristic evaluation and then a number, tHU1.
- Stage 3: Researcher Reduction: Problems with either a frequency of discovery of less than 50% (less than 50% of the groups members identified the problem) or which had a severity rating of below 3, were discarded. This resulted in 8 x 3 reduced lists of usability problems, severity ratings and frequencies of discovery (the percentage of students in the group that had found the problem).
- Stage 4: Damage Index: Following the reduction stage the remaining problems were coded using a new formula devised for prioritising the data sets from multiple evaluations, referred to as the Damage Index. This formula would enable problem sets from multiple evaluations to be reliability aggregated and remove potential bias in the prioritisation of the problem set, as this will be based on the index value that is generated. Using the notion that the damage to an individual user of the software as a result of usability problems would be both psychological and system related, a damage index was apportioned to each usability problem, see Figure 30.

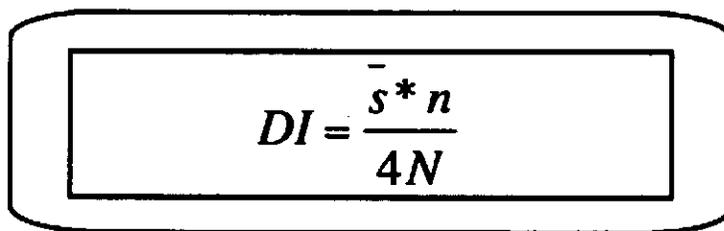

$$DI = \frac{\bar{s} * n}{4N}$$

Figure 30 Damage Index Formula

- Damage Index = DI
- Mean Severity Score = \bar{s}
- Number of groups that identified the problem = n
- Upper bound of severity rating scale = 4
- group size = N

The Damage Index would produce a ratio for each problem and the basis for the formula was problems with a high probability of being discovered and high severity rating are likely to cause the user the most difficulties. This formula if proven to be appropriate would fulfil the objective of developing a mechanism for aggregating problems described in Section 10.1.1. On reflection it would have been feasible to apply the Damage Index earlier in the aggregation process, however the decision was made to apply it at this stage as the data set would be more manageable.

- Stage 5: Type Allocation: Each problem was investigated to determine whether it was type A: Software specific or type B: Test design. The criteria used was a combination of determining whether the problem would exist if the test was paper based or if altering the text provided in the questions and feedback would alleviate the problem. For example, if a problem was in relation to the navigation then this would be classified a Type A, whilst if it related to an ambiguous question it would be Type B.

10.2.6 Recoding the Data

After the initial analysis the data from stage 2 was re-examined and using the same procedure as Chapter 5, each problem was coded with a task step code and unacceptable consequences scale. This would ensure that the coding of the problems within the corpus is consistent. After this stage the coding was modified, instead of using a small letter to represent the software a capital letter was used, for example THU1. Also this data set would be used to establish whether the damage index is effective at identifying problems with unacceptable consequences, and would be used to determine if problems were unique to a single application.

10.3 Results

The results for each stage of the analysis for each of the 3 applications are presented in Table 29 below:

			Stage 1	Stage 2	Stage 3
Software	Group	Students	Problems	Problems	Problems
WebCT	A	5	11		
	B	6	18		
	C	3	12		
	D	5	17		
	E	5	16		
	F	2	4		
	G	3	15		
	H	6	11		
	Total		104		
Questionmark	A	6	14		
	B	4	15		
	C	2	13		
	D	3	13		
	E	3	13		
	F	4	13		
	G	5	14		
	H	4	15		
	Total		110		
TRIADS	A	5	18		
	B	3	10		
	C	4	17		
	D	4	19		
	E	4	14		
	F	5	11		
	G	5	9		
	H	2	7		
	Total		105		

Table 29 Number of problems reported in each of the 3 applications

10.3.1 Stage 1 - Student Aggregation

Prior to the student aggregations the raw data revealed that each heuristic had problems classified to it as shown in Table 30. Unlike Chapters 8-9, there were no heuristics that had no problems classified to them.

Heuristic	WebCT	Qmark	Triads
1	36	35	37
2	29	24	21
3	45	26	49
4	37	39	39
5	36	46	28
6	12	12	17
7	5	15	37
8	16	16	41
9	25	14	17
10	15	15	11

Table 30 Problems classified to each of the heuristics at stage 0

Once the individual sheets had been collected, the evaluators were formed into eight different groups per software application and the individual results were aggregated. The total number of problems found in each application was calculated by simply adding the total of each group (this may have included duplicates at this stage) Questionmark® had a total of 110 problems, TRIADS® had 105 and WebCT® had 104.

There were a similar number of problems reported within all of the environments, see Table 31.

Software	Mean	Standard Deviation
Questionmark	13.75	0.89
TRIADS	13.12	4.51
WebCT	13	4.43

Table 31 The mean number of problems identified per application

10.3.2 Stage 2 - Duplication Treatment

The list of raw data from each group was analysed by the two researchers to remove any duplicate problems, Questionmark® had a total of 33 problems reported in more than one group leaving a total of 77, TRIADS® had 31 duplicates with 74 problems remaining, and WebCT® had 26 duplicates with 77 problems remaining.

10.3.3 Stage 3 - Researcher Reduction

The researchers discarded problems with either a low frequency (less than 50% within the group) or severity rating of less than 3. For example if 40% of people within the group identified the problem and attached a severity rating of 3 this

would be retained but if it had a rating of 2 it would be discarded. This resulted in 42 problems being discarded in Questionmark, 35 in TRIADS and 40 in WebCT. This left a total of 35 problems for Questionmark, 39 for TRIADS and 37 for WebCT.

The reporting format used hindered this stage to some degree, as the aggregated list only reported the problem, frequency and severity rating unlike the form used for the evaluation, which had details about where the problem occurred. In some instances it was difficult for the two researchers to determine if the problems reported were identical due to the different levels of abstraction. It would have been ideal if a member from each of the groups aggregated the list, as they would have been more familiar with the problem set, but this was not feasible.

10.3.4 Stage 4 – Applying the Damage Index

Based on the Damage Index (DI) the problems with the highest 5 ratios for Questionmark® are listed below:

- qHU9 - You could finish the test and submit your answers even if some questions hadn't been attempted - should have prompted you, DI= 0.56
- qHU2 - Q7 onwards does not specify how many boxes to tick, DI= 0.56
- qHU3 - q13-18 required perfect character entries or would be marked wrong, DI= 0.53
- qHU36 - Some questions are worth more marks, DI=0.50
- qHU18 - 13-18 the input boxes had a drop down menu arrow, confusing the user, DI=0.41

There were two problems with a rating of 0.56 and both could have a negative impact on test performance by affecting the results. One of the problems could easily be rectified by specifying how many correct answers there were within the multiple response question stem. Also including the scoring algorithm (marking scheme) within the question text would help alleviate the fact that students were confused with how many marks were associated with each question.

The highest rated problems for TRIADS are listed below:

- tHU6 - When going back to a previous question the question and answer was not displayed, DI=0.56

Chapter 10

- tHU10 - When going back to a previous question if had been wrong you could change it, DI=0.50
- tHU11 - The program took too long to move from one question to another, DI=0.50
- tHU7 - The continue and submit buttons were different throughout the test, DI=0.44
- tHU18 - When getting a result it should be complete or not at all, DI=0.34

Similar to some of the problems within Questionmark® the last problem could easily have been avoided with better test design, ensuring the feedback was appropriate and accurate for each question. The other problems were characteristics of the software which could be addressed through amendments to the program. However, the fact that there were long delays moving between questions may be a concern if used for high stakes exams under timed conditions.

Finally the highest rated problems within the WebCT® environment are rated below:

- wHU1 - Question does not specify how many questions to choose in a multi choice question, DI=0.59
- wHU15 - Does not say it is negatively marked, DI=0.37
- wHU4 - If a small spelling error is made on the end questions a wrong answer is given, DI=0.34
- wHU46 - You can save without actually answering any questions no error message user assume correct, DI=0.34
- wHU10 - Feedback not clear, DI=0.31

It is apparent from examining these results that three of the problems could have been avoided through better test design. The remaining problems relate to validation within the system and may be more difficult to address. When using text entry style questions, the software is unlikely to be able to recognise every permutation of the answer and there would need to be manual checks especially in the case of students with dyslexia or other learning difficulties.

Overall it is evident that a number of the highest rated usability problems within the three environments were attributed to poor test design and could have been easily avoided. By addressing these issues through peer moderation, the test design could be improved, resulting in an overall lower damage index score and improvement in user satisfaction. However, improvement in the interface design and validation of user actions would improve all three environments significantly.

10.3.5 Stage 5 - Type Allocation

The problems were also classified as either test design or a characteristic of the software used to deliver the assessment. Analysis of the data after stage 3, Questionmark® had 10 problems that were classified as being attributed to the test design, TRIADS® had 5 and WebCT® had 10. This difference may have been as a consequence of the analysis method applied, as a number of problems had been discarded due to low frequency or severity. It was anticipated that this would be the same in each environment as the same questions, scoring algorithm and feedback was provided.

10.3.6 Unique Problems

It is claimed that heuristic evaluation can be performed by 3-5 evaluators examining a system and Nielsen and Landauer (1993) suggest the percentage of known usability problems that an evaluator is likely to find based on a lambda value of λ will be 31%. However, after aggregating the results, this study revealed that every group found some unique problems, therefore adding additional evaluators would probably reveal more unique problems, see Table 32.

Software	Group	Problems	Unique problems
WebCT	A	11	3
	B	18	1
	C	12	3
	D	17	6
	E	16	2
	F	4	1
	g	15	5
	H	11	3
	Total	104	24
	Questionmark	A	14
B		15	5
C		13	3
D		13	2
E		13	2
F		13	3
g		14	1
H		15	2
Total		110	19
TRIADS		A	18
	B	10	2
	C	17	3
	D	19	4
	E	14	3
	F	11	3
	g	9	2
	H	7	3
	Total	105	24

Table 32 Number of unique problems reported by group and application

It is questionable whether between 3 and 5 people is sufficient for conducting a heuristic evaluation and this has been found in other studies (Sim, Read *et al.*, 2006b; Woolrych & Cockton, 2001). For example, only one person in one group identified that pressing the back button within the browser would terminate the exam within Questionmark®. This is a severe problem and may not have been identified if only a small sample conducted the evaluation.

10.4 Exploring the use of the Damage Index and Unacceptable Consequences Scale

Using the data from stage 3, it was evident that there was some variability between the severity ratings attached to problems within the three environments, see Table

33. Questionmark® had more severe problems with 40% being classified with a rating of 4, (usability catastrophe) compared to TRIADS® and WebCT®, which were both around 25%.

	Severity Ratings			
Software	4	3	2	1
Questionmark	14	20	1	0
TRIADS	10	23	3	3
WebCT	9	19	8	1

Table 33 The number of problems classified to each of the severity ratings

As the research was primarily concerned with major problems within the environments, it was necessary to analyse the remaining problem set using the Damage Index.

Using this formula a calculation was made for each of the CAA environments to establish the environment with the most frequently predicted severe usability problems, the data used to perform the calculation was after stage 3 of the analysis and the results are presented in Table 34. An overall index value was produced along with a value for the problems specific to the software and the test design.

	Overall		Software		Test Design	
Software	Mean	Standard Deviation	Mean	Standard Deviation	Mean	Standard Deviation
Questionmark	0.201	0.146	0.188	0.131	0.234	0.180
TRIADS	0.169	0.133	0.167	0.138	0.181	0.102
WebCT	0.158	0.119	0.137	0.085	0.216	0.175

Table 34 Results of the damage index applied to the three environments.

Based on this formula Questionmark® is predicted to have the most severe usability problems, followed by TRIADS® and WebCT®. However, an ANOVA test using the Damage Index revealed there was no significant difference between the three CAA environments $F(108,2)=1.012$, $p=0.367$ suggesting there is no unequivocal difference in the predictions of severe usability problems between applications. As the problems categorised to test design were not characteristics of the software, a further ANOVA test was performed purely on the damage index value of the application with the test design problems removed. Again there was no significant difference between the three applications $F(83,2)=1.162$, $p=0.318$ based on the

Damage Index value. The heuristic evaluations would seem to suggest that all three of the environments have a similar level of usability problems associated with them.

10.4.1 Problems Classified with Tasks and Consequences

To ensure the data was in the same format as the other corpus items the problems were coded with the tasks step code and consequences scale. The data used for this coding exercise was the data after stage 2 duplication treatment stage. Further analysis and coding of the problem sets was performed by the same researcher and lecturer as in previous chapters. This process lead to a further reduction in size of the problem sets as additional problems were merged this is shown in Table 35.

Software	Duplicate Treatment	Re-Coded
Questionmark	77	41
TRIADS	74	51
WebCT	77	44

Table 35 Problems recoded with task and consequence

10.4.1.1 Problems with Unacceptable Consequences

Table 36 shows the number of problems in each of the applications that have been rated to the consequences scale discussed in Chapter 4, the data can be found in appendix L - N.

Software	Certain	Probable	Possible	Dissatisfied
Questionmark	1	3	26	20
TRIADS	2	4	26	26
WebCT	4	3	20	23

Table 36 Problems rating to the unacceptable consequences scale

It would appear that many of the problems identified would just lead to the user being dissatisfied rather than any unacceptable consequences. This scale contradicts the results in Section 9.3.5, were 97.1% of Questionmark® problems were rated by the students at greater than 3. For example an evaluator reported *that some questions were worth more marks* and gave this a severity rating of 4, however the student would not have any grounds for appeal and therefore this was coded as dissatisfied. Therefore the accuracy of the severity ratings may be questionable and an alternative approach could be using the Damage Index but it is unclear if the Damage Index identifies problems with unacceptable consequences.

10.4.1.2 Unacceptable Consequences vs. Damage Index

A Damage Index was proposed in Section 10.2.5, and this was applied to the data in order to filter out the problems which have the potential for causing the user the greatest difficulty. Table 37 shows the top 3 problems for each application based on the Damage Index along with their unacceptable consequences rating.

