Instructional Multimedia:

Comparison and Enhancement of Expert Evaluation Methods

Maia Tzokova Dimitrova

Submitted for Examination of Doctor of Philosophy

Centre for Human-Computer Interaction Design
City University
London

VOL I

May, 2002

Volume 1

Table of Contents

Abstract 13

Chapter 1: Introduction 16

1.1 Introduction 17

1.2 Instructional Multimedia: Definition and Evolution 18

1.2.1 Definition of Instructional Multimedia 18

1.2.2 Evolution of Instructional Multimedia 20

1.3 The Quality of Instructional Multimedia 22

1.3.1 Usability 22

1.3.2 Learning Effectiveness 23

1.4 Building Quality Instructional Multimedia 24

1.4.1 The Importance of Ensuring the Quality of IMM 24

1.4.2 A Learner-Centred Approach to the Development of IMM 24

1.4.3 The Process of Formative Evaluation of IMM 26

1.5 Formative Evaluation Methods for Instructional Multimedia 29

1.5.1 Existing Formative Evaluation Methods for IMM 29

1.5.2 Assessing the Effectiveness of Existing Evaluation Methods for IMM 30

1.6 Thesis Scope and Objectives 31

1.7 Research Methods Used 33

1.8 The Organisation of the Thesis 35

1.9 Summary of the Thesis Contribution to the Field 37

Chapter 2: Formative Evaluation Methods and their Effectiveness 38

2.1 Introduction 39

2.2 Current Practice of Formative Evaluation of IMM 40

2.3 Review of Existing Evaluation Methods for IMM 41

2.3.1 Classification of Evaluation Methods 41

2.3.2 User Evaluation Methods for IMM 44

2.3.2.1 One-to-one Evaluations 44
2.3.2.2 Small Group Evaluations 45
2.3.2.3 Formative Experimentation 45
2.3.3 Expert Evaluation Methods for IMM 46
2.3.3.1 Early Heuristic Evaluation 46
2.3.3.2 Socio-Constructivist Heuristic Evaluation 46
2.3.3.3 Pedagogical and User Interface Dimensions for the Evaluation of IMM 47
2.3.3.4 Multimedia Taxonomy 48
2.3.3.5 Multimedia Cognitive Walkthrough 49
2.3.4 Combined Approaches for the Evaluation of IMM 51
2.3.4.1 IMM Evaluation Methodology 51
2.3.4.2 A Formative Evaluation Program for IMM 52
2.4 Reviews of the Effectiveness of Evaluation Methods for Instructional Software 54
2.5 Empirical Studies of the Effectiveness of Traditional UEMs 58
2.6 Towards More Effective Expert Evaluation Methods for IMM 65

Chapter 3: A Framework for Measuring the Effectiveness of Expert Evaluation Methods for IMM 67
3.1 Introduction 68
3.2 Summary of Previously-Used Effectiveness Criteria 69
3.3 Creating an EEM Effectiveness Framework 70
3.4 EEM Effectiveness Framework 72
3.4.1 Validity 74
3.4.2 Thoroughness 76
3.4.3 Reliability 78
3.4.4 Diagnostic Power 78
3.4.5 Tailorability to Context 80
3.4.6 Theoretical Validity of Evaluation Criteria 81
3.4.7 Cost Effectiveness 82
3.4.8 Learnability 82
3.4.9 Ease of Use 83
3.4.10 Expert Specialisation 83

3.4.11 Quality of Documentation 84
3.5 Conclusion 84

Chapter 4: Assessing the Effectiveness of Existing Expert Evaluation Methods for IMM 86
4.1 Introduction 87
4.2 Methodological Approach of Expert Evaluations 88
4.2.1 The Expert Evaluation Methods 88
4.2.2 The IMM Software 89
4.2.3 The Experts 90
4.2.4 Protocol of Expert Evaluations 91
4.2.4.1 Familiarisation with EEMs 91
4.2.4.2 Expert Evaluations 92
4.2.4.3 Post-Evaluation Interviews and Questionnaires 92
4.3 Methodological Approach of User Tests 94
4.3.1 The Users 94
4.3.2 Procedure of One-to-one User Tests 95
4.3.3 Post-Evaluation Interviews 95
4.3.4 Before and After Knowledge Tests 96
4.4 Method of Analysis of the Expert and User Evaluation Data 92
4.4.1 Analysis of Expert Evaluation Data 98
4.4.1.1 Refining of Expert Problem Sets 98
4.4.1.2 Assigning Severity to Expert Problems 99
4.4.1.3 Categorisation of Expert Problems 100
4.4.1.4 Establishing the Reliability of Expert Predictions 102
4.4.1.5 Establishing the Diagnostic Power of each EEM 102
4.4.1.6 Analysis of Expert Interviews and Post-Evaluation Questionnaires 103
4.4.2 Analysis of User Evaluation Data 103
4.4.2.1 Analysis of User Test Data 103
4.4.2.2 Analysis of Knowledge Tests 104
4.4.2.3 Assigning Severity to User Problems 104
4.4.2.4 Categorisation of User Problems 105
4.4.3 Validation of Predicted Problems 105

4.4.3.1 Step 1: Validation against User Test Results 106
4.4.3.2 Step 2: Ratification by Independent Instructional Experts 108
4.4.4 Reverse Problem Matching 109
4.5 Results from the Expert Evaluations 109
4.5.1 Predicted Problem Sets 110
4.5.2 Severity of the Predicted Problems 110
4.5.3 Types of Predicted Problems 111
4.5.4 Reliability of Expert Predictions 113
4.5.4.1 Multimedia Cognitive Walkthrough 113
4.5.4.2 Interactive Multimedia Checklist 113
4.5.4.3 Multimedia Taxonomy 114
4.5.5 Diagnosis of Predicted Problems 114
4.6 Results from the User Tests 116
4.6.1 User Problem Set 116
4.6.2 Severity of User Problems 118
4.6.3 Types of User Problems 119
4.7 Validation of the Predicted Problem Sets 119
4.7.1 Results from Validation Step 1 119
4.7.2 Results from Validation Step 2 120
4.7.2.1 Instructional Problems 120
4.7.2.2 Minor Design Faults 121
4.7.3 Final Validation Results 122
4.7.4 Analysis of False Alarms 123
4.7.4.1 Analysis of False Alarms Predicted Using the MMCW 123
4.7.4.2 Analysis of False Alarms Predicted Using the IMMC 124
4.7.4.3 Analysis of False Alarms Predicted Using the MMT 125
4.8 Actual Thoroughness of Predicted Problem Sets 125
4.8.1 Nature of Correctly Predicted User Problems 126
4.8.1.1 Nature and Severity of the Problems Predicted Using the MMT 127
4.8.1.2 Nature and Severity of the Problem Predicted Using the IMMC 128
4.8.1.3 Nature and Severity of the Problem Predicted Using

the MMCW 128
4.8.2 Nature of Unidentified User Problems 129
4.8.2.1 Learning Support 129
4.8.2.2 Comprehension 129
4.8.2.3 Missed Interaction 130
4.8.2.4 Misleading Functionality 130
4.8.2.5 User Dialogue 131
4.9 Discussion of the Effectiveness of the Three EEMs 131
4.9.1 Validity 131
4.9.2 Thoroughness 132
4.9.3 Reliability 134
4.9.4 Diagnostic Power 135
4.9.5 Tailorability to Context 135
4.9.6 Theoretical Validity of Evaluation Criteria 137
4.9.7 Cost Effectiveness 137
4.9.8 Learnability 138
4.9.9 Ease of Use 139
4.9.10 Expert Specialisation 140
4.9.11 Quality of Documentation 141
4.9.12 Summary of EEM's Effectiveness 142
4.10 Need for Enhancing the Effectiveness of EEMs for IMM 143

Chapter 5: Empirically and Theoretically Based Approach to the Improvement of the Effectiveness of Expert Evaluation Methods for IMM 145
5.1 Introduction 146
5.2 Analysis of User Problems while Learning with IMM Software 148
5.2.1 Analysis of User Interaction Problems 149
5.2.1.1 Categories of Cognitive Breakdowns 152
5.2.1.2 Categories of Behavioural Outcomes 154
5.2.1.3 Categories of Performance Outcomes 154
5.2.2 Analysis of the Comprehension Problems 157
5.2.2.1 Definition of Comprehension Problems 157
5.2.2.2 Comprehension Problem Report Format 158
5.3 Effectiveness Limiting Characteristics of the EEMs for IMM 162

5.3.1 Effectiveness Limiting Characteristics 1 162
5.3.2 Effectiveness Limiting Characteristics 2 164
5.3.3 Effectiveness Limiting Characteristics 3 166
5.3.4 Effectiveness Limiting Characteristics 4 168
5.4 Effectiveness Promoting Factors 169
5.4.1 Prediction of Cognitive Breakdowns 170
5.4.2 Prediction of Behavioural Outcomes 171
5.4.3 Prediction of Comprehension Problems 171
5.5 Theoretical Frameworks of Learning with Multimedia 173
5.5.1 Cognitive Frameworks of Learning 173
5.5.2 Cognitive Theories of Learning with Multimedia 177
5.5.3 The Learner of IMM Applications 181
5.6 Hypotheses for Improving the Effectiveness of EEMs for IMM 183
5.6.1 Hypothesis 1 185
5.6.2 Hypothesis 2 186
5.6.3 Hypothesis 3 187
5.6.4 Hypothesis 4 188
5.7 Methodological Approach for Testing the Four Hypotheses 189
5.7.1 Testing Hypothesis 1 190
5.7.2 Testing Hypothesis 2 191
5.7.3 Testing Hypothesis 3 192
5.7.4 Testing Hypothesis 4 193
5.8 Conclusions 194

Chapter 6: Creation and Validation of a New Expert Evaluation Method for Instructional Multimedia 195
6.1 Introduction 196
6.2 Creation of an Expert Evaluation Method for IMM 198
6.2.1 Rationale for Selecting the MMCW 198
6.2.2 Principles of the EMMCW Method 198
6.2.3 Formulation of Analytic Evaluation Questions 201
6.2.4 Introduction of Usability Report Formats 203
6.2.5 Modification of the Evaluation Procedure 205
6.3 Approach of EEM Validation Study 207