Problem	Software	DI	UC
You could finish the test and submit your answers even if some questions hadn't been attempted – should have prompted you,	Q	0.56	Pos
Q7 onwards does not specify how many boxes to tick,	Q	0.56	Pos
q13-18 required perfect character entries or would be marked wrong,	Q	0.53	Pos
When going back to a previous question the question and answer was not displayed,	T	0.56	Pos
When going back to a previous question if had been wrong you could change it,	T	0.5	Dissat
The program took too long to move from one question to another,	T	0.5	Pos
Question does not specify how many questions to choose in a multi choice question,	W	0.59	Pos
Does not say it is negatively marked,	W	0.38	Prob
If a small spelling error is made on the end questions a wrong answer is given,	W	0.34	Cert

Table 37 Damage index compared to consequences classification

The results show that the majority of problems with the highest Damage Index score are not necessarily the problems that would lead to certain or probable unacceptable consequences. There was 1 problem that was rated as dissatisfied which raises concerns over the effectiveness of the Damage Index in extracting the most severe problems however it does remove bias from the ranking process and for this reason is judged to be an effective tool. The problem *When going back to a previous question if had been wrong you could change it* was rated as dissatisfied because it does not have any real consequences, the user could change their answer but it would not alter their score.

Despite suggested limitations, the process of filtering the problems and attaching a Damage Index appears to identify large number of problems which would be classified as true negatives as they have been discarded from the problem set. Table 38 displays the WebCT® data after the merging and applying the consequences

scale cross tabulated with the data after stage 3. If a problem was discarded in stage 3 it would not have a Damage Index score and was therefore coded with a -1.

Consequence Scale		Damage Index									
		-1	0.06	0.09	0.13	0.22	0.28	0.34	0.38	0.59	Total
Diss	0	13	1	6	3		1				24
Poss	1	7	1	1					1	1	11
Prob	2			1			1	1	1		4
Cert	3				3	1		1			5
Total		20	2	8	6	1	2	2	2	1	44

Table 38 Cross tabulation of the Consequences Scale and Damage Index for WebCT

The data shows that the 20 problems that were discarded in stage 3 would also be discarded using the filtering process based on unacceptable consequences, suggesting that this approach may be effective at reducing the corpus of problems in usability evaluations. Similar results are also found with the other software application see Appendix O.

10.4.1.3 Problems Found in all Three CAA Environments

Using the merged data from Section 10.4.1, Figure 31 shows the number of problems that were found and there distribution between the 3 applications.

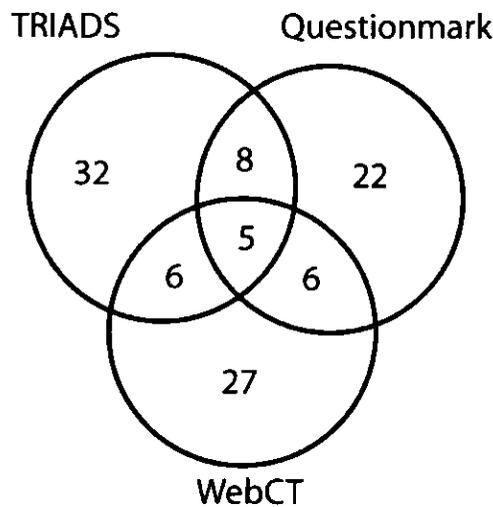


Figure 31 Problems found within each application

It is clear from Figure 29 that many of the problems are unique to a CAA application. There were only 5 problems that appeared in all 3 CAA applications and these are presented in Table 39 below:

Code	Problem	User Task	Consequences
WHU1 THU8 QHU2	Question does not specify how many questions to choose in a multi choice question	D1.2	Poss
WHU3 THU13 QHU3	If a small spelling error is made on the end questions a wrong answer is given	D1.1	Cert
WHU4 THU9 QHU15	Question 17 uses incorrect grammar and makes no sense	D1.1	Poss
WHU16 THU45 QHU9	At end of test is no question answered and clicked finish it brings up a box with ok or cancel	E3	Poss
WHU27 THU23 QHU10	When saved a question you could go back and edit it anyway	D	Dissat

Table 39 Problems that appear in all 3 applications

Four of the five problems that are predicted in each of the three applications would lead to unacceptable consequences. The evaluators reported *When you saved a question you could go back and edit it anyway* as a problem, they also reported that you could change your answer after it had been marked (this would not alter your score), however in most tests situations you would have the opportunity to alter an answer during the test, therefore it was difficult to determine why this was a real problem that would have any unacceptable consequences.

It was anticipated that the evaluators would have identified the test design issues within all three environments as the questions, scoring and feedback were the same throughout but this was not the case. For example the evaluators using WebCT® and Questionmark® identified the fact that it does not say an answer is negatively marked and this was not picked up within TRIADS®.

10.5 Conclusions

The primary objective of this study was to expand the corpus of usability problems and this has been achieved. The study reported here has shown that heuristic evaluations can be used to predict usability problems in three CAA environments using novice evaluators. A large number of problems were initially recorded, over 100 in each of the three environments. This study was mainly concerned with the

most severe and many of the reported problems were duplicates, after a filtering process only a subset of problems were added to the corpus.

Another objective was *To develop a mechanism for filtering and aggregating Usability problems from multiple evaluations*. A filtering process was devised that incorporated several stages and a new formula, the Damage Index, was proposed. The Damage Index allows for problems from multiple evaluations to be prioritised in a repeatable manner thus removing the subjectivity and potential bias which may occur if an individual is performing the task. Law and Hvannberg (2008) suggested that consolidating usability problems is an integral part of usability evaluation but little is known about the mechanisms of this process. The results presented throughout this chapter in relation to the Damage Index, contribute to the knowledge on the process for aggregating usability problems from multiple evaluations. In Section 10.4.1.2 the method was analysed and compared to the consequences scale used in previous chapters and the process appeared to be effective at identifying the most severe problems. Many of the problems that were classified as Dissatisfied were discarded using the filtering process, suggesting that true negatives (not real problems) are being discarded from the problem set. Further work is still required to establish the effectiveness of the Damage Index, however it has the potential for prioritising the problem set (by removing true negatives) if used as part of a development lifecycle to fix known issues as in the RITE method (Medlock *et al.*, 2002). All three environments had a similar level of usability problems based on the Damage Index applied to the reported problems. There was no significant difference between the three environments based on the Damage Index. All three software applications had usability problems that could potentially affect test performance leading to unacceptable consequences.

It is possible to classify the problems based on two criteria, test design and software characteristics. The problems associated with test design could easily be rectified through peer moderation and training in the construction of objective tests. This would reduce the number of problems within the environment and could improve user satisfaction.

There are only 5 problems that appear in all three environments, most are unique to one individual application. Many of the problems stem from inconsistent layout, poor validation and error messages. However, it is possible to amend the templates

of many CAA environments therefore many of the issues identified in this study could potentially be avoided through careful interface design. One commonly found problem that may be more difficult to fix is the error caused by poor spelling. A potential solution is that the marking is not automated for text entry questions, or a second solution is to reduce the number of text input questions. Additionally to facilitate automated accurate marking of free text, one option would be to put a spell checker on the text entry box or enter multiple spelling alternatives in the possible answers.

The final objective was *to establish what effect increasing the number of evaluators may have on new problems found*. In this instance it opens the debate over the number of evaluators required. Some of the major usability problems such as browser buttons terminating the exam within Questionmark® may not have been reported with such a small number of evaluators. It may be that heuristics using between 3 and 5 evaluators should be used in conjunction with other methodologies to demonstrate the existence of the problems and their severity.

Although there was some overlap between problems reported in the 3 applications many of the problems reported were unique to the individual application. Therefore it would not be possible to suggest that the corpus can be generalised beyond the 3 applications (TRIADS®, Questionmark® & WebCT®). In order to use the corpus to synthesise a set of domain specific heuristics, it would need to be expanded to incorporate additional applications and the Damage Index can be used to prioritise the corpus.

10.5.1 Methodological Limitations

A finding from the process of analysing the aggregated results is that it was not possible, due to the data capture form to establish which problems were aggregated together. For example, a single problem may have been reported by three evaluators with different levels of abstraction. Once this problem had been merged in some instances it was not possible to identify the 3 instances on the individual evaluators' sheets. It would have been sensible for evaluators to code each of the problems and use this identification code on the aggregated list to ensure cross referencing could be performed. Due to the reporting format, this caused problems in identifying the problem and the location within the application. This could have an impact on the

ability to accurately calculate the Damage Index. If the Damage Index is used in a single evaluation to prioritise the problem set then the two parameters of mean severity rating and number of evaluators who identified it as a problem is essential. Without the ability to accurately cross reference the data from the evaluators' sheets with the merged data set, calculating the mean severity rating is problematic and therefore it is recommended that the form be altered to alleviate this issue.

Due to the large number of problems identified, managing the data was problematic and the data entry had to be double checked to minimise data input errors. Despite this it was later discovered that one of the problems in the TRIADS® application tHU8 had accidentally been removed from the corpus after stage 3 of the data analysis. Due to the data being analysed using two different approaches, it was later added to the corpus when the data set was reanalysed using the consequences scale, therefore the problem was retained in the corpus to aid the development of the heuristics.

10.5.2 Research Questions

Having expanded the corpus to include problems from three CAA applications the next stage of the research is:

- To expand the corpus to include additional applications.
- To synthesise domain specific heuristics for CAA from the corpus, using the Damage Index as the basis of an Evidence Based Design approach.

Chapter 11 Synthesis of Heuristics for CAA

11.1 Introduction

The objective of this chapter is to further expand the corpus and then synthesise a set of evidence based domain specific heuristics. An evidence based design approach for the synthesis of domain specific heuristics is proposed in Chapter 7. The evidence based design approach will diminish the impact of the evaluator effect reported in previous chapters by incorporating data from multiple evaluations. The Damage Index will enable the prioritisation of the corpus of usability problems. The Damage Index diminishes the aggregation effect as it is quantifiable, repeatable and thus removes subjectivity and bias from the aggregation process. Limitations of the approach are analysed and suggestions for how the method can be improved are proposed. The work in this chapter has been published at the INTERACT conference (Sim *et al.*, 2009).

11.1.1 Objectives

As stated above, the main objective was to expand the corpus and to apply an evidence based design approach to synthesise heuristics for the CAA domain. Other objectives are:

1. *To establish the limitations of the applied methodology.*

By evaluating and critiquing the approach taken in this thesis, limitations and improvements to the methodology will be proposed that can assist future development of heuristic sets.

2. *Identify future research regarding heuristic development.*

By synthesising the heuristics, it is envisaged that additional research questions will emerge.

11.1.2 Scope

This study was designed to synthesise domain specific heuristics for CAA by using an evidence based design approach. The study was constrained by the number of researchers and domain experts available to participate in the card sorting exercise and assist in the synthesis of the newly created heuristic set.

11.1.3 Contributions

The main contributions are:

1. A set of evidence based domain specific heuristics for CAA.
2. An evidence based design approach for the creation of heuristics.

11.1.4 Structure

This chapter is structured in the following way: In Section 11.2 the expansion of the corpus is discussed, the user studies are re-analysed in Section 11.3, the heuristic evaluations are further examined in Section 11.4, expanding the corpus through re-examining the literature is presented in Section 11.5 with the merging of the data sets reported in Section 11.6. In Section 11.7, the synthesis of the domain specific heuristics is presented. The discussion and conclusions are presented with a summary of the findings in Section 11.8, along with, limitations and further research.

11.2 Expanding the Corpus

The previous studies established that many of the usability problems identified are unique to an individual CAA application, and the corpus would need to be expanded to cover problems with other applications. There are a number of possible methods that could be applied, these are listed below:

- Conduct usability evaluations with other applications
- Consult other institutions using CAA applications to see if they would be prepared to perform usability evaluations of the applications
- Analyse the published literature to expand the corpus based on reported usability problems found within the literature.

The first two options would be extremely difficult from a practical basis to apply. For example, the university where this research took place does not have access to other applications. There would be financial implications for acquiring other applications, therefore this option was not feasible. Similarly other universities may be reluctant to perform usability evaluations without any incentive as it is a time consuming process so this option was also dismissed.

Although the literature review in Chapter 3 established that there was little research in the usability of CAA applications, this was judged to be the most feasible approach to expanding the corpus. The literature was re-examined to identify reported usability problems and the process is discussed in Section 10.5 below.

The evidence based design approach presented in Chapter 7 had three stages starting with user studies, see Figure 32.

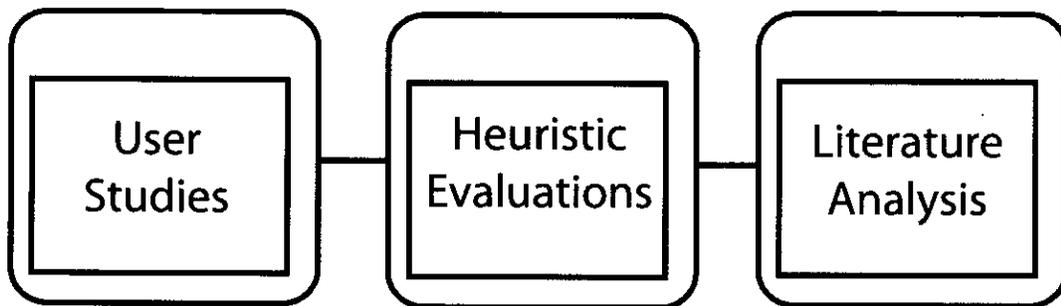


Figure 32 Three stages of investigation used to develop the corpus

11.3 User Studies

In stage 1, user problems were collected using self reporting questionnaires (Chapters 5-6) to establish whether problems reported were usability related and whether these would lead to unacceptable consequences. The results of these studies highlighted that there were severe problems evident in both WebCT® and Questionmark®. A total of 22 usability problems were reported, these problems have been used to assess the effectiveness of Niensens' heuristics in Chapter 8. However, of these 22 problems 6 were judged to have no unacceptable consequences, and for the purpose of synthesising the heuristic were discarded. The removed problems were judged to lead to the individual being dissatisfied with an element of the application or test, but it would not hinder their performance. For

example *Clicked finish and you get a blank screen, needs to say thanks* was removed, this would have no unacceptable consequences as the data would have been saved and the students results would not have been affected. In total 16 problems were thus carried forward from the usability studies.