6.3.1 Pilot Expert Evaluation 207
6.3.2 The Experts 210
6.3.3 Procedure of Expert Evaluations 210
6.3.4 Post Evaluation Interview 211
6.3.5 Method of Analysis of Expert Evaluation Data 212
6.3.5.1 Refining of the Predicted Problem Set 212
6.3.5.2 Validation of the Predicted Problems 212
6.3.5.3 Detailed Problem Matching 213
6.3.5.4 Comparison of the Results to those from the MMCW 214
6.3.5.5 Comparison of the Results to those from all EEMs Studied 214
6.4 Results from the Expert Evaluations using the EMMCW 215
6.4.1 Number of Problems Predicted 215
6.4.2 Validation of Predicted Problems 216
6.4.2.1 Detailed Problem Matching 217
6.4.2.2 Validation of User Interaction Problems 218
6.4.2.3 Validation of Comprehension Difficulties 219
6.4.3 Prediction of Cognitive Breakdowns 220
6.4.4 Prediction of Behavioural Outcomes 222
6.4.5 Prediction of Performance Outcomes 224
6.4.6 Prediction of Comprehension Difficulties 225
6.5 Discussion of the Effectiveness of the EMMCW 228
6.5.1 Comparison of the Results to Original MMCW 228
6.5.1.1 Discussion of Validity 229
6.5.1.2 Discussion of Diagnostic Power 230
6.5.1.3 Discussion of Actual Thoroughness 230
6.5.2 Comparison of the Results to all Existing EEMS Studied 231
6.5.2.1 Discussion of Validity 231
6.5.2.2 Discussion of Diagnostic Power 232
6.5.2.3 Discussion of Actual Thoroughness 233
6.5.3 General Comments about the Effectiveness of the EMMCW 234
6.6 Implications of Results for Enhancing the Effectiveness of EEMs for IMM 235
6.6.1 Conclusions Regarding Hypothesis 1 235

6.6.2 Conclusions Regarding Hypothesis 2 236
6.6.3 Conclusions Regarding Hypothesis 3 236
6.6.4 Conclusions Regarding Hypothesis 4 237
6.7 Conclusions 238

Chapter 7: Discussion and Conclusions 239
7.1 Summary 240
7.2 Contributions and Implications of Research 241
7.2.1 Producing a Systematic Framework for Measuring the Effectiveness of Formative EEMs for IMM 242
7.2.2 Demonstrating the Effectiveness of Existing EEMs for IMM 245
7.2.2.1 Supporting EEM Selection Decisions 245
7.2.2.2 Demonstrating the Limitations of EEMs for IMM 246
7.2.3 Producing a Systematic Approach to Enhancing EEMs' Effectiveness 247
7.2.3.1 Definition of Effectiveness-limiting and Effectiveness-enabling Characteristics 248
7.2.3.2 Value of the Taxonomy of User Problems 249
7.2.3.3 Review of Theories of Learning with IMM 249
7.2.4 Supporting Effective Expert Evaluation of IMM 251
7.2.4.1 The Value of the Evaluation Questions in the EMMCW for Usability Evaluation 251
7.2.4.2 Demonstrating the Effectiveness of the EMMCW 252
7.3 Limitations of Thesis Research 254
7.4 Future Directions 259
7.4.1 Development of a Theoretical Framework of Learning with IMM 259
7.4.2 Further Improvement of the EMMCW 261
7.4.3 Further EEM Effectiveness Improvement Studies 262
List of Abbreviations 264
Glossary 265
References 268
Bibliography 279

Thesis Figures

Figure 1.1 Sample Instructional Multimedia Applications 19
Figure 1.2: Concurrent Formative Evaluation of IMM Software 27
Figure 1.3: Scope of Thesis Research 32
Figure 1.4: Structure of the Thesis 36
Figure 2.1: Instructional Multimedia Design Process 40
Figure 2.2: Multimedia Taxonomy 48
Figure 2.3: An Extract of an Attentional Graph 50
Figure 2.4: UEM Effectiveness Tree 64
Figure 3.1: Effectiveness Measurement Criteria Previously Used and the Number of Studies in Which They Were Used 69
Figure 3.2: Expert Evaluation Method Effectiveness Framework 73
Figure 4.1: Overview of EEM Effectiveness Assessment 88
Figure 4.2: Sample Screens from the Maths Learning Tool 89
Figure 4.3: Two-step Validation of Predicted Problems 106
Figure 4.4: Number of Problems Predicted per Problem Category 112
Figure 4.5: Experienced Problems by Category 119
Figure 4.6: Validity Calculations of Predicted Problem Sets 121
Figure 4.7: Sample False Alarm predicted by the Experts 124
Figure 4.8: Learner Problems Predicted and Unidentified by the Experts 126
Figure 4.9: Coverage of User Problems 127
Figure 4.10: Sample Screens from the Maths Application Illustrating Typical User Problems Successfully Predicted by the Experts 128
Figure 4.11: An Example of a *Missed Interaction* Not Identified by the Experts 130
Figure 5.1: Hypothetico-deductive Approach to Improving EEM's Effectiveness 147
Figure 5.2: Components of a User Interaction Problem 150
Figure 5.3: Components of a Comprehension Problem 158
Figure 5.4: Limitations in representing Scaling 168
Figure 5.5: Sample Screens Illustrating Inconsistency of Maths Notation Used 172
Figure 5.6: Five Cognitive Phases of Learning 175
Figure 5.7: Components, Processes and Sub-processes of Conceptual Learning 176
Figure 5.8: Overview of MM Cognitive Framework 179
Figure 5.9: Usability Problem Components and the Hypothesis Supporting the Prediction of Each One 184

Figure 6.1: Method of Assessing the Effectiveness of the EMMCW 197
Figure 6.2: Analytic Evaluation Procedure Proposed in the EMMCW 205
Figure 6.3: Comparison of Original MMCW and the EMMCW 206
Figure 6.4: Validation of Problems Predicted Using the EMMCW 217
Figure 6.5: Problem Matching Results 218
Figure 6.6: Distribution of Cognitive Breakdowns 220
Figure 6.7: Distribution of Behavioural Outcomes 222
Figure 6.8: Distribution of Performance Outcomes 224
Figure 6.9: Distribution of Comprehension Difficulties 226
Figure 6.10: Coverage of User Problems per EEM 233
Figure 7.1: EEM Effectiveness Framework and Factors Addressed 243
Figure 7.2: A Multidimensional Framework for the Evaluation of IMM 260

Thesis Tables

Table 2.1 Example Relationships between Usability and Learning Evaluation Heuristics 47
Table 2.2: Example Questions in QUALITY Dimension 49
Table 4.1: Experts Involved Using Each EEM 90
Table 4.2: Propositions Tested in Post-exposure Knowledge Tests 97
Table 4.3: Severity Rating Schema 100
Table 4.4: Problem Categorisation Schema 101
Table 4.5: Criteria for extracting user interaction problems 104
Table 4.6: Additional Problem Categories 105
Table 4.7: Rules for Matching Predicted and Experienced Problems 107
Table 4.8: Scale of Degree of Certainty of Problem Matches 108
Table 4.9: Problems Predicted by the Experts Using Each Evaluation Method 110
Table 4.10 : Number of Predicted Problems per Severity Category 111
Table 4.11: Diagnosed vs. Under-Specified Predicted Problems 114
Table 4.12: Number of User Problems per Problem Identification Criterion 116
Table 4.13: Comprehension Test Results 117
Table 4.14: Number of Experienced Problems per Severity Category 118
Table 4.15: Intermediate and Final Validity Calculations for Each EEM 122
Table 4.16: Reliability Findings for each EEM 134
Table 4.17: Cost-benefit Ratio Calculations 138
Table 4.18: A Summary of the Effectiveness of the three EEMs 142
Table 5.1: User Interaction Problem Components and their Prediction Rate by experts 156
Table 5.2: Conditions for Accepting and Rejecting Each Hypothesis 190
Table 6.1: Problems Predicted by Evaluators using the EMMCW 216
Table 6.2: Comparison of the Validity of Problems Predicted Using the Original MMCW vs. those predicted using the EMMCW 230
Table 6.3: User Data Validation of Predicted Problems for Each EEM 232
Table 6.4: Diagnosed vs. Under-Specified Predicted Problems for all Four EEMs 232

Acknowledgements

I would like to thank my first supervisor, Dr Helen Sharp, for her dedicated assistance, guidance, and encouragement throughout my PhD. I would also like to thank my second supervisor, Ms Stephanie Wilson, for her support, advice and encouragement. I also would like to express my gratitude to the School of Informatics and the Centre for Human-Computer Interaction Design at City University for funding this research. Thanks as well to all multimedia designers, subject matter experts, and students, who took part in the usability evaluations. I also would like to acknowledge the help and support of a number of people at various stages of this research, namely Prof Alistair Sutcliffe, Dr Neil Maiden, Dr Paola Amaldi, Dr Kulwinder Kaur Deol, Dr John Dowell, Dr Mark Ennis, Dr Peter Faraday, Dr Julia Galliers, Mr Alexis Gizikis, Dr Sanowar Khan, Ms Marina Krumbholz, Mr Alistair Mavin, Ms Julia Meek, Dr Shailey Minocha, Ms Raquel Monja, Dr Cornelius Ncube, Ms Sajal Patel, Mr Peter Pavan, Ms Michelle Ryan, and Ms Xiaohong Zhu. Finally, I wish to thank friends and family for their ongoing support.

Declaration

The author grants the power of discretion to the university library to allow the thesis to be copied in whole or in print without further reference to the author.

Abstract

Instructional Multimedia (IMM) applications use a variety of representation media with which the user interacts in order to achieve certain learning goals. IMM software has special needs for usability evaluation techniques beyond the typical desktop, office work applications that the majority of the traditional evaluation methods have been developed for. There have been a limited number of evaluation methods developed specifically for the evaluation of IMM applications. Their effectiveness, however, has not been studied empirically neither in a laboratory nor the field. Thus, no definitive conclusions can be made regarding how effective they are in uncovering valid usability and learning problems, what is the nature of the problems they identify, or how they compare in cost effectiveness.