11.4 Existing Heuristics

Nielsen's heuristics have been applied to a wide variety of domains, including hypermedia browsers (Connell & Hammond, 1999), edutainment applications (Embi & Hussain, 2005) and to improve the hardware of musical products (Fernandes & Holmes, 2002). This wide application suggested that the effectiveness of Nielsen's heuristics was worth evaluating in the context of CAA. The effectiveness was evaluated in Chapter 8, the results suggest that they are relatively ineffective. Despite the ineffectiveness of these heuristics, some problems were found and the data from this study was used to expand the corpus.

Despite the limitations, but without a suitable alternative, a series of further heuristic evaluations (Chapters 9-10) were conducted using Nielsen's heuristics with the sole purpose of expanding the problem corpus. These studies covered three different CAA environments in order to discover as many problems as possible ensuring that problems unique to one or more systems could be found. In total over 300 usability problems were reported within the three environments including several with unacceptable consequences; for example: *Using back button in browser exits the test rather than returning to previous question* and *The programme took too long to move from one question to another*, both of these were judged to have unacceptable consequences. Table 40 below shows the number of problems reported in each of the applications from the studies reported in Chapters 8-10, and also shows the final number after the data sets had been merged and filtered.

	Chapter 8	Chapter 9	Chapter 10		
	Questionmark	Questionmark	Questionmark	TRIADS	WebCT
Total	47	52	110	105	104
Merged	33	44	41	51	44
Filtered	16	20	20	26	23
Problems	17	24	21	25	21

Table 40 Problems remaining after filtering process

Chapter 11

In this table, the figures aligned to ‘total’ represent the raw number of problems reported, ‘merged’ is the number after the removal of any duplicates, ‘Filtered’ is the number of least severe problems removed - problems coded as ‘dissatisfied’ and the final code ‘problems’ is the remaining number of usability problems. At this stage some of the problems are likely to be duplicated between studies and these are merged in Section 11.6.

The difference in the numbers from ‘merged’ to ‘problems’ would suggest that about 50% of problems reported in the heuristic evaluations would not lead to any unacceptable consequences.

The studies dealt with the following known factors that might affect the quality of the corpus:

- Question styles – Chapters 8, 9 & 10
- Evaluator effect – Chapter 9
- Context Summative of formative – Chapters 8 & 9
- Cohorts – Chapter 6 (user studies)
- Software applications – Chapters 8-11

Despite the depth of the resulting corpus it is not feasible to synthesise a set of heuristics and generalise about the appropriateness for evaluating CAA applications because of the limited number of applications and question styles evaluated. Therefore the next stage was to re-consult the literature to further expand the corpus.

11.5 Literature Review

A literature review of CAA was conducted in Chapter 2, and was published in ALT-J (Sim, Holifield *et al.*, 2004). However, to establish any additional usability problems that may not have been uncovered within the user studies and heuristic stages, more recent publications were reviewed to improve corpus quality. There have been two literature reviews published in the area of CAA (Conole and Warburton, 2005; Sim *et al.* 2004) and these were used as the initial focal point along with searches in digital libraries such as Ingenta, ACM, AACE and analysis of the conference proceedings of the International CAA Conference. Although there is very limited research specifically focusing on usability and CAA the review

revealed a body of evidence from studies that suggested usability problems existed in applications. The review also revealed a number of published guidelines (BS7988, 2002; ISO/IEC23988, 2007) for the implementation of CAA and these were also consulted.

11.5.1 Data Set from Literature Review

The objectives when analysing the literature were to identify reported problems from other applications and questions styles. If problems were reported in applications already evaluated in Chapters 5-10 these would also be recorded to provide additional evidence that the problem was a true positive (real problem) enhancing the quality of the corpus.

A total of 22 publications were identified that offered some evidence of usability problems within CAA environments. When a problem had been identified from the literature it was recorded in a spreadsheet using the same procedure as the previous chapters. The problems identified from the literature are displayed in Appendix P. There were a total of 24 problems or guidelines identified relating to CAA covering a number of additional applications, for example TOIA and V32. However 5 of these problems or guidelines were classified as dissatisfied, leaving 19 problems that would have unacceptable consequences. One of the problems was *cases have been reported of direct copying from external sites*, this was judged not to be an unacceptable consequence but an important issue to prevent plagiarism but this did not align to the unacceptable consequences scale. The reason this was not judged to be an unacceptable consequences is because the student committing the offence would not have grounds for appeal, neither would the other students taking the test, their test performance may not be affect however it would be more of a concern to the academic.

11.6 Merger of the Data Sets

At this stage there was data from 3 sources, the user studies, the heuristic evaluations and the literature review. Figure 33 represents the process of coding and analysing the data from the various studies.

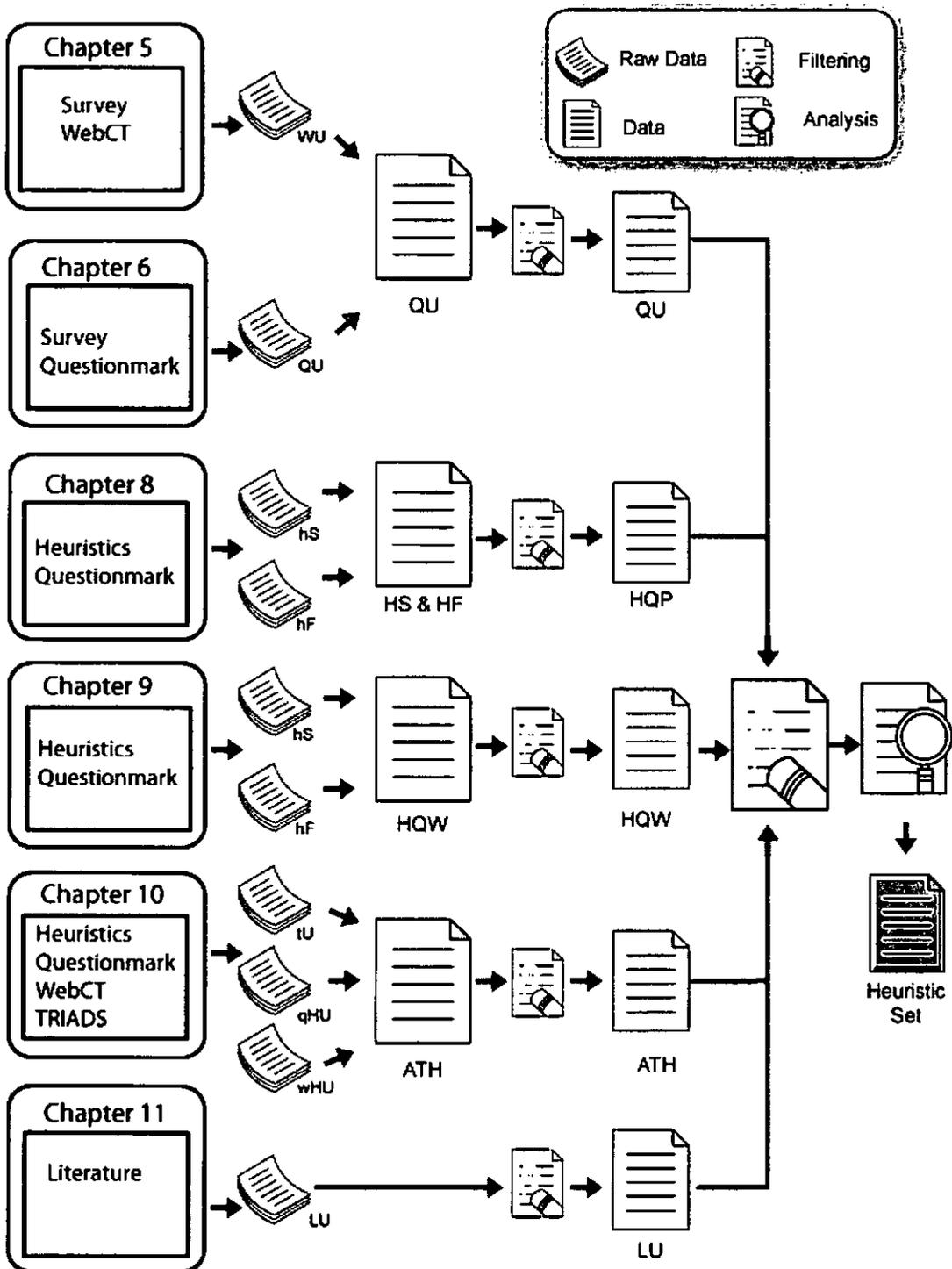


Figure 33 Triangulation of data from the various studies

For example the data sets from the three heuristic evaluations were merged into a single list and recoded with ATH (all three heuristics) to indicate the problems came from this study. Following this initial analysis stage, the data was merged using a two phased approach, utilising the expertise of an educational technologist, HCI lecturers and research students. The first phase involved the researcher and

educational technologist merging the data based on the task step code and consequences this would enable a Damage Index score to be calculated. This score would enable the problem set to be prioritised for the synthesis of the heuristics.

11.6.1 Merging – Phase One

The 67 problems from the 3 applications in Chapter 10 were aggregated into a single list to ensure that matched problems were allocated to the same task step code. When merged problems had a different task step code and consequence scale a decision was made to determine the final classification by examining the problems collectively then agreeing a final code. This resulted in an aggregated list of 47 problems for the 3 applications evaluated in Chapter 10 with unacceptable consequences, see appendix Q.

In order to merge the problem sets two steps were taken, the first, problems were grouped together based on the user task step code. For example all problems with a code of D are grouped together, this is represented in Figure 34. Following this the problem set was analysed between task step codes and problems were merged again.

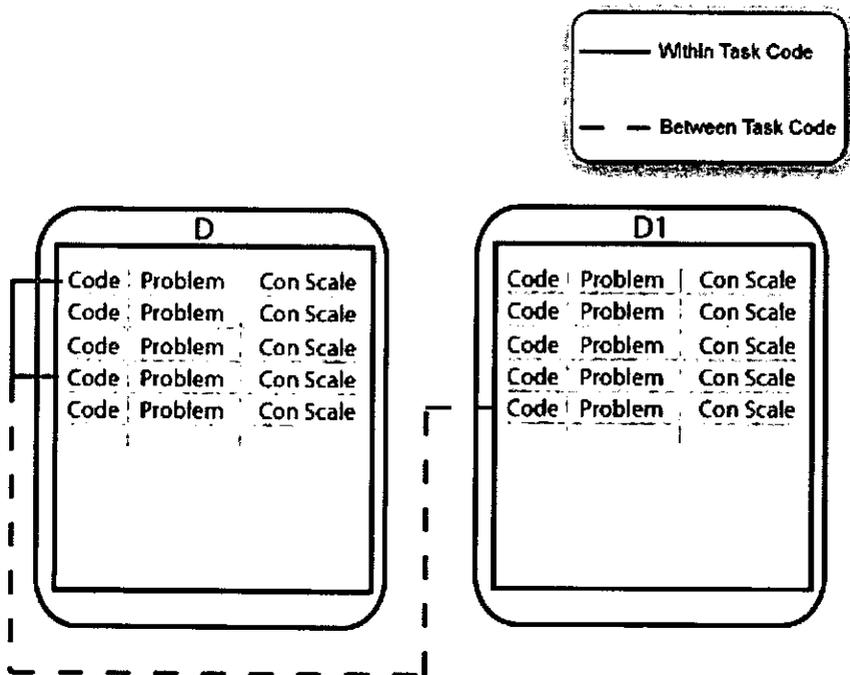


Figure 34 Merging of data within and between task step codes

Table 41 below shows the total number of problems reported from all the data sets and their corresponding user tasks step code after the within analysis.

It is clear that the majority of the problems with unacceptable consequences occur during the test and therefore the majority of the heuristics should deal with these issues. It would not be sensible to have several heuristics dealing with starting the test as there are only 2 problems reported and in designing domain specific heuristics Paddison and Englefield (2004) suggest limiting the number of heuristics in order to not overwhelm the evaluators.

User Task Step Code	Within Task Problems	Problems Remaining
Test (T)	11	7
Start Test (S)	2	1
Access (S1)	0	0
Login (S2)	0	0
Select Test (S3)	0	0
During Test (D)	29	10
Answer Question (D1)	9	5
Understand how to answer (D1.1)	11	4
Construct an answer (D1.2)	16	7
Confirm answer (D1.3)	8	4
Review / edit question (D2)	6	3
Navigate through questions (D3)	10	6
Feedback (D4)	2	2
End (E)	3	3
Awareness of finish (E1)	1	1
Check answers (E2)	1	1
Submit answer (E3)	7	2
Feedback (E4)	4	4

Table 41 Number of problems mapped to the task step code and merged within

It became apparent when merging the problems within the task step codes that some problems appear under more than one code. For example

- Task Step Code E3 - Accidentally finished the test
- Task Step Code D - Lost exam answers message came up 'page has expired'

Therefore to minimise overlap between task step codes a second stage of merging took place. The researcher and educational technologist examined each problem to determine whether it existed in another task step and merged any duplicates, this was necessary before a set of heuristics could be synthesised from the corpus to ensure appropriate coverage. The final number of problems attributed to each task step is displayed in Table 42.

User Task Code	Final Problem Groupings
Test (T)	7
Start Test (S)	1
Access (S1)	0
Login (S2)	0
Select Test (S3)	0
During Test (D)	8
Answer Question (D1)	2
Understand how to answer (D1.1)	1
Construct an answer (D1.2)	3
Confirm answer (D1.3)	1
Review / edit question (D2)	2
Navigate through questions (D3)	3
Feedback (D4)	1
End (E)	1
Awareness of finish (E1)	0
Check answers (E2)	1
Submit answer (E3)	1
Feedback (E4)	3

Table 42 Problems remaining after between task step code analysis

Through a series of systematic mergers the problem set has been refined, leaving a small number of problems from which a set of heuristics will be synthesised. The problems were allocated a total number to indicate how many instances had been reported through the various studies and the unacceptable consequence scale attributed to each instance of the problem.