The first part of this thesis investigates empirically the effectiveness of three expert evaluation methods for IMM. The performance of one cognitive walkthrough, one checklist, and one taxonomy-based approaches was studied against a set of eleven effectiveness criteria, such as method's validity, thoroughness, reliability, and cost effectiveness. The empirical study found that the three methods are not as effective as practitioners and researchers would like them to be, particularly in identifying valid usability problems. The empirical study also highlighted certain characteristics of the expert evaluation methods which limit their effectiveness.

The second part of the thesis deals with improving the effectiveness of expert evaluation methods for IMM. A set of hypotheses for improving the performance of evaluation methods are formulated, particularly for improving their validity. The hypotheses are informed by the results of the empirical study, as well as theoretical work regarding cognitive and pedagogical implications of multimedia design. The hypotheses were tested empirically by developing a new evaluation method addressing the main limitations of the existing methods, which was then applied by expert evaluators. Improvement in the prediction of valid usability and learning problems, particularly those regarding learner comprehension was observed, together with a decrease in the number of false alarms predicted using the developed expert evaluation method.

The thesis fulfils its goal of providing an improved understanding into the effectiveness of existing expert evaluation methods for IMM and defining the characteristics that constitute an effective expert evaluation method for IMM. The results of the research can be used to inform further effort for developing and assessing the effectiveness of evaluation methods for IMM.

Chapter 1

Introduction

Chapter 1

Introduction

1.1 Introduction

Multimedia (MM) is the combination of a variety of representational media, such as text, graphics, sound, speech, animation and video to deliver information to the user via a computer. Multimedia presentations have been used increasing in the past couple of decades to educate and train people. Particularly with the advancement of Information and Communication Technologies, Instructional Multimedia (IMM) solutions delivered online via the Internet or a local Intranet have increasingly been developed. The usability and learning effectiveness of IMM solutions need to be ensured as poorly designed IMM applications have been found to have detrimental effect on students' learning performance and their satisfaction. Having effective design and usability evaluation procedures in place can support IMM developers in making decisions as to how to build and improve the quality of their designs, before they are released to learners. In particular, usability evaluations during early stages of IMM development can ensure that potential usability and learning effectiveness problems have been identified and rectified early in the development process. Different usability evaluation techniques have been developed to aid the process of usability evaluation of IMM. However, to be of real use, such techniques need to be effective.

There are three main problems that make the investigation of the effectiveness of evaluation methods (EMs) for IMM important:

- Firstly, user tests are expensive and time consuming to administer, and learners are not always easily available. IMM is typically developed under tight budget constraint, which makes the production of high quality multimedia a challenge, and often usability evaluation activities can be overlooked. Therefore, alternative usability evaluation methods that are less time-consuming and equally effective need to be investigated and developed.

- Secondly, expert-based evaluation techniques for IMM, such as checklists and guidelines, take less time and resource to conduct, however, they are not widely used in practice. A major reason for this is the unknown validity of the results produced using such EMs, as they have not been subjected to rigorous empirical tests to establish their effectiveness. Therefore, practitioners do not know how effective they are at predicting valid learner problems, what types of problems they are best suited at predicting, and how cost-effective they are. Furthermore, researchers do not know what aspects of the effectiveness of such EMs need to be improved.

- Finally, new EMs for IMM are being developed, however no sound conceptual basis exists for their development in terms of what characteristics enhance their effectiveness and thus are desirable to implement, and what factors constrain their effectiveness and need to be avoided.

In summary, the goal of this thesis is the iterative improvement of the effectiveness of expert evaluation methods (EEMs) for the evaluation of IMM. Firstly, the thesis aims to investigate empirically the effectiveness of existing EEMs for IMM. Secondly, it aims to produce a conceptual approach for improving the effectiveness of such EMs based on the findings from the empirical study. Thirdly, by implementing the conceptual approach, to develop a new EEM for IMM that improves the effectiveness of existing EEMs. Finally, to empirically investigate the effectiveness of new EEM and to measure the extent of the improvement achieved.

1.2 Instructional Multimedia: Definition and Evolution

1.2.1 Definition of Instructional Multimedia

As already mentioned, one area where multimedia is believed to bring benefits is in multimedia applications used for education and training, called Instructional Multimedia applications. Reeves (1993) defines an IMM application as:

> "...a computerised database that allows users to access information in multiple forms, including text, graphics, video, and audio. IMM is specifically designed with linked nodes of information to allow users to access the information

according to their unique needs and interests... The goals of using IMM should be nothing less than fundamentally improving the conditions of teaching and learning in education and training."

One main assumption about IMM applications is that the combination of static media (e.g. text, graphics) and dynamic media (e.g. speech, animation) can provide more effective ways of representing information to the user than a single medium (e.g. text or video alone) (Large et al., 1995; Mayer and Anderson, 1992). The example shown in Figure 1.1 (a) uses a combination of still images and text to represent a model of the electronic structure of an atom, whereas Figure 1.1 (b) shows how animation, text and voice narration are used to represent the concepts of Branding. In addition, many claims have been made about the gains that arise from the ability to interact with multimedia in ways not possible with conventional linear media, such as books and videos (Najjar, 1996; Sims, 1997). One of the main differences is that multimedia allows for the rapid and non-linear access of multiple representations of information, which encourages active exploration of information rather than passive reception (Bell and Johnson, 1993). Therefore, Instructional Multimedia applications, employ a variety of representational media with which the user interacts in order to achieve certain learning goals.

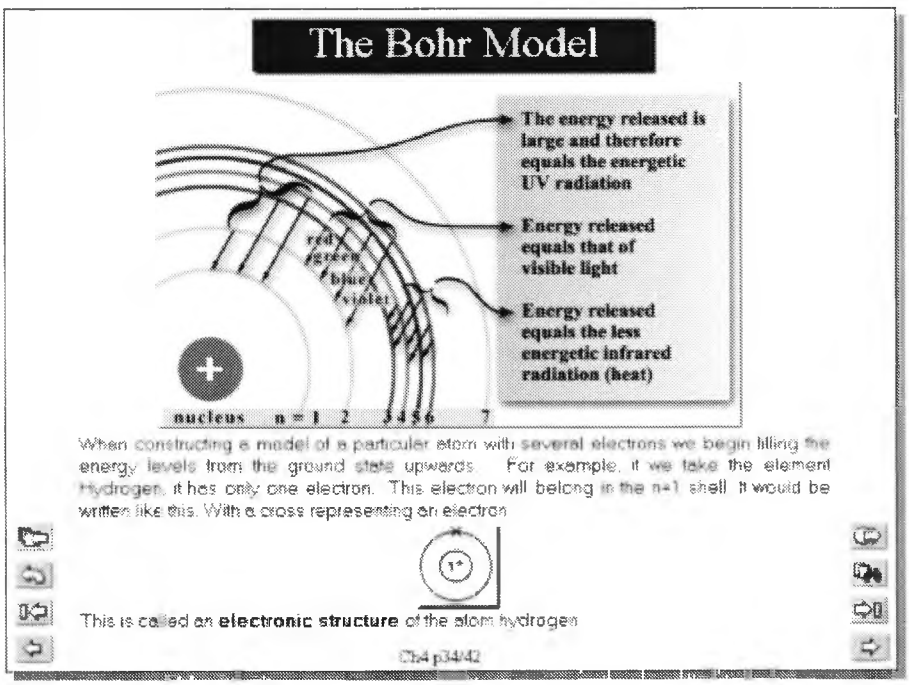
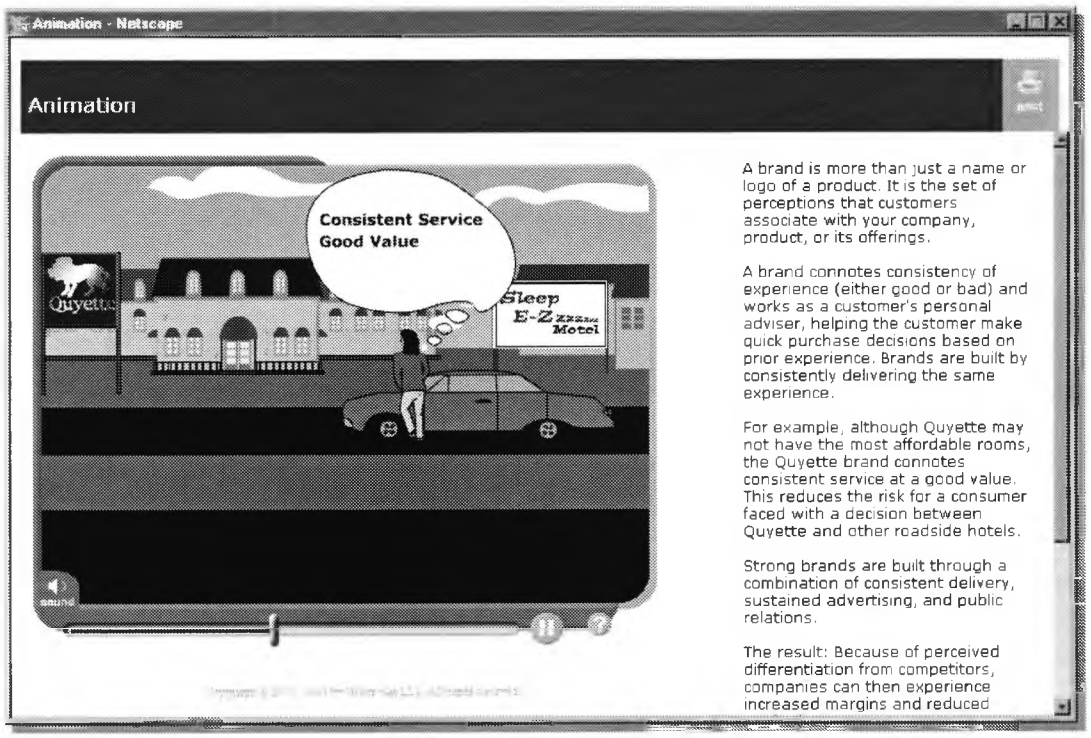


Figure 1.1 (a): A sample screen from "GCSE Chemistry" by Aircom Education © 1998



"A brand is more than just a name or logo of a product. It is ..."

Figure 1.1 (b): A sample screen from "Principles of Marketing: Branding" at www.cardean.edu © 2001

1.2.2 Evolution of Instructional Multimedia

The past few years have seen considerable evolution of IMM and the way it is used for educational and training purposes. The main aspects of this evolution are:

- *Evolution of IMM technologies*

  Over the past couple of decades IMM technologies have evolved considerably from linear drill and practice applications, primarily text based, through interactive video and integrated multimedia CD-ROMs, to highly realistic simulated learning environments. The advancement of information communication technologies (ICT) and the Internet have also made educational software geographically more widely accessible than ever before. This makes possible virtual classrooms, such as those used by Warwick Business School (www.wbs.ac.uk) and Henley Management College (www.henleymc.ac.uk) for administering online tutorials and student group discussions. The advancement of ICT also enables the existence of virtual universities, two examples of which are Capella University (www.capella.edu) and Cardean University (www.cardean.edu). Figure 1.1 (b) shows an example screen

from an online Marketing course provided at Cardean University, which enables learners anywhere in the world to undertake the course at time and place convenient to them.

- *Diversification of user population*

  The diversity of people using educational software, in terms of age, cultural background and abilities, continues to expand considerably with the advent of online learning. Users' expectations from advanced educational software have changed too in regards to the quality of the delivery media and interaction styles, as well as the diversity of the available information and speed of access (Georganas, 1997).

- *Changes of learning tasks*

  Apart from their more traditional educational purposes, learning technologies are increasingly used for life-long learning (Sumner and Taylor, 1998) and on-the-job training. Taken together, the increase in the diversity of learners and the radical changes in learning tasks present significant challenges for the design of IMM software.

- *Emergence of novel interaction styles*

  Furthermore, the novelty of interaction styles based on novel features of the emerging technologies pose the question of their effectiveness. As Sims (1999) emphasises, because they are novel or more advanced does not guarantee their effectiveness, especially their pedagogical efficacy.

- *Change of approach to learning*

  As traditional tutor-driven classroom-based teaching is gradually being superseded by new forms of learning, such as distance and e-learning, learners are becoming more proactive in what they learn and less dependent on immediate tutor support and guidance (Schelin, 2001). Therefore, advanced learning software is required more than ever before to provide adequate learning support, and to keep learners motivated.

All these changes necessitate a change in our attitude towards the effectiveness of advanced educational technologies, as well as the approaches used to measure their effectiveness.

1.3 The Quality of Instructional Multimedia

The quality of IMM is primarily determined by whether the target learners are effectively facilitated in achieving their learning goals, as the confusion resulting from using poorly designed IMM software can be particularly detrimental for learning performance. Therefore, the *learning effectiveness*, as well as the *usability* of IMM, need to be established to ensure high quality of IMM, as software that is easy to use is not necessarily educationally effective (Squires and McDougall, 1996). Due to the integrated nature of learning and interaction with IMM, the usability and the learning effectiveness of IMM cannot easily be separated. Therefore, they need to be considered synergistically.

1.3.1 Usability

Usability is a measure of the quality or the effectiveness of a software product. Traditionally, usability has been defined as

> "... the extent to which a product can be used by specified users to achieve specified goals with *effectiveness*, *efficiency* and *satisfaction* in a specified context of use" (ISO 9241-11, 1998).

It determines the quality of the user experience or the user interaction with the software product. In this traditional context, usability is perceived in terms of the design of the user interface which facilitates efficient and effective completion of well-defined, relatively limited tasks.

The quality of IMM is considered to be more complex than the usability of traditional desktop office applications, as there is a need to assess how effective the design of IMM is in supporting learners in acquiring new knowledge and skills alongside assessing its ease-of-use, ease-of-interaction and user satisfaction (Lee et al., 1995). The traditional view of usability as defined above does not pay attention to the complexity and multi-dimensionality of the learning process during interaction with IMM software. Therefore, the learning effectiveness of IMM also needs to be established.

1.3.2 Learning Effectiveness

Squires and McDougall (1996) and Squires and Preece (1996, 1999) advocate the need to study the interaction between usability and learning effectiveness. Squires and Preece (1999) propose a socio-constructivist approach to evaluating the quality of educational software. In their approach the authors strongly advocate that the synergy between usability and learning effectiveness should be considered when measuring the quality of educational software. In particular, they believe that the integration between the learning processes and the user's interaction with the software needs to be explored for the achievement of learning goals. Their approach is founded on the Constructivist paradigm of learning. Constructivism considers learning as an active, constructive, cognitive and social process by which the learner strategically manages available cognitive, physical, and social resources to create new knowledge by interacting with information in the environment and integrating it with information already stored in their memory (Shuell, 1988; Cunningham, Duffy and Knuth, 1993). The Constructivist view of learning urges the consideration of how learners construct meaning while interacting with IMM software.

Very often, when instructional applications are evaluated, the quality of the user interface design is considered separately from the learning effectiveness (e.g. in Draper et al., 1996). The use of the software is typically considered in isolation from users' learning processes and the outcomes of their interaction with IMM on their learning performance and behaviour is also overlooked. There is no consideration of the implications for user interface features on the use of the package to achieve learning goals. Such separation of usability and learning effectiveness, however, leads to superficial evaluations of the quality of the design of IMM software, as such evaluations overlook the effect of users' interaction on their knowledge acquisition and skill development processes.

Based on the above, the quality of IMM can be defined as a complex multi-faceted measure of the effectiveness of the interaction between the learner, the design of the IMM and the wider context of learning. The quality of the interaction depends to a great extent on the learner and the learning context, not only on the design of the IMM. Therefore, quality is not simply a property of the user interface design. Instead, it is the

interaction between the media design features, the learner, and the context of learning, in which the learner's role is central to determining the effectiveness, efficiency and satisfaction of the IMM software.

1.4 Building Quality Instructional Multimedia

1.4.1 The Importance of Ensuring the Quality of IMM

The quality of IMM software is crucial for its success and the satisfaction of its users, as the confusion involved in using a poorly designed educational application can be particularly detrimental for the learning performance of its users. In the past few years, there has been a huge increase in the development of multimedia systems because it became possible to integrate different types of media on more powerful computer platforms. The initial excitement resulted in the production of a plethora of multimedia systems. The development of such systems, however, paid little attention to their usability, as Laurillard (1993) acknowledges, and therefore it is not surprising that such systems have been criticised for being 'boring, restricted, insubstantial and flashy' (Pham, 1998). To prevent deploying a considerable amount of time, effort and money on producing ineffective IMM software, and to ensure that the software produced meets its learning objectives and satisfies its users, development teams should pay special attention to the usability or the learning effectiveness of the software. As Preece and Shneiderman (1995) emphasised six years ago, the key principles for survival of MM applications are mainly those of good usability and fulfilling a real need. Therefore, measuring and improving the effectiveness of such software products is vital for ensuring their high quality.

1.4.2 A Learner-Centred Approach to the Development of IMM

Norman and Spohrer (1996) advocate a learner-centred approach to constructing effective educational software. At the heart of their approach is the idea that people learn most effectively when engrossed in the topic and when motivated to seek new knowledge and skills they need to solve problems. The goal is active exploration, construction and learning rather than the passivity of lecture attendance and textbook reading.

This thesis proposes four major elements of the design of IMM that are vital to consider in order to ensure the usability and learning effectiveness of IMM software -. These factors are: the *learner* and their characteristics and abilities, the *role of the external media representations,* the *user interaction* with the IMM, and the *context of learning.* Each of these is discussed in the following paragraphs.

- Most importantly, in a learner-centred approach to developing instructional multimedia, the primary focus should be on the *learners* and their needs (Soloway, 1998), their existing knowledge and experience, as well as the learner's pre-determined learning style, and individual approach to learning (Wild and Quinn, 1998). The design of high quality user interfaces should take the user into account during early phases of design (Morris et al., 1994). The target audience for the system must be identified and described so that their physical and cognitive characteristics are understood. According to Jih and Reeves (1992) learning is influenced by three dimensions of individual differences: a) personalistic factors, like prior knowledge and experiences, b) affective factors, such as motivation and attitudes, and c) physiological factors, e.g. eye-hand co-ordination and visual acuity. Furthermore, users' motivation and affective feelings while using IMM are also important (Peterson, 1998), as they have an effect on their learning performance. Chan and Ahern (1999) emphasise that when people are intrinsically motivated to learn, they not only learn more, they also have a more positive experience. The authors argue against the conventional perception that good quality instructional design, especially one employing a number of media, is by itself motivating. Chan and Ahern (1999) claim that multimedia can be 'a double-edged sword for instructional design', as it can confuse students as well as motivate them to learn if not designed appropriately. The authors also criticise traditional instructional design for not considering motivational and affective issues and provide theoretically underpinned evidence that aspects of multimedia design, such as the structure of the learning activity, influence students' motivation. Therefore, it is important to consider the effect of such multimedia design features on learners' motivation to learn and subjective feelings, as a measure of the effectiveness of IMM.

- Secondly, it is also vital to consider the *role of the external media representations* of the subject matter on shaping learners' internal mental models of the content when

exploring the learning effectiveness of IMM. Following the constructivist paradigm of learning, Kozma (1994) suggests that there is a constant interaction between the learner's cognitive resources and aspects of the external environment, of which the multimedia presentation is a major part. A number of authors have proposed that media inherently have characteristics which interact with and influence learners' cognitive processes (Salomon, 1978; Kozma, 1991, 1994; Rogers and Scaife, 1997). Such characteristics include the symbolic system and processing capabilities of each medium, which make them more or less effective for certain learners, learning goals and tasks (Clark, 1994).

- Thirdly, the learners' *interaction with the external MM environment* is vital for shaping their understanding of the subject matter. Multimedia technology offers an unprecedented opportunity to create richly interactive learning environments, promoting active learning (Kirsh, 1997). IMM software needs to employ complex forms of interactivity to suit the learning strategy adopted (Sims, 1997). Therefore, the level and style of interaction should be adequate and relevant to support the learning tasks and to facilitate the acquisition of knowledge and the development of new skills (Squires and Preece, 1996; Sims, 1997).