To determine the Damage Index, the mean severity rating score was calculated based on the unacceptable consequences scale ranging from 0 for Dissatisfied and 3 for Certain. Therefore the formula needed to be modified, instead of having $4N$ this was altered to yN to represent the Consequences Scale.

$$DI = \frac{\bar{s} * n}{yN}$$

Figure 35 Revised Damage Index Formula

In the original formula containing the parameter 4 this represented the upper bounds of Nielsen's severity rating scale and therefore could only be applied to data that

utilised this scale and by altering this parameter to y the formula can be adapted across multiple domains / disciplines.

It became evident that some of the problems had not been adequately merged in the studies, for example ATH6, ATH22 and ATH32 were judged to have been reporting the same issue and therefore at this stage a mean severity score was calculated prior to applying the Damage Index. The decision was made to have 5 groups as the data sets from each of the individual studies had been merged at this stage, shown in Figure 30, for example the user studies had been merged and are reported in Appendix C. The 5 groups used in the formula were, 1 group for the user studies, 3 groups for each of the heuristic evaluations and the final group the literature review. This data would be used to highlight the common problems and prioritise the corpus when forming the heuristics. An example of how the problems were merged for task step code T is presented in Table 43 below. The 'Merged' column shows problems merged within a single task step code, in this instance T, and 'Between Tasks' displays the problems merged from different tasks, for example ATH6 has been merged with problem HQW2 from task step code D. 'Total' represents the number of groups that identified the problems.

Code	Problem	Merged	Between Tasks	Total	D.I.
QU6	Don't trust the computer think something will go wrong + not process			1	0.07
ATH6	Help feature is not very helpful	ATH22 ATH32 HQP19	D(HQW2)	3	0.20
ATH9	Does not say it is negatively marked		D1.1(QU15), D1(HQW4)	4	0.33
ATH18	text on screen is not re-sizeable could be hard for user to read		D1(QU3), D1.1(ATH41)	4	0.30
ATH22	There wasn't any help information throughout the test				
ATH26	Unable to scroll down in firefox incompatible	LU22		2	0.27
ATH32	No explanation on the functions of the button				
HQP19	No help				
LU10	In the event of the test crashing users should be able to resume the test from the point where they left off (Jacobsen & Kremer, 2000; Stephens <i>et al.</i> , 1998).		D(QU13),E(A TH34), E3(QU8)	5	0.92
LU20	Consideration should be given to any potential time delays. For example, every time a student answers a question there is a connection to the server, this takes time and loading image files could also impact on latency (Ashton <i>et al.</i> , 2003).		D3(ATH39, E(ATH44)	3	0.20
LU22	There is a large array of different browsers available and compatibility needs to be ensured. (Pain & Le Heron, 2003) found issues with WebCT working in Netscape and IE in some computer laboratories. If multimedia elements are used such as Flash ensure that plugins are installed (Herd & Clark, 2002; Sim <i>et al.</i> , 2005).				

Table 43 Example of the merging process for the task step code T

When the problems were merged a description was synthesised which encompassed the problem by the researcher and educational technologist, see Appendix R for the final problem set and descriptions. LU10 was the only problem that was identified in all the studies and was therefore identified as the most severe problem based on the Damage Index.

11.6.2 Merging – Phase Two

The final 34 problems that remained after phase one were then further analysed by two lecturers in HCI, two research students and the researcher. The researcher emailed each person the remaining corpus which contained the unique code, the description and the total number of instances within the various evaluations. Individually they were asked to merge any problems which they thought were similar or related. Following this they had to create a theme that the problems could be categorised too, with a maximum of 12 themes, these would then be used to help synthesise a set of heuristics. The number of themes was limited to 12 in order to try and limit the number of heuristics.

11.6.2.1 Card Sorting

A week after the problems were emailed to the individuals a card sorting exercise was arranged. The five individuals brought their merged problem sets and themes to the meeting. Everyone was briefed by the researcher about the context the heuristics are to be used in and the process that will be used for the card sorting exercise.

Initially the problem corpus was going to be merged as a group based on the 34 problems and then themes collectively synthesised, however it became quickly apparent that each individual had approach the task from a slightly different perspective. For example, one of the individuals had used Nielsen's heuristics as a basis for his/her themes, another person based his/her themes on prior experience as a web developer, this meant that each problem was merged differently and no agreement could easily be reached. Therefore an alternative approach was adopted using the individual themes instead of the individual problems. Each problem had originally been classified to one of the themes therefore the merging of the themes would ensure that coverage of the corpus is retained.

Each person was given a unique coloured paper to write down all their themes on and these were then grouped together. Initially 10 themes emerged after discussion

with a further 8 themes remaining, these were later merged after further analysis and debate. The process of grouping the themes took over two hours as there was a great deal of debate and themes emerged and altered. For example, it was debated whether accessing the test is the same as accessibility issues, the final decision was that other themes would encapsulate accessibility and accessing the test would remain as an individual theme. Therefore it was decided to remove the two instances of accessibility from the card sorting exercise as these would be encapsulated into other themes. Other areas of debate centred upon interface design and whether navigational issues should be incorporated, it was decided that they should remain separate in an attempt to ensure these elements are not overlooked. The data was merged into a total of 12 themes and are shown in Table 44.

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
TH6. Protecting Answers	Saving Issues
TH7. Access	Access Accessing Test
TH8. Test Design	Unclear Information in test Teacher Issues Test related Tutor
TH9. Psychological / Perception	Comparability with paper Trust Stupidity Perception
TH10. Physical	Online Issues Hardware x2
TH11. System Feedback	Provide Help X2 Feedback for actions Feedback x2 Confirm all actions Inadequate information for users Feedback and support

Table 44 Final themes merged from groups individual themes

11.7 Synthesis of Heuristics for CAA

The final themes were then used to start the process of synthesising a set of heuristics for CAA. In Chapters 8 and 9 it was highlighted that some of Nielsen's heuristics were possibly redundant within the CAA domain, however some seemed appropriate. The researcher and an educational technologist then re-examined Nielsen's heuristic set and the themes that had emerged in Section 11.6.2.1 to compare and contrast. The purpose was to aid the synthesis of the CAA heuristics through close attention to terminology and description before translating the themes into a heuristic set. For example having a heuristic called 'access' (TH7, Table 44) would be rather ambiguous and not aid the evaluators when performing an evaluation. The initial heuristic set is displayed in Table 45 and 46.

Theme	Heuristic	Description
Same as Nielsen's Heuristics		
TH3. Reduce Errors	H3. Error prevention	Prevent errors from affecting test performance.
TH5. User Freedom	H5. Maximise 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.

Table 45 Initial Heuristic Set based on Retaining and Modifying Nielsen's

Of the 11 heuristics, 3 were based on Nielsen's original set, 2 were modifications, and 6 were new heuristics specific to CAA, see Table 46. The process of creating the heuristics from the themes was rather complex. Appropriate terminology was

important to encapsulate problems in the way that breaches to a heuristic could clearly be established. For example *Psychological and Perception* (TH9) proved to be difficult for the researcher and educational technologist to establish how a violation against this would be established 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 original heuristic set, Aesthetics and Minimalist Design would give rise to similar issues, so therefore the heuristic *Design should inspire trust and doesn't unfairly penalise* was named to capture the psychological theme.

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	H8. Accessing the test should be clear and intuitive	Users 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 the user 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 doesn't unfairly penalize	The user 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.

Table 46 New Heuristics synthesised not in Nielsen's set

11.7.1 Mapping Problems to Heuristic Set

With the initial heuristic set synthesised, the researcher and educational technologist then went through the process of cross-checking every problem in the final corpus against the new CAA heuristics. Each heuristics was numbered 1 to 11 and the problems were mapped to a heuristic, a decision was made to enable a problem to be mapped to more than one heuristic as in previous studies. Zuk *et al.*, (2006) claim that this enables heuristics to support each other by revealing problems from different standpoints. During this process one of the heuristics was extended to enable the incorporation of problem *LU10 – Recovery from errors*. This problem had the highest Damage Index score (0.92) therefore it was paramount that the heuristic set dealt with this issue. The heuristic error prevention was modified to *Prevent errors and aid recovery* and the final heuristic set is shown in table 46 along with the number of problems classified to the heuristic.

Number	Heuristic	Number of Problems
1	Navigating within the application and terminating the exam should be intuitive	8
2	Ensure appropriate interface design characteristics	7
3	Prevent errors and aid recovery	3
4	Answering question should be intuitive	4
5	Maximise user control and freedom	5
6	Prevent loss of input data	3
7	Accessing the test should be clear and intuitive	2
8	Use clear language and grammar within questions and ensure the score is clearly displayed.	5
9	Design should inspire trust and doesn't unfairly penalize	3
10	Minimise external factors which could affect the user	4
11	Ensure appropriate help and feedback	9

Table 47 Final Heuristic Set and problems classified to each heuristic

To ensure that the heuristics offered better coverage than Nielsen's heuristic it was important that each of the problems could be classified to at least one heuristic and this was achieved. In Section 11.6.1 the majority of the problems occurred with the task step code D – *during the test* and the synthesised heuristic set reflect this.

11.8 Conclusions

The primary objective of this chapter was to synthesise a set of evidence based domain specific heuristics and this has been achieved. The heuristic set that was synthesised 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. As discussed in Chapter 7, the validity of the new heuristics cannot be validated against user testing, as user testing is not well suited to the CAA domain. 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 from real world CAA applications, in contrast Nielsen (1994a) used 249 problems and these 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, it is expected that the 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.

Another objective was *To establish the limitations of the applied methodology* and a number of limitations to the method were discovered through its application within the CAA domain. Improvements could be made to the aggregation of the data gathered in each of the stages. As only two researchers performed the analysis of the problems and classification to the heuristics, there is the potential for bias to occur, especially if the same evaluators participate in each of the first three stages. Their prior experience may influence the process of mapping the problems to the heuristics. Using different evaluators throughout the various stages may help address this issue but unfortunately this was not feasible due to limited resources.

Within the literature audit stage it is difficult to verify that all the key literature has been examined, particularly as there has been limited research in the area of CAA

and usability. Often the evidence is hidden in journals or conference papers that do not directly relate to CAA. Having more than one domain expert perform a literature review may have helped alleviate this issue and uncover additional usability problems. However the user studies and heuristic evaluations will help compensate for this possible limitation. Additionally, if it is an emerging domain there will be limited research published or no reported usability problems, which would make this stage redundant. If this occurred then greater emphasis would need to be placed on the primary research stage with more than one application being evaluated.

Careful consideration is needed in the user study stage in relation to test task design. It may be useful to perform more than one evaluation of the application in order to try and maximise the number of problems identified within the system to help ensure appropriate coverage. If only a small percentage of problems are revealed then the heuristics may not accurately represent the domain thus making them ineffective.

Using the evidence based design approach to synthesise heuristics will require more time in the initial development stage compared to other techniques that just use a single method. A set of heuristics have been developed that ensure coverage of the severe problems within CAA applications however correctness, effectiveness and ease of use within the domain of CAA still need to be established. The heuristics could be used by educational technologists to evaluate the appropriateness of CAA applications.

The Damage Index proved to be a useful tool as the initial heuristic set was modified in Section 10.7.1 to ensure the problem with the highest ratio was adequately represented by the heuristic set. Without the ability to prioritise the corpus, then this problem may well have been overlooked, jeopardising the effectiveness of the final set.

11.8.1 Methodological Limitations

The researcher and educational technologist, after merging the problems, summarised the problem, which proved rather difficult in incorporating the merged problems into a single statement. This data was then used for the card sorting exercise and if the summary didn't accurately reflect all the problems this may lead to ambiguity and the development of inappropriate themes. However, any ambiguity

in the meaning of problems was discussed in the card sorting exercise and consensus was reached over the final set.

For the card sorting exercise it may have been beneficial to video record the process. Notes were made about the key decisions but the process was a lot more complex than initially anticipated and it was difficult to reflect back on the discussion after the event. The same issue arose with the task of converting the themes into heuristics, in some cases it took in excess of 60 minutes to devise the terminology especially for the perceptual / psychological theme.

The heuristics ensure coverage of all the known severe problems within CAA applications however no heuristic evaluation has been conducted using the new heuristic set. Further research would be required to establish correctness, effectiveness and ease of use. It is anticipated that the proposed heuristics may be further modified or enhanced at a later date to reflect the evolution of the technology and expansion of emerging question styles. This modification has been done in other studies for example Nielsen's original heuristic set was modified after a number of years to include additional heuristics (Nielsen, 1994a).

Chapter 12 Conclusions

12.1 Introduction

The work in this thesis was conducted over a six year period and was exploratory in nature. It was anticipated that the initial hypotheses would be revised and new areas of research would emerge. This proved to be the case and as a result the research methods adopted had to be continually refined in order to meet the objectives of the research. This chapter summarises the thesis by first re-examining the research approach in Section 12.2 and then summarising the major contributions in Section 13.3. Finally the chapter concludes with an outline of possible further research presented in Section 12.4.

12.2 Research Approach

The initial objective of the thesis outlined in Chapter 1 was to establish “*If severe usability problems exist that can cause users difficulties and dissatisfaction with unacceptable consequences whilst using existing commercial CAA software applications?*”. From this the following two hypotheses were formulated:

- Usability problems exist which could have an impact on students’ test results thus leading to unacceptable consequences.
- Students are satisfied with commercial CAA applications.