- Finally, the *wider context of learning* also needs to be considered to ensure the effectiveness of IMM. As part of their socio-constructivist approach to usability evaluation, Squires and Preece (1999) postulate that learning is situated in specific learning contexts. Therefore, the effects of using multimedia information technology on learning will depend on the context in which it is used, with all the components of a learning environment, such as people and artefacts, interacting and contributing to the learning processes.

1.4.3 The Process of Formative Evaluation of IMM

To ensure that an IMM application is of high quality, its usability and learning effectiveness need to be built into its design throughout its development. The process of gathering information for the purpose of validating the design decisions and improving the design of software during its development is called *formative evaluation* (Flagg, 1990). Formative evaluations are conducted throughout the development cycle in a laboratory, at the developer's site, or in the final setting (Heller, 1995). They are used as

a vehicle for obtaining valuable information about potential problems and how improvements can be made to the product's design before it has been released. Formative evaluation can inform developers about potential usability and learning problems early in the development process, thus making it more cost effective to change the software. It is particularly important to identify and fix problems early in the development of IMM, as the cost of redesigning artwork and reproducing video and audio resources can be very substantial.

There are three main aspects to the process of formative evaluation of IMM that are important:

- *The process should be ongoing.*

Northrup (1995) and Thornton and Phillips (1997) propose models for formative evaluation of IMM concurrent with all stages of software development. Northrup's (1995) framework of concurrent formative evaluation is shown in Figure 1.2. The model suggests performing formative evaluation at every step of system development, including analysis, design, development and implementation. For every stage of the development cycle, a set of evaluation techniques is proposed for use by involving different stakeholders in the process.

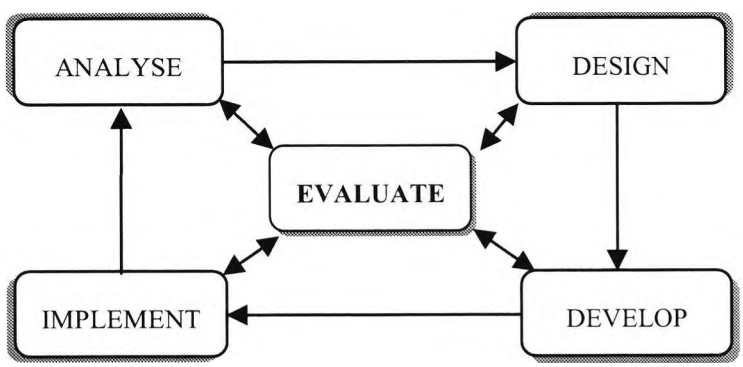


Figure 1.2: Concurrent Formative Evaluation of IMM Software
(Northrup, 1995)

A methodology for developing IMM software featuring iterative design, allowing input from potential users in the early stages of development is also proposed by Morris, Owen, and Fraser (1994). The authors emphasise that ongoing evaluation performed at the end of each stage of development is a

crucial part of such a development process. It can ensure that the project is completed within budget, because design problems can be identified early enough to fix them, before valuable effort and money are expended. Northrup (1995) stresses that if a first version of a product is created before a formative evaluation is conducted, then major modifications will not occur during later stages of development even if they are required. The reason being that too much money, effort and time would have been expended on the development of the product to allow a major rework to take place.

- *The process should be learner-centred.*

As already emphasised, learners should play an important role during the design and evaluation of IMM. Reeves (1993) argues that the quality of the interaction is determined by the skills and experience of the learners with the medium and the degree to which the medium has been designed to support the interaction. Thus, Reeves (1993) emphasises that it is imperative to evaluate IMM software within the context of its use. This implies that the interaction between the learners and aspects of the multimedia user interface should be a major focus for evaluation, as well as the effects of the media design on users' learning processes. This also means that individual differences amongst learners with respect to aptitude, knowledge, skills, attitudes, physical abilities, previous experience, and motivation must be taken into consideration during formative evaluation of IMM.

- *The process should be effective.*

The process of formative evaluation also needs to effectively fulfil its purpose of gathering information regarding the usability and learning effectiveness of the design of IMM, identifying potential usability problems, and supporting designers in making decisions as to how to rectify them. Formative evaluations not only need to identify as many usability and learning problems as early as possible during the product's development, but these problems also need to be *valid* (Sears, 1997). Evaluations need to identify genuine issues, which will impact the target users and their interaction with the application. Furthermore, evaluations also need to be *thorough* (Sears, 1997), focusing on as a wide range of design issues as possible, such as whether the most appropriate media are

selected to represent the content, and whether all media components are integrated effectively to support users in achieving their learning tasks (Morris, Owen, and Fraser, 1994). Finally, evaluations need to help the identification of effective redesign solutions to the problems predicted, which will improve the design of the software, rather than introduce new problems.

The following section reviews existing usability evaluation techniques for IMM.

1.5 Formative Evaluation Methods for Instructional Multimedia

1.5.1 Existing Formative Evaluation Methods for IMM

To ensure the usability and learning effectiveness of IMM software, effective evaluation techniques are required. As stated previously, IMM software has special needs for evaluation methods beyond these of typical desktop, office work applications that the majority of the established usability evaluation methods have been developed for. Evaluation methods for IMM should address the instructional effectiveness of the application in terms of how successful users are in achieving their learning goals, and should verify how effective the multimedia presentation is in supporting the cognitive processes of learning.

Traditional usability evaluation methods (UEMs), such as Heuristic Evaluation (Nielsen, 1993) and Cognitive Walkthrough (Wharton, Rieman, Lewis and Polson, 1994), do not emphasise issues of the learning effectiveness of interactive applications. A limited number of methods developed specially for the formative evaluation of IMM have been proposed to date. They include a number of user-based and expert-based EMs. User-based (or empirical) EMs involve potential users to provide feedback and identify usability and learning problems, and evaluators to reason about the causes of the problems and devise redesign suggestions. Formative experimentation (Reeves, 1992) is an example of a user-based EM for IMM. Expert-based (or analytic) EMs, on the other hand, rely on experts to reason about the effectiveness of the user interface. Example expert EMs for IMM include Interactive Multimedia Checklist (Barker and King, 1993) and Multimedia Taxonomy (Heller and Martin, 1999c). These differ

significantly from each other in the data collection techniques they use and the evaluation results they produce.

The information gathered using such EMs aims to help instructional designers and developers in making the right decisions about the design or implementation of IMM software. The information gathered during formative evaluations plays a vital role for the quality of the software developed (Heller, 1995). Thus, as argued earlier, it is essential to ensure the effectiveness of the evaluation process, and the validity and thoroughness of the results produced. This will greatly depend on the effectiveness of the evaluation methods used.

1.5.2 Assessing the Effectiveness of Existing Evaluation Methods for IMM

As EMs for IMM proliferate, evaluators need to know how to choose between methods for their particular project, and researchers need to know which ones need further development and in what areas to concentrate their research efforts. Unlike the traditional UEMs, evaluation techniques for IMM have seldom been a subject of empirical investigation. Hence, there is very little empirical evidence of their effectiveness, especially in comparison to one another. For example, there is little understanding of how good they are at uncovering usability and learning problems, what kinds of problem they are best suited to identify, and how they compare in cost benefit.

The unknown effectiveness of these EMs is pointed to as a major reason why such methods are not widely used for assessing the quality of IMM applications (Tergan, 1998). Because the validity and reliability of the predictions made using the majority of the EMs for IMM is unknown (Tergan, 1998), there is a danger that the evaluation results produced using such methods can be of limited usefulness. Such evaluations typically make little positive improvement to the quality of the software being developed, while incurring considerable development costs (John and Marks, 1997). This can greatly undermine the credibility of the evaluation process amongst software developers.

Finally, as the strengths and limitations of the existing EMs for IMM are not well understood, researchers and method developers do not know what aspects of the

methods need addressing. This means that no systematic actions for improving the performance or usefulness of such methods can be taken.

Therefore, there is a need for further research into the effectiveness of EMs for IMM. In particular, the validity, reliability and thoroughness of such EMs have to be investigated, as well as their cost benefit and usefulness. This can only be achieved by applying adequate measures and procedures for systematically gathering information regarding how effective EMs are at assessing the usability and learning efficacy of IMM interfaces. Such investigations can reveal the characteristics of EMs that enable their effective or ineffective evaluation performance. Based on such evidence, actions could be taken to enhance the performance of such evaluation methods.

1.6 Thesis Scope and Objectives

This thesis addresses the problem of effective formative evaluation of IMM applications, by investigating the effectiveness of existing expert evaluation methods and proposing ways of enhancing their effectiveness. The scope of the first part of this thesis is the investigation of the effectiveness of existing EEMs for IMM, since little empirical research has been carried out to validate this type of EM. EEMs are also of particular interest because these methods can be potentially very useful in practice as they allow interface designers and other experts to evaluate designs quickly without the need to employ users. The research does not investigate the effectiveness of user-based evaluation methods, as they are fundamentally different from expert ones in the way data is collected and decisions are made regarding potential usability and learning problems. Furthermore, the thesis focuses on the effectiveness of the predictions made using EEMs for *formative* evaluation of IMM, where the primary aim is to gather information regarding the usability and learning effectiveness of such interfaces in order to improve their design before the application is released. Therefore, issues regarding the effectiveness of summative evaluations of IMM are not considered.

During the second part of this thesis, existing cognitive models of learning with IMM and their implications for the effective design of such applications are used to develop evaluation criteria suitable for assessing the usability and learning effectiveness of IMM designs. Although the evaluation of aspects of IMM usability are considered, such as

quality of screen layout and ease of navigation, particular emphasis is given to assessing how effective the design of multimedia presentations is for supporting learner's knowledge construction processes, such as attention to and comprehension of information. The main reason is that none of the existing EEMs adequately assesses the learning effectiveness of multimedia designs, as found in the first part of this thesis. Although, the Constructivist learning paradigm is advocated for creating highly interactive explorative learning environments that promote effective learning, evaluations to test and enhance the effectiveness of EEMs are conducted with a tutorial-style IMM application. It is important to consider how the effectiveness of such applications can be assessed, as a considerable proportion of the instructional software developed in industry is still based on such a pedagogical approach, rather than on Constructivist learning paradigm. Finally, the research is more applicable to IMM for single-user learning, as issues of how IMM can support group learning are not considered. Figure 1.3 shows the scope of this thesis research, as described above.