As a result of the literature review and analysis of various evaluation methods the decision was made to use a survey based approach. Without the co-operation of module leaders incorporating CAA into their modules it would not have been possible to use survey methods which may have hindered the research approach. However, this did constrain the research design as the author had no control over the question styles, number of questions or scoring algorithm. In Chapters 5 and 6 the results from the survey tool revealed a small number of usability problems that would certainly have unacceptable consequences from the students’ perspective. Thus the first hypothesis was proven and the survey results from Chapter 5 indicated that the students were satisfied with the CAA application, proving the second hypothesis to be true. The author considers that the initial primary aims of the

research were fulfilled however additional research questions emerged as a result of the limitations of the survey tools. The survey approach was ineffective at identifying usability problems, there was low inter-group consistency and the yield per evaluator was low at 0.06 (22/397), therefore another approach was required. The only viable approach, as identified in Chapter 3 was the use of inspection methods, in particular the heuristic evaluation.

Heuristic evaluations were judged to be a viable alternative to the survey methods as they alleviates many of the ethical and practical concerns of using alternative methodologies discussed in Chapters 3 and 4. Nielsen's heuristics (Nielsen & Mack, 1994) are the most widely cited and applied, however, based on the literature review of heuristics in Chapter 7, it was anticipated that Nielsen's heuristics would be ineffective within the CAA domain as they are too generic, therefore domain specific heuristics would be required. This assumption was then used to deduce the following hypothesis:

- Nielsen's heuristics are ineffective within the CAA domain.

Using the corpus of usability problems from Chapters 5 and 6 as the actual problem set, the data from the heuristic evaluations was analysed using the formula proposed by Hartson *et al.*, (2003) to establish the effectiveness. There was little overlap between the heuristic and survey data, the effectiveness score was also very low at 0.06 and there were a number of problems identified which could not be classified to a heuristic. This combination of results supported the hypothesis and Nielsen's heuristics were judged to be ineffective thus the decision was made to synthesise a set of domain specific heuristics for evaluating CAA applications. The next objective was to:

- Use the Evidence Based Design approach to Synthesise a set of domain specific heuristics

The research strategy at this point then focussed on the synthesis of evidence based domain specific heuristics for CAA, reported in Chapter 7 and applied in Chapter 11. The review of the literature in Chapter 7 revealed that there was no consensus upon a suitable method for synthesising domain specific heuristics therefore the evidence based design approach was formulated. It became evident in Chapters 5-9 that the evaluator effect could have an impact on corpus quality, as many of the

problems were unique to an individual evaluator. The evidence based design approach relied on a mixed method research strategy to diminish the evaluator effect and by conducting multiple evaluations a mechanism for the aggregation of data sets was required. This became the new objective of the thesis after Chapter 9:

- To devise a mechanism to enable the effective combination of results from different usability studies.

In Chapter 10 a Damage Index formula was proposed that enabled the data from multiple studies to be aggregated in a quantifiable and repeatable way, thus enabling the prioritisation of the problem set whilst alleviating bias from the aggregation process. This would enable the corpus to be prioritised and this data could then be used to synthesise the heuristic set to maximise coverage of the most severe problems within CAA.

Having established a mechanism for aggregating and prioritising the corpus, in Chapter 11, the next objective was to devise a set of domain specific heuristics using the evidence based design approach and Damage Index. Utilising experts in HCI and E-learning, the corpus was aggregated using card sorting techniques and the corpus prioritised using the Damage Index. A set of domain specific heuristics were synthesised and thus the main objective of the research outlined in Chapter 7 was fulfilled:

- Use the Evidence Based Design approach to Synthesise a set of domain specific heuristics

The ability to prioritise the corpus by utilising the Damage Index enabled the modification of the initial heuristic set thus ensuring adequate coverage of the most severe problems within the CAA domain. The heuristics were judged to offer better coverage of the severe problems within the CAA domain than Nielsen's set.

12.3 Contributions to Knowledge

The thesis is that '*severe usability problems exist in CAA applications that cause unacceptable consequences and that using an evidence based design approach CAA heuristics can be devised.*' There are four major contributions to knowledge from this research.

12.3.1 The Corpus of Usability Problems

Through surveys, heuristic evaluations and a literature review the thesis identifies severe usability problems in CAA applications that may lead to unacceptable consequences and these are summarised in Section 12.4.1. The most frequently occurring severe problem related to *Recovery from Errors*. This problem was identified in all studies, as incidents arose whereby students had lost their answers and in some more severe cases had to restart the test. The corpus is a contribution as it is the first published corpus of usability problems within the CAA domain; it is anticipated that this corpus could be used by software manufactures to improve future CAA applications; if software is improved then institutions may be more willing to incorporate CAA into their assessment strategies; for students, their test results may not be affected by poor usability.

12.3.2 Damage Index

A Damage Index was proposed in Chapter 10 to facilitate the aggregation of data from multiple evaluations, and was modified in Chapter 11 due to the use of a different severity rating scale. The Damage Index is represented in Figure 36.

$$DI = \frac{\bar{s} * n}{yN}$$

Figure 36 Damage Index Formula

- Damage Index = DI
- Mean Severity Score = \bar{s}
- Number of groups that identified the problem = n
- The upper bound of the severity rating scale = y
- Group size = N

This modification enables the formula to be generalisable, as the value of y can represent the upper bounds of any severity scale. The Damage Index is original within the context of usability evaluation methods and enables the prioritisation of data in a repeatable and quantifiable way, thus alleviating bias from the evaluators in

the aggregation process. It is envisaged that the formula has multiple applications, as it could be used to prioritise the results from a single study, whereby the number of groups is replaced by the number of evaluators, or it could be used to prioritise data sets from multiple evaluations as used within the context of this research.

12.3.3 Evidence Based Design

To establish how domain specific heuristics have been synthesised a critique of the literature was performed. This revealed that no established method existed and led to the synthesis of the evidence based design approach reported in Chapter 7 using a mixed method research strategy. In using the evidence based design approach, at stage 1, the objective was to determine the necessity for domain specific heuristics. An amalgam of empirical evidence was produced to demonstrate that Nielsen's heuristics were ineffective within the CAA domain. In Section 8.3.7 the formula proposed by Hartson *et al.*, (2003) was used to establish the effectiveness of the heuristics and they were shown to be ineffective. Further evidence was provided in Section 8.3.3, as predicted problems could not be classified to an appropriate heuristic. In Section 9.3.3 there was redundancy in the heuristic set with no problems being classified to *Support recognition rather than recall* and in Section 8.3.4 no problems were classified to *Flexibility and efficiency of use*. This led to the development of a corpus of usability problems that were then aggregated using a combination of techniques, including card sorting and the Damage Index, to enable the heuristics to be synthesised.

The evidence based design approach was judged to be a contribution as it was successfully applied within the CAA domain and it is anticipated that it could be adopted for the synthesis of heuristics within other domains such as Child Computer Interaction.

12.3.4 Heuristics for CAA

Another original contribution is a set of CAA heuristics synthesised by applying the evidence based design approach and the final heuristics are:

- Navigating within the application and terminating the exam should be intuitive
- Ensure appropriate interface design characteristics

- Prevent errors and enable recovery
- Answering question should be intuitive
- Maximise user control and freedom
- Prevent loss of input data
- Accessing the test should be clear and intuitive
- Use clear language and grammar within questions and ensure the score is clearly displayed.
- Design should inspire trust and doesn't unfairly penalise
- Minimise external factors which could affect the user
- Ensure appropriate help and feedback

The heuristics were shown to be valid in the study described in Chapter 11, the validation criteria was that they would offer coverage of the severe problems within the CAA domain. The heuristic set is original to the CAA domain and should enable the process of evaluating CAA applications to be more efficient than existing methods, thus improving the development of future applications.

A secondary contribution associated with the heuristics is a severity rating scale based on unacceptable consequences:

- Dissatisfied – the user would be unsatisfied but it is unlikely to affect the overall test performance
- Possible – there is a possibility that the problem may affect the users test performance
- Probable – it would probably affect the users test performance
- Certain – It would definitely affect the test performance of the user

One of the limitations of this scale is that no formal evaluation has been performed and no direct comparison has been made between this scale and the severity rating scale proposed by Nielsen.

12.4 Discussion

The work conducted in this thesis is relevant to both the Educational Technology and HCI communities. The majority of the work relating to usability and CAA has been published at Educational Technology conferences whilst the evidence based design approach would be of interest to those in HCI and has been published at INTERACT. The research approach could be used to develop heuristics for other domains outside CAA.

This section revisits the four main contributions to the thesis: the corpus, the heuristic, the Damage Index and evidence based design approach.

12.4.1 Corpus Revisited

Through the use of inspection methods, surveys and analysis of the literature a corpus of usability problems was synthesised. This corpus offers the potential to aid in the design of future CAA applications. By analysing the corpus developers can understand the severe issues inherent in existing application, therefore enabling the creation of more usable software. For example *LU10 – Recovery from errors* had the highest Damage Index score (0.92) and was reported in all the studies. Software developers therefore need to find ways of improving this process by preventing errors occurring and minimising the potential loss of data. Without this understanding it is unlikely that applications will improve and there may be scepticism amongst students and academics over the suitability of CAA due to poor usability.

Some of the reported problems were associated with issues relating to test design, such as poorly worded questions and negative marking (Chapter 5-10). Adebisin, *et al.* (2009) state that e-learning applications should be evaluated for pedagogical effectiveness and this argument could be applied to CAA whereby it is important to evaluate the test design. Although many problems are not system related and could be addressed through staff development, it is important for learning technologists to understand the issues associated with test design which are prevalent in CAA. This understanding will aid in procurement decisions and the effective training of staff. If these issues remain, ultimately they could have a negative affect on user satisfaction, or cause misinterpretation of the question leading to errors, resulting in barriers to uptake within institutions.

12.4.2 Heuristic Evaluations Revisited

While Nielsen's heuristics may be regarded as dated and inspection methods as inadequate, heuristic evaluation remains the best option for CAA where it is not possible to 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 instructor training and for use as part of a development lifecycle. Heuristic evaluations can be used to predict problems users may experience in a CAA environment and the new domain specific heuristics for CAA will improve the evaluation process, making it more efficient and effective. From the students perspective the user experience of future CAA applications may improve as software manufacturers now have an inspection method to evaluate applications as part of the development life cycle. The potential improvements could increase confidence in CAA as a viable assessment technique, thus improving the adoption of CAA within Schools, Further and Higher Education. Over the past decade pilot studies have been conducted in schools (Ashton & Bull, 2004) and universities (Sim & Holifield, 2004b), however CAA has not become integrated into institutions learning and teaching policies. In many institutions there still remains disparity between the uses of technology in the learning and assessment. For example in the authors own institution all modules are expected to use the learning management system, many assignments are submitted electronically, yet the summative assessments are administered through paper based exams. The potential the heuristics have for improving future CAA applications may help shorten the gap between the students learning and assessment experience.

In order for educational technologists to effectively use the new heuristics set, training in the evaluation process will be required to ensure problems reported are true positives and adequate coverage is achieved. The heuristic evaluation method relies upon the judgement of the evaluators and training is an integral part of the technique to ensure the effectiveness of the method.

It was clear from the literature that whilst domain specific heuristics have emerged, severity ratings have tended to be overlooked with reliance upon using Nielsen's original set. A set of domain specific severity ratings have been synthesised in Chapter 4 based on the consequences to the end user (students) and whether they would have grounds for appeal. It is anticipated that this will enable evaluators to

distinguish between the boundaries of the scale more effectively than using Nielsen's more generic scale. For academics or educational technologists the inter-rater consistency may improve in classifying the severity of a problem based on using this scale. This should enable more accurate and informed decisions to be made about the suitability of a CAA application.

12.4.3 Damage Index

A Damage Index was synthesised for prioritising the most severe problems when using a large number of evaluators or from multiple evaluations, Section 10.2.5, and this enables the prioritisation of usability problems in a systematic way. This should prevent resources being wasted on re-designing or fixing problems that would have no unacceptable consequences for the user. The Damage Index is not limited to the development of heuristics or heuristic evaluations, but can be generalised across many evaluations methods and domains. It could be used to prioritise usability problems from observational or user studies, providing a severity rating can be attributed to the problem.

The formula could be modified to be used in other domains, not just for prioritising the data from usability studies. For example it could be used in the area of computer security whereby problems are identified and ranked based on the severity of the potential threat (Whitman, 2003).

12.4.4 Evidence Based Design Approach Revisited

It was apparent from the literature that there was no consensus on how heuristics are developed and as a result of the limitations of current approaches the evidence based design approach was synthesised, reported in Section 7.7, and its application to CAA is discussed in Chapter 11.

The evidence based design approach is applied in a linear sequence, although it is feasible to complete stages in a different order. Reflecting upon the development of this method, it would be interesting to compare results of the application of the approach to a domain if stages were altered and see if this would influence the results.

The development process validated the heuristics to certain criteria, however, the reliability of the approach was never examined in this thesis but it is unlikely that

the exact same set of heuristics would be derived if two independent researchers examined the same domain. Both heuristic evaluations and the evidence based design approach rely on the judgement of the evaluators and there will inevitably be some variance. Despite the limitations, by gathering data from various sources and mapping these to the heuristics, the validity of the heuristics has been established for coverage.

Although there are several user groups the focus was on the students and their interaction with the application. Educational technologists would be able to use the heuristics to evaluate applications from the students' perspective but not other user groups. For example, it would not be possible to determine any usability problems an invigilator may encounter whilst the students are conducting the test. The initial set of heuristics presented in section 11.7.1 may need to be expanded to take into account other user groups. Also the evidence based design approach focused primarily on objective testing, although text entry style questions have been examined, no automated essay marking systems have been evaluated to inform the synthesis of the heuristic set. It may be feasible to further extend the evidence based design approach to incorporate data from automated assessment software, such as E-Rator (Powers *et al.*, 2002) and emerging fields such as Interactive Computer Marked Assessment (Jordan *et al.*, 2007).

12.5 Future Work

As the work in this thesis crossed two domains, Educational Technology and HCI and utilised a mixed method research strategy incorporating surveys and heuristics evaluations, a number of possible directions for future research have emerged.