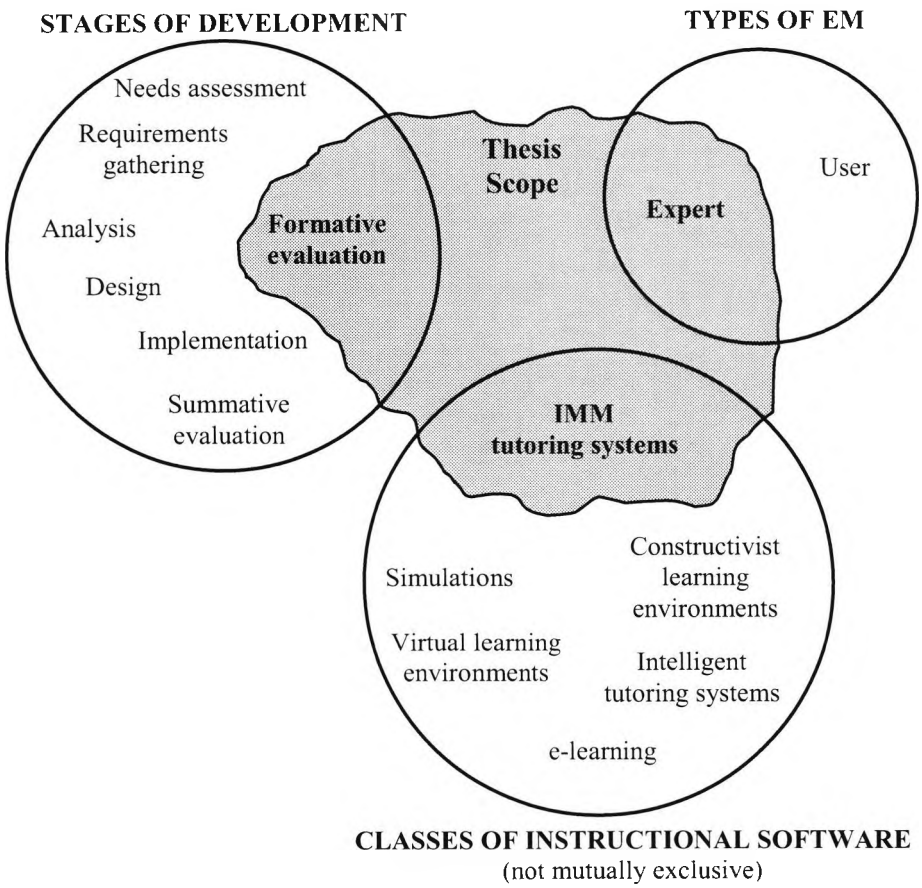


Figure 1.3: Scope of Thesis Research

The primary aim of this PhD thesis is to study the effectiveness of existing EEMs for formative evaluation of IMM, and to develop a method for formative evaluation of IMM applications which improves on the effectiveness of the existing EEMs.

To achieve this aim the following five objectives were set:

Objective 1: To define a framework for measuring the effectiveness of formative expert evaluation methods for IMM.

Objective 2: To measure empirically the effectiveness of existing EEMs for IMM using the effectiveness framework created in Objective 1.

Objective 3: To formulate hypotheses for enhancing the effectiveness of EEMs for IMM based on the empirical findings produced in Objective 2 and on theories of learning with IMM.

Objective 4: To develop an EEM for IMM, which aims to improve the performance of the existing EEMs by implementing the characteristics specified in the hypotheses defined in Objective 3.

Objective 5: To validate the effectiveness of the EEM developed as part of Objective 4.

1.7 Research Methods Used

To achieve objective 1, existing criteria used for measuring the effectiveness of traditional UEMs are reviewed and adapted to address specific characteristics of the evaluation process for IMM. Furthermore, method effectiveness requirements found in literature on the effectiveness of existing evaluation methods for Computer-Based Learning software are used to formulate new effectiveness measurement criteria.

The method for achieving objective 2 involves conducting two controlled evaluation studies of the usability and learning effectiveness of a commercially-developed IMM CD-ROM. The first study applies existing EEMs using user interface and subject matter

experts to reason about the quality of the IMM and predict potential usability and learning problems. The second study involves performing user tests of the same IMM application with representative users. One-to-one user tests are conducted involving think-aloud protocols and retrospective interviews to gather data regarding the usability problems that exist in the interface. 'Before' and 'after' comprehension tests are also used to identify learning problems users encounter. The evaluation reports produced by the experts are validated against the problems experienced by the users and ratified by two independent judges. Finally, the effectiveness of the EEMs is measured and compared using the framework formulated in the previous objective.

A hypothetico-deductive approach is adopted in order to formulate hypotheses for enhancing the effectiveness of EEMs for IMM (objective 3). The approach includes deduction of research hypotheses from the findings of the empirical study conducted in the previous objective and from existing cognitive models of learning with multimedia. The hypotheses define desirable EEM characteristics and relate them to potential performance outcomes.

Objective 4 is achieved by integrating the desirable EEM characteristics specified in the hypotheses into an EEM for the evaluation of IMM interfaces. Evaluation criteria are formulated based on existing guidelines for effective IMM design and theoretical research into effective use of IMM for learning.

Finally, the method for testing the effectiveness of the developed EEM (objective 5) involves applying the EEM to evaluate the same commercial IMM CD-ROM using interface design and subject matter experts. The evaluation predictions made are first compared to the user test results to establish the ability of the new EEM to predict valid user problems. Secondly, the results are compared to those produced using the existing EEMs to establish whether improvement has been achieved in terms of the evaluation performance of the new EEM.

1.8 The Organisation of the Thesis

This thesis consists of seven chapters. The next chapter, Chapter 2, describes existing research relevant to the thesis. It includes a review of existing techniques suitable for the formative evaluation of IMM, what is known about their effectiveness, and details of how the effectiveness of traditional UEMs is measured, including criteria and measurement procedures used. Chapters 3 through 6 then describe the research conducted to meet the thesis objectives.

Chapter 3 defines a framework for assessing the effectiveness of EEMs for IMM. It describes eleven effectiveness measurement criteria, together with procedures for assessing each criterion. Chapter 4 describes how the effectiveness of three existing EEMs for IMM was investigated. Two studies were conducted, the first of which presents expert evaluations of an existing commercial IMM application using the three EEMs. The second study presents user tests of the same IMM interface, the results of which are used to validate the predictions made using the three EEMs. The methods' performance is assessed and compared in terms of each of the criteria defined in the effectiveness framework. Conclusions are then drawn as to their effectiveness and usefulness for the formative evaluation of IMM interfaces.

Chapter 5 presents an analysis of the usability and learning problems identified in the user tests. The main characteristics limiting the evaluation power of the three EEMs are revealed together with performance enhancing characteristics, identified from the findings of the expert evaluations described in Chapter 4. A review of existing cognitive models of learning with multimedia is also given, after which four hypotheses are formulated as to how to enhance the ability of existing EEMs to predict valid user problems.

In Chapter 6 the four hypotheses are tested by firstly integrating the desirable EEM characteristics specified in the hypotheses into an expert evaluation method. The effectiveness of the developed EEM is empirically tested using four experts, and conclusions are drawn regarding the improvement achieved over the performance of the existing EEMs, whose effectiveness was studied in Chapter 4.

Finally, Chapter 7 summarises this research and concludes with a discussion of its implications and possible future research directions. The research provides an improved understanding of how effective three EEMs are for the evaluation of IMM interfaces, and the characteristics that enable and constrain the effective performance of such methods. Future directions include further development of the new EEM, as well as the development of a more systematic theoretical framework of learning with IMM.

Figure 1.4 shows the structure of the thesis.