12.5.1 Heuristics

The results from the heuristic evaluations published in Chapters 8 - 10 indicate that the majority of problems are classified to heuristics 1 to 5. The order the heuristics are presented to the evaluators may affect the classification of reported problems. Once a usability problem is identified the evaluator may classify this to the first heuristic they encounter, disregarding the others. This may hinder accurate identification of the problems or limit the ability of the software developer to rectify the problem. Future studies will change the order in which the heuristics are

presented to the evaluators to determine the effect it has on their classification of problems to a specific heuristic.

Inter-rater consistency is low with respect to severity ratings and the ability of evaluators to accurately classify a reported problem. There has been little published work on severity ratings and further research will analyse the effectiveness of the new severity rating scale, to establish the inter-rater consistency compared to Nielsen's severity rating scale. It is anticipated that the consequences scale will enable evaluators to distinguish between boundaries more effectively thus improving their classification of the severity of a problem.

With CAA there are different contexts of use such as formative, summative and diagnostic, it would be of interest to investigate severity ratings within different context. Evaluators may be able to use the consequences scale effectively within one context but not another. Further modification of the consequences scale may be necessary to enable them to be more generalisable within the area of assessment.

Although the heuristics have been validated for coverage they have not been applied within CAA. Despite this, the author feels that the heuristics adequately represent the domain and heuristic sets have been derived and published without being applied (Squires & Preece, 1999). Future work will use the heuristic set to evaluate the heuristic against other validation criteria discussed in Section 7.5, such as thoroughness (Sommervell & McCrickard, 2005).

12.5.2 Evidence Based Design

The approach has been applied to one domain, so questions still remain on its effectiveness at generating heuristics for other applications. Within the authors institution there is considerable research in Child Computer Interaction (CCI) and corpuses of usability problems are available from research studies using specific technologies such as mobile devices. These data sets could be used to synthesise a heuristic set within the CCI domain and this would enable the method to be refined and further identify any limitations of the approach. Comparisons could then be made by using the data from the heuristic evaluation and the existing corpus.

12.5.3 Comparing Evaluation Methods

Research in the thesis compared the effectiveness and efficiency of survey tools with heuristic evaluations in the context of CAA. A formula for evaluating the effectiveness of evaluation methods has been proposed by Hartson *et al.* (2003), although this is widely cited it does not appear to be widely applied. This formula was applied within the thesis, but there are limitations as no falsification testing could be performed and the problem set cannot have closure as additional studies will inevitably lead to additional problems. Further research will be conducted to establish alternative methods for comparing the results of different evaluation methods and examining the effectiveness and efficiency of these. This may be based on a combination of time yield and the formula proposed by Hartson *et al.*, (2003).

12.6 Concluding Remarks

When the research began, there was very little published literature on the usability of CAA applications and the resulting publications from this work have gone some way to address this, but the area is still largely ignored. A great deal of research has been published relating to CAA but mainly focusing on pedagogical challenges rather than usability. It is hoped that the contributions within this thesis will encourage new researchers in Educational Technology or HCI to investigate the usability of CAA applications. Through further analysis and publication of results of usability studies this might have an impact on the quality of future systems. As expected the technology used in the studies has also evolved and new versions of the software have come onto the market.

When starting this research, there was a vast amount of literature relating to heuristic evaluations and the domain is clearly understood. The two main areas of research within the domain focus upon improving the method and the creation of domain specific heuristics. However, there is no established methodology for creating a set of heuristics and the contributions made in this thesis go some way to meeting this challenge.

It is anticipated that the Damage Index will be adopted within the HCI community as it enables the formulaic merging of data sets from multiple evaluations in a reliable and systematic way. Within the HCI community it is acknowledged that the

aggregation of data from usability studies needs further research, as there is a tendency in publication to not reveal the aggregation process.

To conclude CAA is used in education institutions on a global scale and in internationally recognised certification programmes such as Apple Final Cut®. The author has shown throughout the thesis that usability matters within the context of CAA. Usability problems have been reported that could affect students' grades and potentially their degree classification. Although many of the problems reported may have been resolved through the evolution of the software it is important to further expand the research in this area. The new heuristic set will enable Educational Technologists and software developers to appropriately evaluate the usability of CAA applications, thus aiding their decision making process and lessening the prospects of students' grades being affected by poor usability.

Appendices

See CD for the data.

Appendix A – Chapter 5 WebCT® problems

Appendix B – Chapter 6 Questionmark® reported problems

Appendix C - Merged problems from WebCT® and Questionmark®

Appendix D – Chapter 8 merged data from summative context heuristic evaluation

Appendix E – Chapter 8 merged data from formative context heuristic evaluation

Appendix F – Chapter 9 Stage 1 aggregated sheet returned to evaluators to attach severity ratings

Appendix G - Chapter 9 formative problem set

Appendix H - Chapter 9 summative problem set

Appendix I - Chapter 9 stage 2 aggregation summative context

Appendix J - Chapter 9 merged problems sets

Appendix K – Front sheet of data capture form

Appendix L – Chapter 10 WebCT® data

Appendix M - Chapter 10 Questionmark® data

Appendix N – Chapter 10 TRIADS® data

Appendix O - Damage Index vs Consequence Scale

Appendix P - Literature review problems

Appendix Q – Merged problems from Chapter 10

Appendix R – Final Corpus

References

- Adams, A., Lunt, P., & Cairns, P. (2008). A qualitative approach to HCI research. In P. Cairns & A. L. Cox (Eds.), *Research Methods for Human Computer Interaction* (pp. 138-157). Cambridge: Cambridge University Press.
- Adebesin, T. F., De Villiers, M. R., & Ssemugabi, S. (2009). Paper presented at the Southern African Computer Lecturers' Association, Mpekweni Beach Resort, South Africa.
- Al-Amri, S. (2007). *Computer-Based vs Paper-Based Testing: Are they the same?* Paper presented at the 11th Computer Assisted Assessment Conference, Loughborough.
- Albion, P. R. (1999). *Heuristic Evaluation of Educational Multimedia: From Theory to Practice*. Paper presented at the ASCILITE, Brisbane.
- 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.
- Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing. A revision of blooms taxonomy of educational objectives.*: Longman.
- Ardito, C., Costabile, M. F., De Angeli, A., & Lanzilotti, R. (2006). *Systematic Evaluation of e-Learning Systems: An Experimental Validation*. Paper presented at the NordiCHI, Oslo.
- Ashton, H. S., & Bull, J. (2004). Piloting E-Assessment in Scottish Schools - Building on Past Experiences. *International Journal of E-Learning, April-June*, 74-84.
- Baker, K., Greenberg, S., & Gutwin, C. (2002). *Empirical Development of a Heuristic Evaluation Methodology for Shared Workspace Groupware*. Paper presented at the CSCW, New Orleans.
- Baranchik, A., & Cherkas, B. (2000). Correcting grade deflation caused by multiple-choice scoring. *International Journal of Mathematical Education in Science and Technology*, 31(3), 371-380.
- Barnum, C. (2003). The 'magic number 5' Is it enough for web testing? *Information Design Journal and Document Design*, 11(2/3), 160-170.
- Bennett, R. E., Goodman, M., Hessinger, J., Kahn, H., Liggett, J., Marshall, G., et al. (1999). Using multimedia in large-scale computer-based testing programs. *Computers in Human Behaviour*, 15(3), 283-294.
- Berg, G. A. (2000). Human-Computer Interaction (HCI) in Educational Environments: Implications of Understanding Computers as Media. *Journal of Educational Multimedia and Hypermedia*, 9(4), 347-368.
- 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 Human - Computer Studies*, 59, 823-835.
- Bertini, E., Gabrielli, S., & Kimani, S. (2006). *Appropriating and Assessing Heuristics for Mobile Computing*. Paper presented at the AVI, Venice.
- Besnard, D., & Arief, B. (2004). Computer security impaired by legitimate users. *Computers & Security*, 23(4), 253-264.

References

- Black, T. R. (1999). *Doing quantitative research in social sciences. An integrated approach to research design, measurement and statistics*. London: SAGE.
- Bloom, B. S. (1956). *Taxonomy of Educational Objectives: The classification of educational goals. Handbook 1. Cognitive Domain*: Longman.
- Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on Formative and Summative Evaluation of Student Learning*: McGraw-Hill Book.
- Bonham, S. W., Titus, A., Beichner, R. J., & Martin, L. (2000). Education Research Using Web-Based Assessment Systems. *Journal of Research on Computing in Education*, 33(1), 28-45.
- Boyle, A., & O'Hare, D. (2003). *Finding appropriate Methods to assure quality Computer-Based Assessment Development in UK Higher Education*. Paper presented at the 7th International Computer Assisted Assessment Conference, Loughborough.
- 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.
- BS7988. (2002). *Code of practice for the use of information technology (IT) in the delivery of assessment* (No. BS7988).
- Bull, J., & McKenna, C. (2001). *Blueprint for computer-assisted assessment*: Loughborough University.
- Burke, D. (2002). Assessment in basic training: what is the matrix? *Australian Psychiatry*, 10(3), 217-221.
- Burns, R. B. (2000). *Introduction to research methods* (4th Edition ed.). London: SAGE.
- Burstein, J., Leacock, C., & Swartz, R. (2001, 2001). *Automated evaluation of essays and short answers*. Paper presented at the Fifth International Computer Assisted Assessment Conference, Loughborough.
- Burton, R. F. (2001). Quantifying the effects of chance in Multiple Choice and True/False Tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, 26(1), 41-50.
- Burton, R. F., & Miller, D. J. (1999). Statistical Modelling of Multiple-Choice and True/False Tests: ways of considering, and of reducing, the uncertainties attributed to guessing. *Assessment & Evaluation in Higher Education*, 24(4), 399-411.
- Bush, M. (1999, 1999). *Alternative marking schemes for on-line multiple choice tests*. Paper presented at the 7th Annual Conference on the Teaching of Computing, Belfast.
- Byrne, M. D., Wood, S. D., Sukaviriya, N., Foley, J. D., & Kieras, D. E. (1994). *Automating Interface Evaluation*. Paper presented at the Human Factors in Computing Systems: Proceedings of CHI'94, Reading.
- Cairns, P. (2007). *HCI... Not as it should be: Inferential statistics in HCI research*. Paper presented at the HCI2007, Lancaster.
- Callea, D. a. K., T. (1997). Using computer-based tests. *ALT-J*, 5(1), 27-32.
- Card, S., Moran, T., & Newell, A. (1993). *The psychology of human computer interaction*: Lawrence Erlbaum.

References

- Carter, J., Ala-Mutka, K., Fuller, U., Dick, M., English, J., Fone, W., et al. (2004). *How shall we assess this?* Paper presented at the ITiCSE 04, Leeds.
- Cassady, J. C., & Johnson, R. E. (2002). Cognitive Test Anxiety and Academic Performance. *Contemporary Educational Psychology*, 27(2), 270-295.
- Chalmers, D., & McAusland, W. D. M. (2002). Computer-assisted Assessment. In J. Sloman & C. Mutchell (Eds.), *The Handbook for Economic Lecturers* (pp. 1-19): Economics LTSN.
- Chapman, G. (2006). *Acceptance and Usage of e-Assessment for UK Awarding Bodies - a research study*. Paper presented at the 10th International Computer Assisted Assessment Conference, Loughborough.
- Christic, J. R. (1999). *Automated Essay Marking for both content and style*. Paper presented at the 3rd Annual Computer Assisted Assessment Conference, Loughborough.
- CIAD. (2003). *Summary of Question Styles*. Retrieved 30/06/03, 2003, from <http://www.derby.ac.uk/ciad/ciastyles.html>
- Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: key factors associated with test mode effect. *British Journal of Educational Technology*, 33(5), 593-602.
- Cockton, G., Lavery, D., & Woolrych, A. (2007). Inspection-Based Evaluation. In A. Sears & J. A. Jacko (Eds.), *The Human Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (2nd Edition ed., pp. 1171-1190): Lawrence Erlbaum Associates.
- Cockton, G., Woolrych, A., & Hindmarch, M. (2004). *Reconditioned Merchandise: Extending Structured Report Formats in Usability Inspection*. Paper presented at the CHI 2004, Vienna.
- Cohen, L., Manion, L., & Morrison, K. (2001). *Research methods in Education*. London: RoutledgeFalmer.
- Connell, I. W., & Hammond, N. V. (1999). *Comparing usability evaluation principles with heuristics: problem instances versus problem types*. Paper presented at the Human-Computer Interaction - INTERACT '99, Edinburgh.
- Conole, G., & Fill, K. (2005). A learning design toolkit to create pedagogically effective learning activities. *Journal of Interactive Media in Education* (8), 1-16.
- Conole, G., & Warburton, B. (2005). A review of computer-assisted assessment. *ALT-J*, 13(1), 19-33.
- Cosemans, D., Van Rentergem, L., Verburch, A., & Wils, A. (2002, 2002). *A campus wide set up of question mark perception (v2.5) at the Katholieke Univeriteit Leuven-Facing a large scale implementation*. Paper presented at the Sixth International Computer Assisted Assessment Conference, Loughborough.
- Cox, K., & Clark, D. (1998). The Use Of Formative Quizzes for Deep Learning. *Computers & Education*, 30(3), 157-167.
- Coyle, C. L., Iden, R., Kotval, X. P., Santos, P. A., & Vaughn, H. (2007). *Heuristic Evaluations at Bell Labs: Analyses of Evaluator Overlap and Group Sessions*. Paper presented at the CHI, San Jose.
- Cresswell, J. W. (2003). *Research design, qualitative, quantitative and mixed methods approaches*. London: SAGE.
- Cresswell, J. W. (2007). *Designing and Conducting Mixed Methods Research*. London: Sage Publications Ltd.
- Crisp, B. R. (2002). Assessment methods in social work education: a review of the literature. *Social Work Education*, 21(2), 259-269.