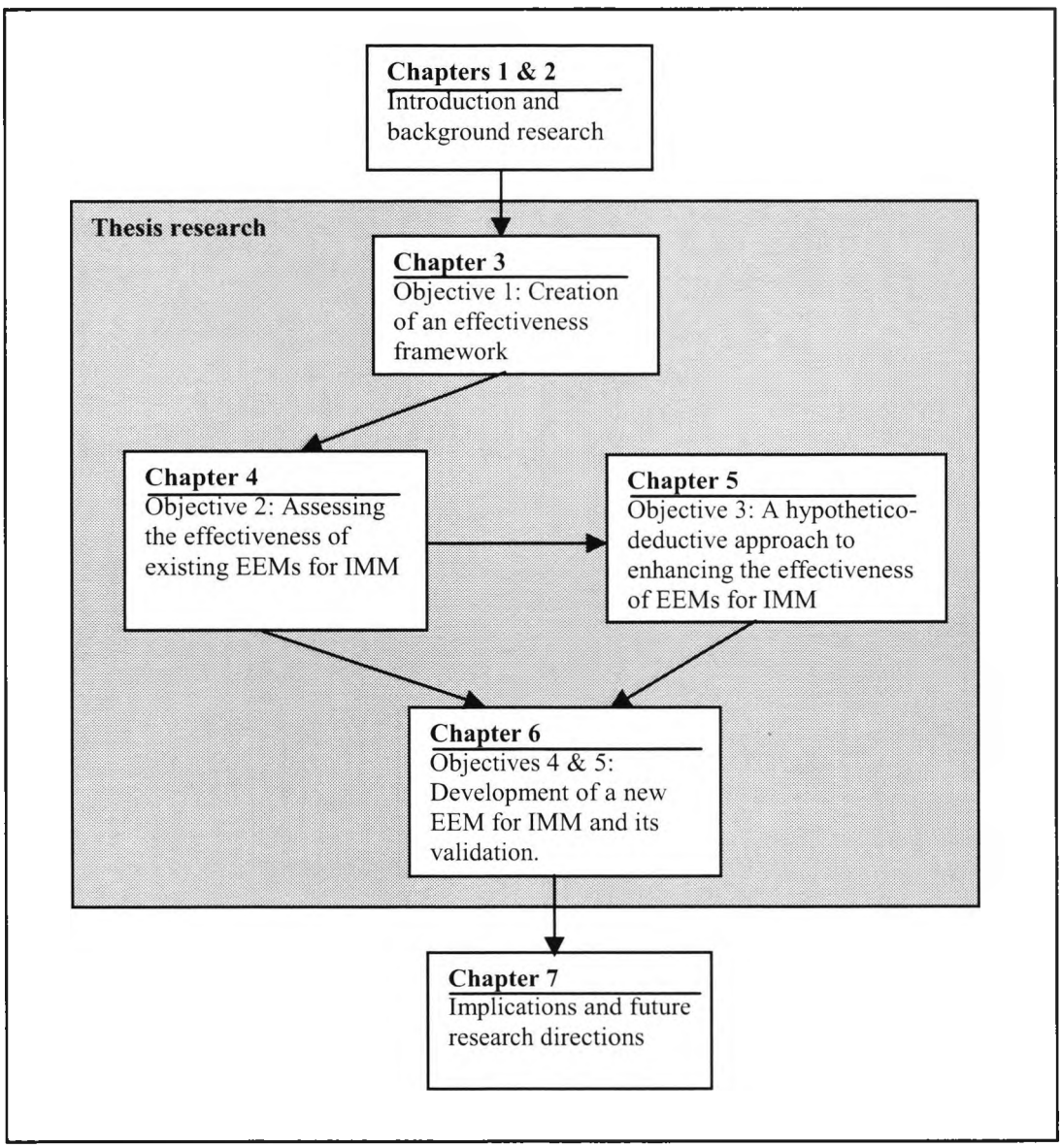


Figure 1.4: Structure of the Thesis

1.9 Summary of the Thesis Contribution to the Field

This thesis makes contributions to research in usability evaluation methods for IMM and to the practical usability evaluation of IMM. The major contributions are:

- A systematic framework for evaluating the effectiveness of usability evaluation methods for IMM that can be used by practitioners and researchers in the field to assess the quality of such methods, particularly expert ones.
- Substantial empirical evidence of the effectiveness of three existing EEMs for IMM, revealing their strengths and limitations. The evidence promotes understanding of their effectiveness and usefulness, and can provide pointers to researchers in the field as to where improvements are urgently needed. The findings can also inform usability practitioners of the costs and benefits of employing such methods so they can decide which ones to use.
- A thorough analysis of the performance limiting and performance enhancing characteristics of EEMs for IMM, which can be of use to method developers to improve the effectiveness of newly-developed EEMs.
- A taxonomy of user problems which promotes understanding of the nature and types of problems users encounter while learning with IMM. The taxonomy can be used by instructional software evaluators to describe the problems experienced by users during user tests in a uniform and complete matter.
- An empirically- and theoretically-based approach to the improvement of the effectiveness of EEMs for IMM, particularly of their ability to predict valid usability and learning problems.
- An improved version of an expert usability evaluation method for IMM, which can be used to effectively predict valid usability and learning problems of IMM interfaces.
- A demonstration of the predictive power of the developed EEM, verifying its benefits for the usability evaluation of IMM and demonstrating that the effectiveness of the existing EEMs can be enhanced based on the conceptual approach adopted.

Chapter 2

Formative Evaluation Methods and their Effectiveness

Chapter 2

Formative Evaluation Methods and their Effectiveness

2.1 Introduction

The importance of ensuring the usability and learning effectiveness of IMM applications produced was emphasised in the previous chapter. To attain high quality IMM applications that satisfy the needs of their learners and adequately support them in achieving their learning goals, effective procedures for designing, evaluating, and implementing such software are required. Formative evaluation during early stages of development can ensure potential usability and learning problems are identified early in the process and rectified effectively before major development costs have been incurred. Mack and Montaniz (1994) emphasise that usability analysis conducted by usability engineers and designers in particular can play a prominent role in user-centred design. Such expert usability evaluations are a valuable option when a company is lacking the funds for conducting user tests. Expert usability evaluations have an advantage of reducing the cost and time of usability activities, and the results produced can play a vital role in developing quality software products. Therefore, it is essential to ensure the effectiveness of expert evaluation methods (EEMs) intended to support the formative evaluation process of IMM interfaces. Evidence of the validity and reliability of expert predictions made using such EMs is required in order to inform usability and instructional design practitioners and researchers of the effectiveness of these methods, who wish to use or further develop such techniques. Furthermore, practitioners would also like to know how such EMs compare in cost-benefit and usefulness to be able to decide on the most appropriate method to use for the needs of IMM development projects.

This chapter firstly presents an overview of the current practice of formative evaluation of IMM. Section 2.3 then reviews existing techniques for assessing the quality of IMM software. Emphasis is given to expert-based EMs as they are the focus of this thesis

research. The essential characteristics of each technique are discussed, together with the approach of generating each method, highlighting whether it has been founded in practical experience or theoretical frameworks of effective instructional design. Section 2.4 presents comparative reviews of the usefulness of some of the EEMs for educational software. Section 2.5 surveys empirical studies of the effectiveness of traditional UEMs, emphasising on the effectiveness criteria used.

2.2 Current Practice of Formative Evaluation of IMM

A study of the current practice of developing and evaluating IMM applications was carried out. Semi-structured interviews were conducted with eight multimedia instructional designers from four educational multimedia organisations within the UK. The designers were asked to describe the processes carried out for the development of a recent IMM project, in which they took part. Although the design methods varied across companies and individual projects, the following four main stages were typically performed: Specification, Conceptual Design, Prototyping and Presentation Design, which are illustrated in Figure 2.1.

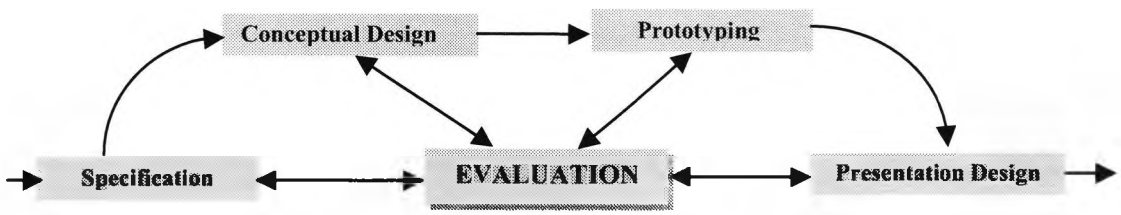


Figure 2.1: Instructional Multimedia Design Process

As can be seen from Figure 2.1, evaluations are performed after each stage of development, however they consisted predominantly of reviews of the specification and design documentation. They were conducted by subject matter experts, external teachers or clients with the aim of approving their contents. In half of the cases users were involved to assess prototypes of IMM software. That was done in an informal manner, and the feedback collected usually consisted of high-level comments and recommendations regarding the colour scheme or the look of the graphics. Only in a third of the cases users were observed while working with a prototype to gather more concrete feedback regarding the design of the IMM interfaces and identify usability problems. Checklists and in-house created guidelines were used in half of the cases,

however predominantly towards the end of the development for acceptance testing. The checklists contained questions mainly regarding the quality of individual media, such as the spelling of text, or the quality of video material.

From the study, it appears that the current evaluation practice lacks a systematic quality engineering approach. Systematic ways of gathering usability and learning effectiveness data are rarely employed, and no formal evaluation methods are used. Most of the designers were not aware of any formal evaluation methods specific for IMM, although a limited number of such techniques exist, as will be reviewed in the following section. One reason why such EMs are not used in practice could be the lack of evidence about their validity, effectiveness and cost-benefit beyond the projects they have been developed for.

2.3 Review of Existing Evaluation Methods for IMM

As highlighted in Chapter 1, Instructional Multimedia applications differ from other types of software in that they use a variety of representation media, with which the user interacts, in order to achieve certain learning goals. The quality requirements of such software differ from those of commercial desktop applications, as they necessitate the investigation of the learning effectiveness of the IMM design alongside other usability characteristics. IMM software, therefore, has special needs for evaluation methods beyond what is required for typical desktop, office work applications that most traditional UEMs have been developed for. This view is also supported by Peterson (1998), who acknowledges the need for special methods for assessing IMM systems.

A number of methods, suitable for evaluating IMM applications, have been developed to date. The following sub-sections present a review of the main EMs available.

2.3.1 Classification of Evaluation Methods

Evaluation methods can be two types – user or expert. These differ significantly in the data collection techniques used and the evaluation results produced.

- *User Evaluation Methods*

These methods involve potential users to provide feedback and identify usability problems, and evaluators to reason about the causes of the problems and devise redesign suggestions. The data collection techniques used include a variety of user tests, such as interviews, questionnaires, observation, pre- and post-tests and record analysis. These methods concentrate either on gathering users' attitudes towards various aspects of the interface or on measuring their task performance. In the first case, high-level user feedback is typically generated (e.g. 18% of the students indicated that the software was difficult or frustrating to use, Watkins et al. 1995), and the evaluators are left to infer what aspects of the material frustrated the learners. In the second case usability problems are uncovered by observing the users interacting with a prototype of the software (e.g. difficulty in seeing system response, Kaufman and Lee 1993). Again the evaluators are left to infer the possible reasons for the problems.

User EMs require the participation of at least a few users, and thus rely on users' availability. Such methods also can be time consuming in terms of planning and conducting the user test, as well as analysing the data gathered (Heller, 1995).

- *Expert Evaluation Methods*

These methods rely on experts to reason about the effectiveness of the user interface. Typically, they examine specific features of the interface or aspects of the interaction, and attempt to infer usability and learning problems. Analytic or expert evaluation methods are usually easier to conduct, and require less time and resources to administer. EEMs can be either question-based or guideline-based. The first category presents experts with a set of questions regarding important issues of interface design. In most cases experts are not given a rating scale to measure the degree of fulfilment of the criteria, thus the approaches rely on the expertise of the evaluators to assess the user interface and infer potential problems. Guideline-based EEMs encapsulate design expertise regarding user interface features. The experts are required to review features of the user interface against the guidelines. The evaluation results produced by the two types of EEM differ. The question-based approaches generally identify limitations of user interface components and the content, and some usability

problems. In comparison, guideline-based approaches typically reveal design problems by analysing different interface features. In addition, design recommendations are generally easier to provide using design guidelines.