References

- Croft, A. C., Danson, M., Dawson, B. R., & Ward, J. P. (2001). Experience of using computer assisted assessment in engineering mathematics. *Computers & Education*, 37(1), 53-66.
- Curtis, P. (2003, 14/05/03). Missing paper sparks exam reprint. *Guardian*.
- Daly, C., & Waldron, J. (2002, 2002). *Introductory programming, problem solving and computer assisted assessment*. Paper presented at the Sixth International Computer Assisted Assessment Conference, Loughborough.
- Davies, P. (2002, 2002). *There's no confidence in multiple-choice testing*. Paper presented at the Sixth International Computer Assisted Assessment Conference, Loughborough.
- Desurvire, H., Caplan, M., & Toth, J. A. (2004). *Using heuristics to Evaluate the Playability of Games*. Paper presented at the CHI, Vienne.
- Desurvire, H., Kondziela, J., & Atwood, M. (1992). *What is gained or lost when using evaluation methods other than empirical testing*. Paper presented at the CHI, Monterey.
- Diener, E., & Crandall, R. (1978). *ethics in Social and Behavioural Research*. Chicago: University of Chicago Press.
- Dix, A., Finlay, J., Abowd, G. D., & Beale, R. (2004). *Human-Computer Interaction* (Third ed.): Pearson Education Limited.
- Dowsing, R. D. (1998, 1998). *Flexibility and the Technology of Computer Aided Assessment*. Paper presented at the ASCILITE 1998, Wollongong.
- DuBois, P. H. (1964, 1964). *A test dominated society: China, 1115 B.C.1905A.D.* Paper presented at the The 1964 Invitational Conference on Testing Problems.
- Dyson, M. C., & Kipping, G. J. (1997). The legibility of screen formats: are three columns better than one? *Computers & Graphics*, 21(6), 703-712.
- Ebel, R. L. (1972). *Essentials of Educational Measurement*: Prentice-Hall.
- Eckersley, C. (2004). *Evaluation of a Learning Object*. Paper presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Lugano.
- Embi, Z. C., & Hussain, H. (2005). 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*, 24(1), 27-42.
- Ergun, E., & Namlu, A. G. (2004). *Factors Effecting University Students' Computer Assisted Assessment Success and Student Perceptions*. Paper presented at the Society for Information Technology and Teacher Education International Conference, Albuquerque.
- Evans, C., Gibbons, N. J., Shah, K., & Griifin, D. K. (2004). Virtual learning in the biological sciences: pitfalls of simply putting notes on the web. *Computers & Education*, 43(1-2), 49-61.
- Evans, C., & Sabry, K. (2003). Evaluation of the Interactivity of web-based learning systems: Principles and Process. *Innovations in Education and Teaching International*, 40(1), 89-99.
- Fereday, J., & Muir-Cochrane. (2006). Demonstrating Rigour Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods*, 5(1), 1-11.
- Fernandes, G., & Holmes, C. (2002). *Applying HCI to music related hardware*. Paper presented at the CHI 2002, Minneapolis, Minnesota.

References

- Frohlich, R. (2000, 2000). *Keeping the wolves from the door, wolves in sheep clothing, that is*. Paper presented at the 4th International Computer Assisted Assessment Conference, Loughborough.
- Frokjaer, E., Hertzum, M., & Hornbaek, K. (2000). *Measuring Usability: Are effectiveness, efficiency and satisfaction really correlated*. Paper presented at the Conference on Human Factors in Computing Systems, The Hague.
- Fu, L., Salvendy, G., & Turley, L. (2002). Effectiveness of user testing and heuristic evaluation as a function of performance classification. *Behaviour & Information Technology*, 21(2), 137-143.
- Fulcher, G. (2003). Interface design in computer-based language testing. *Language Testing*, 20(4), 384-408.
- Giza, B., & Awalt, C. (2005). *A longitudinal Study of Technology Infusion in a Major Teaching Institutions*. Paper presented at the SITE, Pheonix.
- Graham, D. (2004). *A Survey of assessment methods employed in UK Higher Education programmes for HCI courses*. Paper presented at the 7th HCI Educators Workshop, University of Central Lancashire.
- Gray, W. D., & Salzman, M. C. (1998). Damaged Merchandise? *Journal of Human-Computer Interaction*, 13(4), 203-262.
- Gunn, C. (1995). *An example of Formal Usability Inspections in Practice at Hewlett-Packard Company*. Paper presented at the Conference on Human Factors in Computing Systems, Denver.
- Haladyna, T. M. (1996). *Writing Test Items to Evaluate Higher Order Thinking*: Allyn & Bacon.
- Harper, R. (2002, 2002). *Allowing for guessing and for expectations from the learning outcomes in computer-based assessments*. Paper presented at the Sixth International Computer Assisted Assessment Conference, Loughborough.
- Harper, R. (2003). Correcting computer-based assessment for guessing. *Journal of Computer Assisted Learning*, 19(1), 2-8.
- Hartson, H. R., Andre, T. S., & Williges, R. C. (2003). Criteria for Evaluating Usability Evaluation Methods. *International Journal Human Computer Interaction*, 15(1), 145-181.
- Hatton, S., Boyle, A., Byrne, S., & Wooff, C. (2002, 2002). *The use of PGP to provide secure email delivery of CAA results*. Paper presented at the Sixth International Computer Assisted Assessment Conference, Loughborough.
- Hawe, E. (2003). It's pretty difficult to fail; the reluctance of lecturers to award a fail grade. *Assessment & evaluation in higher education*, 28(4), 371-382.
- Heinrich, E., & Wang, Y. (2003, 2003). *Online marking of Essay-type assignments*. Paper presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Hawaii.
- Herd, G., & Clark, G. (2002). *Computer Assisted Assessment Implementing CAA in FE Sector in Scotland: Question Types*: Glenrothes College.
- Hertzum, M. (2006). Problem Prioritization in Usability Evaluation: From Severity Assessment towards Impact on Design. *International Journal Human Computer Interaction*, 21(2), 125-146.
- Hertzum, M., & Jacobsen, N., E. (2001). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 13(4), 421-443.
- HESA. (1995). *Students in Higher Education Institutions 1994/95*.
- HESA. (2002). *Students in Higher Education Institutions 2001/02*.

References

- HESA. (2008). *First year UK domiciled HE students by qualification aim, mode of study, gender and disability 2007/08*.
- Hibert, D. M., & Redmiles, D. F. (2000). Extracting Usability Information from User Interface Events. *ACM Computing Surveys*, 32(4), 384-421.
- Hindle, S. (2003). Careless about privacy. *Computers & Security*, 22(4), 284-288.
- Hodson, P., Saunders, D., & Stubbs, G. (2002). Computer-Assisted Assessment: Staff Viewpoints on its Introduction within a New University. *Innovations in Education and Teaching International*, 39(2), 145-152.
- Horney, W. (2003). Assessing Using grade-related criteria: a single currency for universities? *Assessment & Evaluation in Higher Education*, 28(4), 435-454.
- Howitt, D., & Cramer, D. (2003). *An introduction to statistics in psychology* (Revised 2nd Edition ed.). Harlow: Pearson education.
- 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).
- Ivory, M. Y., & Hearst, M. A. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys*, 33(4), 470-516.
- 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.
- Jafarpur, A. (2003). Is the test constructor a facet? *Language Testing*, 20(1), 57-87.
- Jefferies, P., Constable, I., Kiely, B., Richardson, D., & Abraham, A. (2000, 2000). *Computer Aided Assessment using WebCT*. Paper presented at the Fourth International Computer Assisted Assessment Conference, Loughborough.
- Jeffries, R., & Desurvire, H. (1992). Usability Testing vs Heuristic Evaluation: Was there a contest? *SIGCHI Bulletin*, 24(4), 39-41.
- Jeffries, R., Wharton, C., & Uyeda, K. M. (1992). *User interface evaluation: A comparison of four techniques*. Paper presented at the CHI 91, New Orleans.
- John, B. E., & Kieras, D. E. (1996). Using GOMS for user interface design and evaluation: Which Technique? *ACM Transactions on Computer-Human Interaction*, 3(4), 287-319.
- Johnson, C. M., Johnson, T., & Zahang, J. J. (2000). Increase productivity and reducing errors through usability analysis: a case study and recommendations. *Journal of the American Medical Informatics Association*, 394-398.
- Jordan, S., Brockbank, B., & Butcher, P. (2007). *Extending the pedagogic role of online interactive assessment: providing feedback on short free-text responses*. Paper presented at the Assessment Design for learner responsibility, Glasgow.
- Karat, C. M. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. *CHI 92*, pp. 397-404.
- King, T., & Duke-Williams, E. (2002, 2002). *Using Computer Aided Assessment to Test Higher Level Learning Outcomes*. Paper presented at the Fifth International Computer Assisted Assessment Conference, Loughborough.
- Kleeman, J., & Osborne, C. (2002, 2002). *A practical look at delivering assessment to BS7988 recommendations*. Paper presented at the Sixth International Computer Assisted Assessment Conference, Loughborough.

References

- Knight, P. (2001). *A briefing on key concepts formative and summative, criterion and norm-referenced assessment* (No. Series No 7): LTSN Generic Centre.
- Korhonen, H., & Koivisto, E. M. (2006). *Playability Heuristics for Mobile Games*. Paper presented at the MobileHCI, Helsinki.
- Kurosu, M., Sugizaki, M., & Matsuura, S. (1999). *A comparative study of sHEM*. Paper presented at the HCI International 99.
- Latu, E., & Chapman, E. (2002). Computerised adaptive testing. *British Journal of Educational Technology*, 33(5), 619-622.
- Lavery, D., Cockton, G., & Atkinson, M. P. (1997). Comparison of evaluation methods using structured usability problem reports. *Behaviour & Information Technology*, 16(4/5), 246-266.
- Law, E. L.-C., & Hvannberg, E. T. (2008). *Consolidating Usability Problems with Novice Evaluators*. Paper presented at the NordiChi, Lund.
- Lilley, M., Barker, T., & Britton, C. (2004). The development and evaluation of a software prototype for computer-adaptive testing. *Computers & Education*, 43(1), 109-123.
- Ling, C., & Salvendy, G. (2005). Extension of Heuristic evaluation method: a review and reappraisal. *International Journal of Ergonomics and Human Factors*, 27(3), 179-197.
- Liu, M., Papathanasiou, E., & Hao, Y. (2001). Exploring the use of multimedia examination formats in undergraduate teaching: results from the fielding testing. *Computers in Human Behaviour*, 17(3), 225-248.
- Lloyd, D., Martin, J. G., & McCaffery, K. (1996). The introduction of computer based testing on an engineering technology course. *Assessment & Evaluation in Higher Education*, 21(1), 83-90.
- Luck, M., & Joy, M. (1999). A Secure On-Line Submission System. *Software - Practice and Experience*, 29(8), 721-740.
- Lynch, B. K. (2001). Rethinking assessment from a critical perspective. *Language Testing*, 18(4), 351-372.
- Macdonald, J., & Twining, P. (2002). Assessing activity-based learning for a networked course. *British Journal of Educational Technology*, 33(5), 603-618.
- Mackenzie, D. (1999, 1999). *Recent Developments in the Triartite Interactive Assessment Delivery System (TRIADS)*. Paper presented at the 3rd Annual Computer Assisted Assessment Conference, Loughborough.
- Mackenzie, D., Hallam, B., Baggott, G., & Potts, J. (2002, 2002). *TRIADS experiences and developments. A Panel Discussion*. Paper presented at the Sixth International Computer Assisted Assessment Conference, Loughborough.
- Mackenzie, D., & O'Hare, D. (2002, 2002). *Empirical Prediction of the measurement scale and base level 'Guess Factor' for advanced computer-based assessment*. Paper presented at the 6th International Conference of Computer Aided Assessment, Loughborough.
- Mackenzie, D., O'Hare, D., Paul, C., Boyle, A., Edwards, D., Willimas, D., et al. (2004). Assessment for Learning: The Triads assessment of learning outcomes project and the development of a pedagogically friendly computer-based assessment system. In D. O'Hare & D. Mackenzie (Eds.), *Advances in Computer Aided Assessment* (pp. 11-25). Birmingham: SEDA.
- Mankoff, J., Dey, A. K., Hsieh, G., Kientz, J., Lederer, S., & Ames, M. (2003). *Heuristic Evaluation of Ambient Displays*. Paper presented at the SIGCHI, Ft. Lauderdale.

References

- Masemola, S. S., & De Villiers, M. R. (2006). *Towards a framework for usability testing of interactive e-learning applications in cognitive domains, Illustrated by a case study*. Paper presented at the SAICSIT, Cape Winelands.
- Mason, S. (2003). Electronic Security is a Continuous Process. *Computer Fraud & Security*, 2003(1), 13-15.
- Mayer, R. E. (2002). A taxonomy for computer-based assessment of problem solving. *Computers in Human Behaviour*, 18(6), 623-632.
- McAlpine, M. (2002). *Principles of Assessment*: CAA Centre.
- McCabe, M., & Barrett, D. (2003, 2003). *CAA Scoring Strategies for Partial Credit and Confidence Levels*. Paper presented at the 7th International Computer Assisted Assessment Conference, Loughborough.
- McKenna, C., & Bull, J. (2000). Quality assurance of computer-assisted assessment: practical and strategic issues. *Quality Assurance in Education*, 8(1), 24-31.
- McLaughlin, P. J., Fowell, S. L., Dangerfield, P. H., Newton, D. J., & Perry, S. E. (2004a). Development of Computerised Assessment (TRIADS) in an undergraduate medical school. In D. O'Hare & D. Mackenzie (Eds.), *Advances in Computer Aided Assessment* (pp. 25-32). Birmingham: SEDA.
- McLaughlin, P. J., Fowell, S. L., Dangerfield, P. H., Newton, D. J., & Perry, S. E. (2004b). 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.
- Medlock, M. C., D., W., Terrano, M., R., R., & B., F. (2002). *Using the RITE Method to improve products: a definition and a case study*. Paper presented at the Usability Professionals Association, Orlando.
- Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2002). *Computer-Based Testing. Building the Foundation for Future Assessments* (First ed.). Mahwah: Lawrence Erlbaum Associates.
- Moore, A., Parr, G., Logan, M., Neeley, H., Roesner, D., & Durer, U. (2001). Developing a European internet and kiosk-based health information system. *Journal of Medical Internet Research*, 3(1), 1-6.
- Morgan, M. R. J. (1979). MCQ: An interactive computer program for multiple-choice self testing. *Biochemical Education*, 7(3), 67-69.
- Moskal, B. M., & Leydens, J. A. (2000). Scoring rubrics development: Validity and Reliability. *Practical Assessment, Research and Evaluation*, 7(10).
- NAA. (2008). *Changes to national curriculum tests and teacher assessments*, from http://www.naa.org.uk/libraryAssets/media/Changes_to_national_curriculum_tests_and_teacher_assessments.pdf
- Nielsen, J. (1992). *Finding usability problems through heuristic evaluation*. Paper presented at the Proceedings of the SIGCHI conference on Human factors in computing systems, Monterey.
- Nielsen, J. (1994a). *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. (1994b). *Usability Engineering*: Morgan Kaufmann.
- Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 206-213). Amsterdam, The Netherlands: ACM Press.

References

- 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.
- Noyes, J., Garland, K., & Robbins, L. (2004). *Paper-based versus computer-based assessment: is workload another test mode effect?* *British Journal of Educational Technology*, 35(1), 111-113.
- O'Hare, D. (2001). *Students views of formative and summative CAA*. Paper presented at the 5th International Computer Assisted Assessment Conference, Loughborough.
- O'Hare, D., & Mackenzie, D. (2004). Production of a CAA - advice for the academics. In D. O'Hare & D. Mackenzie (Eds.), *Advances in Computer Aided Assessment* (Vol. Paper 116, pp. 118): SEDA.
- O'Leary, R., & Cook, J. (2001, 2001). *Wading through treacle: CAA at the University of Bristol*. Paper presented at the 5th International Computer Assisted Assessment Conference, Loughborough.
- Paddison, C., & Englefield, P. (2003). *Applying Heuristics to Perform a Rigorous Accessibility Inspection in Commercial Context*. Paper presented at the CUU, Vancouver.
- Paddison, C., & Englefield, P. (2004). Applying heuristics to accessibility inspections. *Interacting with Computers*, 16(2), 507-521.
- Pain, D., & Le Heron, J. (2003). WebCT and Online Assessment: The best thing since SOAP? *Educational Technology & Society*, 6(2), 62-71.
- Parahoo, K. (2006). *Nursing research: principles, process and issues* (Second Edition ed.). Basingstoke: Palgrave Macmillan.
- Parlangeli, O., Marchigiani, E., & Bagnara, S. (1999). Multimedia systems in distance education: effects of usability on learning. *Interacting with Computers*, 12(1), 37-49.
- Paterson, J. S. (2002, 2002). *Linking on-line assessment in mathematics to cognitive skills*. Paper presented at the Sixth International Computer Assisted Assessment Conference, Loughborough.
- Patterson, A., & Bellaby, G. (2001). *An initial experiment with computer aided assessment in level one systems analysis*. Paper presented at the 3rd Annual Conference of the LTSN Information & Computer Science, Loughborough.
- Paxton, M. (2000). A Linguistic Perspective on Multiple Choice Questioning. *Assessment & Evaluation in Higher Education*, 25(2), 109-119.
- Phipps, L., & McCarthy, D. (2001). *Computer Assisted Assessment and Disabilities*. Paper presented at the 5th International Computer Assisted Assessment Conference, Loughborough.
- Piguet, A., & Peraya, D. (2000). Creating web-integrated learning environments: An analysis of WebCT authoring tools in respect to usability. *Australian Journal of Educational Technology*, 16(3), 302-314.
- Pinelle, D., Wong, N., & Stach, T. (2008). *Heuristic Evaluation for Games: Usability Principles for Video Game Design*. Paper presented at the CHI2008, Florence.
- Polson, P., Lewis, C., Rieman, J., & Wharton, C. (1992). Cognitive walkthroughs: A method for theory-based evaluation of user interface. *International Journal of Man-Machine Studies*, 36, 741-773.

References

- Pommerich. (2004). developing computerized versions of paper-and-pencil tests: mode effects for passage-based tests. *The journal of technology, learning and assessment*, 2(6), 3-44.
- Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rator: challenging the validity of automated essay scoring. *Computers in Human Behaviour*, 18(2), 103-134.
- 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.
- Quinn, C. N. (1996). *Pragmatic Evaluation: Lessons from Usability*. Paper presented at the ASCILITE, Adelaide.
- Race, P. (1995). The art of assessing. *The New Academic*, 4(3).
- Rafilson, F. (1991). The case for validity generalization. *Practical Assessment, Research and Evaluation*, 2(13), 1-3.
- Read, J. C., MacFarlane, S. J., & Casey, C. (2001). *Measuring the Usability of Text Input Methods for Children*. Paper presented at the HCI2001, Lille, France.
- Reeves, T. C., Benson, L., Elliott, D., Grant, M., Holschuh, D., Kim, B., et al. (2002). *Usability and Instructional Design Heuristics for E-learning Evaluation*. Paper presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Denver.
- Reid, N. (2002, 2002). *Designing online quiz questions to assess a range of cognitive skills*. Paper presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Denver.
- 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.
- 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.
- Ricketts, C., & Zakrzewski, S. (2004). *How do the risks of web-based CAA system differ from those of a closed network system?* Paper presented at the 8th International Computer Assisted Assessment Conference, Loughborough.
- Ridgway, J., & McCusker, S. (2004). *Literature Review of E-assessment*.
- Ripley, M., Tafler, J., Ridgway, J., Harding, R., & Redif, H. (2009). *Review of Advanced e-Assessment Techniques: JISC*.
- Rowntree, D. (1987). *Assessing Students: How shall we know them*. New York: Kogan page.
- Rudner, L. M. (1998). *An on-line, interactive, computer adaptive testing tutorial*. Retrieved 09/12/02, 2002, from <http://ericae.net/scripts/cat/catdemo.htm>
- Sabar, N. (2002). Towards principled practice in evaluation: Learning from instructors' dilemmas in evaluating graduate students. *Studies in Educational Evaluation*, 28(4), 329-345.
- Salvia, J., & Ysseldyke, J. (1991). *Assessment* (First ed.): Houghton Mifflin.
- Sapsford, R. (1999). *Survey Research*. London: SAGE.
- Schenkman, B., Fukuda, T., & Persson, B. (1999). Glare from monitors measured with subjective scales and eye movements. *Displays*, 20, 11-21.
- Sclater, N., & Howie, K. (2003). User requirements of the 'ultimate' online assessment engine. *Computers & Education*, 40(3), 285-306.
- Sears, A. (1997). Heuristic Walkthroughs: Finding the problems without the noise. *International Journal Human Computer Interaction*, 9, 213-234.
- SENDA. (2001). Special Educational Needs and Disability Act 2001.

References

- Shackel, B. (1986). *Ergonomics in design for usability*. Paper presented at the HCI 86 Conference on People and Computers II, Cambridge.
- Shneiderman. (1998). *Designing the User Interface* (3rd Edition ed.). Massachusetts: Addison Wesley Longman.
- Sim, G., & Holifield, P. (2004a). *Computer Assisted Assessment: All those in favour tick here*. Paper presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Lugano.
- Sim, G., & Holifield, P. (2004b). *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.
- Sim, G., MacFarlane, S., & Read, J. C. (2006). All Work and No Play: Measuring Fun, Usability and Learning in Software for Children. *Computers and Education*, 46(3), 235-248.
- Sim, G., Malik, N. A., & Holifield, P. (2003). *Strategies for large-scale assessment: an institutional analysis of research and practice in a virtual university*. Paper presented at the 7th International Computer Assisted Assessment, Loughborough.
- Sim, G., Read, J. C., & Cockton, G. (2009). *Evidence based Design of Heuristics for Computer Assisted Assessment*. Paper presented at the 12th IFIP TC13 Conference in Human Computer Interaction, Uppsala.
- Sim, G., Read, J. C., & Holifield, P. (2006a). *Evaluating the user experience in CAA Environments: What affects user satisfaction?* Paper presented at the 10th International Computer Assisted Assessment Conference, Loughborough.
- Sim, G., Read, J. C., & Holifield, P. (2006b). *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.
- Smythe, C., & Roberts, P. (2000). *An overview of the IMS question & test interoperability specification*. Paper presented at the 4th International Computer Assisted Assessment Conference, Loughborough.
- Sommervell, J., & McCrickard, D., S. (2005). Better discount evaluation: illustrating how critical parameters support heuristic creation. *Interacting with Computers*, 17(5), 592-612.
- Sommervell, J., Wahid, S., & McCrickard, S. (2003). *Usability Heuristics for Large Screen Information Exhibits*. Paper presented at the Human-Computer Interaction - INTERACT, Zurich.
- Spool, J., & Schroeder, W. (2001). *Testing Websites: Five Users is Nowhere Near Enough*. Paper presented at the CHI, Seattle.
- Squires, D., & Preece, J. (1999). Predicting quality in educational software: Evaluating for learning, usability and the synergy between them. *Interacting with Computers*, 11, 467-483.

References

- Stephens, D. (1994). Using computer-assisted assessment: time saver or sophisticated distractor? *Active Learning*, 1, 11-15.
- Stephens, D., Bull, J., & Wade, W. (1998). Computer-assisted Assessment: suggested guidelines for an institutional strategy. *Assessment & Evaluation in Higher Education*, 23(3), 283-294.
- Stevenson, A., Sweeny, P., Greenan, K., & Alexander, S. (2002, 2002). *Integrating CAA within the University of Ulster*. Paper presented at the Sixth International Computer Assisted Assessment Conference, Loughborough.
- Tannenbaum, R. S. (1999). *Theoretical foundations of multimedia*. New York: W.H.Freeman and Company.
- Taras, M. (2002). Using Assessment for Learning and Learning from Assessment. *Assessment & Evaluation in Higher Education*, 27(6), 501-510.
- Thomas, P., Price, B., Paine, C., & Richards, M. (2002). Remote electronic examinations: student experience. *British Journal of Educational Technology*, 33(5), 537-549.
- Thompson, M., & Radigan, J. (2002). *Listservs Facilitate a cross-sectional study of an instructional technology class for preservice teachers*. Paper presented at the SITE, Nashville.
- Valenti, S., Cucchiarelli, A., & Panti, M. (2002). Computer Based Assessment Systems Evaluation via the ISO90126 Quality Model. *Journal of Information Technology Education*, 1(3), 157-175.
- Van Veenendaal, E. (1998). *Questionnaire based usability testing*. Paper presented at the European Software Quality Week, Brussels.
- Walker, D. M., & Thompson, J. S. (2001). A note on Multiple Choice Exams, with Respect to Students Risk Preference and Confidence. *Assessment & Evaluation in Higher Education*, 26(3), 261-267.
- Walsh, W. B., & Betz, N. E. (1985). *Tests and Assessment*: Prentice-Hall.
- Warburton, B., & Conole, G. (2003). *CAA in UK HEIs - The state of the art?* Paper presented at the 7th International Computer Assisted Assessment Conference, Loughborough.
- Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The Cognitive Walkthrough Method: A Practitioner's Guide. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 105-141). New York: John Wiley & Sons.
- Whitefield, A., Wilson, F., & Dowell, J. (1991). A framework for human factors evaluation. *Behaviour and Information Technology*, 10(1), 65-79.
- Whitman, M. E. (2003). Enemy at the gate: threats to information security. *Communications of the ACM*, 46(8), 91-95.
- Whittingham, D. (1999). Technical and Security Issues. In S. Brown, J. Bull & P. Race (Eds.), *Computer Assisted Assessment in Higher Education* (pp. 205). London: Kogan Page.
- Wiles, K. (2002). Accessibility and computer-based assessment: a whole new set of issues? In L. Phipps, A. Sutherland & J. Seale (Eds.), *Access All Areas: disability, technology and learning* (pp. 88): JISC TechDis Service ALT.
- Wiles, K., & Ball, S. (2003). *Constructing accessible CBA: Minor works or major renovations?* Paper presented at the 7th International Computer Assisted Assessment Conference, Loughborough.
- Wilson, R., Shortreed, J., & Landoni, M. (2004). *A study into the usability of e-encyclopaedias*. Paper presented at the Symposium on Applied Computing, Melbourne.

References

- Wiltfelt, C., Philipsen, P. E., & Kaiser, B. (2002). Chat as Media in Exams. *Education and Information Technologies*, 7(4), 343-349.
- Wixon, D., Jones, S., Tse, L., & Casaday, G. (1994). Inspections and Design Reviews: Framework, History, and Reflection. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 77-98). New York: John Wiley & Sons.
- Wood, D. A. (1960). *Test Construction* (First ed.). Columbus: Charles E. Merrill Books Inc.
- Woolrych, A., & Cockton, G. (2000). *Assessing Heuristic Evaluations: Mind the Quality, not just Percentages*. Paper presented at the British HCI Group HCI 2000 Conference, London.
- Woolrych, A., & Cockton, G. (2001). *Why and When Five Test Users aren't enough*. Paper presented at the Proceedings of IHM-HCI 2001 Conference, Toulouse.
- 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.
- Woolrych, A., Cockton, G., & Hindmarch, M. (2004). *Falsification Testing for Usability Inspection Method Assessment*. Paper presented at the HCI2004.
- Yorke, M., Barnett, G., Bridges, P., Evanson, P., Haines, C., Jenkins, D., et al. (2002). Does grading method influence honours degree classification? *Assessment & Evaluation in Higher Education*, 27(3), 269-279.
- Zakrzewski, S., & Steven, S. (2003). Computer-based assessment: quality assurance issues, the hub of the wheel. *Assessment & Evaluation in Higher Education*, 28(6), 609-623.
- Zhang, J., Johnson, T. R., Patel, V. L., Paige, D. L., & Kubose, T. (2003). Using usability heuristics to evaluate patient safety of medical devices. *Journal of Biomedical Informatics*, 36(1), 23-30.
- Zuk, T., Schlesier, L., Neumann, P., Hancock, M. S., & Carpendale, S. (2006). *Heuristics for Information Visualization Evaluation*. Paper presented at the BELIV, Venice.