A considerable number of evaluation methods have been proposed in the past few years, however only a few of these are suitable for assessing the usability and learning effectiveness of IMM. The following four categories of evaluation methods are distinguished:

- *Traditional Usability Evaluation Methods*

A number of methods have been developed to date, some of which are based on formal analysis of user interfaces or performance prediction, such as TAG: Task-Action Grammar (Payne and Green, 1989) and GOMS: Goals, Operators, Methods, and Selection Model (Card, Moran, and Newell, 1983). Other techniques involve various kinds of reviews and inspections of different aspects of user interface design, which include Heuristic Evaluation (Nielsen, 1993) and Cognitive Walkthrough (Wharton, Rieman, Lewis, and Polson, 1994). These evaluation techniques are more suitable for assessing the usability of desktop, office work applications. They do not directly address aspects of the learning effectiveness of IMM designs, and therefore are outside the scope of this thesis research.

- *Evaluation Methods For Educational Software*

A number of evaluation techniques have been developed for assessing the effectiveness of educational software. Some examples include the software evaluation model proposed by Zahler, Reiser, Dick and Gill (1992), a methodology for software evaluation developed by Comer and Geissler (1998), and six evaluation designs described in Foshay (1999). These methods propose models for assessing different aspects of learning software, including the clarity and depth of the content, and the adequacy of the instructional approach adopted. Although suitable for the evaluation of IMM, these methods do not consider issues pertinent to the effectiveness of the MM and their effect on the user's learning processes. Therefore, these methods are not specific enough to be considered in this thesis research.

- *Evaluation Methods for Multimedia Software*

A number of methods for assessing the quality of multimedia presentations have also been proposed, including Multimedia Cognitive Walkthrough (Faraday and Sutcliffe, 1997) and MUMMS Questionnaire (1996). Some of these techniques consider the cognitive processes involved while using multimedia user interfaces, including comprehension and knowledge formation, which makes them suitable for evaluating IMM. Such methods will be reviewed later in this section. On the other hand, other multimedia evaluation methods concentrate on issues of the media design, the synchronisation between dynamic media resources, user control of media, and how much the product captures user's emotional responses. Such techniques, including the MUMMS Questionnaire, do not consider aspects of the learning effectiveness of the user interaction with multimedia presentations, which makes them unsuitable for a thorough evaluation of IMM applications. Thus, they will not be covered in the review which follows.

- *Evaluation Methods for Instructional Multimedia Software*

A limited number of evaluation methods specifically for the evaluation of IMM software have been proposed to date. These are the main focus of this thesis and are reviewed in the following section.

The following is a review of existing user and expert EMs for the evaluation of IMM software.

2.3.2 User Evaluation Methods for IMM

The following is a review of the main user EMs that can be used to gather usability data about IMM interfaces.

2.3.2.1 One-to-one Evaluations

Involve an evaluator monitoring a learner working through learning tasks using an instructional application. As described by Byrum (1992), the evaluator observes the user interaction while recording any verbal and non-verbal responses relevant to the instruction. After the learner has completed the task, the evaluator administers a post-

test to gather comments and redesign suggestions from the learner. Because an evaluator is in contact with the learner during the evaluation, learning problems can be identified while the subject is using the material. However, this method can be time consuming and is subject to the idiosyncratic responses of individual learners and evaluators in specifying interaction problems and feedback.

2.3.2.2 Small Group Evaluations

This method is a field testing approach, which can be used to collect learner data (Byrum 1992). Typically, a group of learners work together on a task, and their interaction with the software and discussions are recorded. The evaluator analyses the data and the error patterns, identifies possible causes and decides on revisions of the design. Therefore, an extensive amount of learner data is collected before any revision is made. Evaluators also examine the scores on the embedded and post-test items to identify usability problems. Data is collected from a wide range of users in ability and interest, and also provides a setting closer to the intended use of the software. However, as there is no face-to-face interaction between the learners and the evaluator, it is difficult to pinpoint the exact cause of learners' problems with the instruction.

2.3.2.3 Formative Experimentation

Reeves (1992, 1993) proposes this method as an approach for formative evaluation of IMM software. Formative experiments take place in a real setting with meaningful instructional goals. Such experiments aim at a particular outcome and observe the process by which the learning goal is achieved. The primary data collection techniques are ethnographic ones, such as interviews, observations and record analysis. The evaluators use a number of unobtrusive monitoring procedures, like automated response-capture routines to track the progress of learners, and *before* and *after* observations to measure learners' performance. In addition, the evaluators can assess the environment in which the IMM is used by the learners. Unobtrusive observations of learner behaviour are also suggested, especially if the IMM is to be used by two or more learners collaboratively. Capturing the nature of interactions among learners can be useful for understanding their perceptions of IMM and the type of cognitive processes in which they engage. Finally, questionnaires and interviews can be used at various points in the experiment to provide information about specific aspects of the IMM and its integration into the training environment.

2.3.3 Expert Evaluation Methods for IMM

A number of expert approaches have been developed to date, ranging from usability heuristics and guidelines, to evaluation checklists and walkthrough techniques. These methods differ in the scope and depth of the evaluation criteria used for measuring the effectiveness of the IMM software. However, they all rely on experts to reason about the effectiveness of the IMM design. The following is a review of the main approaches developed.

2.3.3.1 Early Heuristic Evaluation (Squires, 1997)

This heuristic approach aims to help expert teachers in "...assessing the quality and potential uses of a software application prior to its use with students". A set of heuristics is presented. Each heuristic contains a set of relevant evaluation questions, which expert evaluators should answer depending on the context of use of the software. The questions cover a mixture of issues regarding the learning effectiveness of the application, user control and navigation. Example questions are: 'How appropriate is the content to the curriculum?', 'Are learners motivated when using the software?', and 'How navigable is the software?' The approach is based on a review of evaluation criteria used by expert practitioners and theoretical work on designing effective IMM. No evidence has been presented of the formal validation of how valid or reliable are the predictions made using the method.

2.3.3.2 Socio-Constructivist Heuristic Evaluation (Squires and Preece, 1999)

The authors adapt the notion of heuristic evaluation introduced by Nielsen (1993), and relate the ten usability heuristics proposed to socio-constructivist criteria for learning. The aim is to support evaluators in considering the integration between usability and learning issues during usability evaluations. The authors propose 19 interactions between Nielsen's heuristics and the instructional concepts of cognitive and contextual authenticity. The concept of cognitive authenticity suggests that during learning experiences students are assisted in some way to construct and refine conceptions in ways meaningful to them. The notion of contextual authenticity, on the other hand, implies that learning experiences are situated in specific learning contexts, and are influenced by components of the learning environment, such as people and physical artefacts. Three example interactions are presented in Table 2.1.

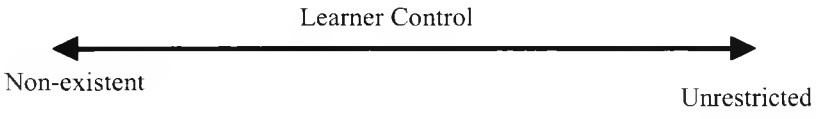
Table 2.1 Example Relationships between Usability and Learning Evaluation Heuristics (Squires and Preece, 1999)

	Cognitive authenticity			Contextual authenticity	
	Credibility	**Complexity**	**Ownership**	**Collaboration**	**Curriculum**
System status	Feedback & designer/ learner models	Navigation			
Consistency		Symbolic representation		Consistent protocols	Subject content
Error recovery	Interaction flow	Pedagogical techniques	Metacognition		

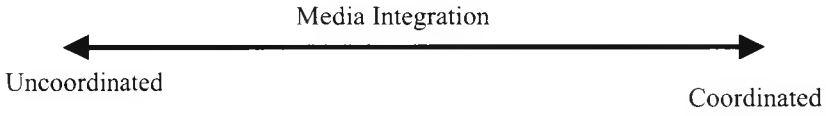
The authors describe all interactions, however, do not propose a specific approach to applying them in practical usability evaluations. This method is theoretically grounded, but as the authors acknowledge, the interactions between the concepts have not yet been validated empirically.

2.3.3.3 Pedagogical and User Interface Dimensions for the Evaluation of IMM (Reeves and Harmon, 1994)

The authors propose fourteen pedagogical and ten user interface dimensions for the systematic evaluation of IMM. The pedagogical dimensions are concerned with those aspects of the IMM design that directly affect learning. They aim to evaluate the effectiveness of the capabilities of IMM to enable effective learning interactions, the techniques for monitoring learner progress, whether the design accommodates individual learner differences and promotes co-operative learning. Some pedagogical dimensions include pedagogical philosophy, instructional sequencing, motivation, and learner control. Each dimension can be assessed by instructional specialists on a scale of two extreme values. An example of such a scale is given below:



The user interface dimensions are concerned with those aspects of IMM design that ensure the learner can engage in a meaningful interactive experience with the application. Example user interface dimensions include ease of use, media integration, navigation and cognitive load. Each dimension is again represented as a continuum described by two extremes, as shown next.



As the authors acknowledge, the dimensions are firstly not comprehensive, and secondly their validity has not been tested formally to establish the effectiveness of the evaluations produced using the dimensions.

2.3.3.4 Multimedia Taxonomy (Heller and Martin, 1999a, 1999b, 1999c)

The Taxonomy, depicted in Figure 2.2, is a three-dimensional categorisation framework of multimedia issues, which can be applied in both design and evaluation of multimedia applications. The first dimension represents different *Media Types*, including Text, Graphics, Sound, Motion and Multimedia. The *Expression* of each medium makes up the second dimension, and is represented by four categories - General, Elaboration, Representation, and Abstraction- ordered in increasing level of abstraction. Finally, the third dimension, *Context*, comprises six discrete categories: Audience, Purpose, Interactivity, Quality, Usefulness, and Aesthetics. For each of these categories a list of specific attributes can be drawn. For instance, the category of Audience could focus on attributes such as age, gender, and cultural background.

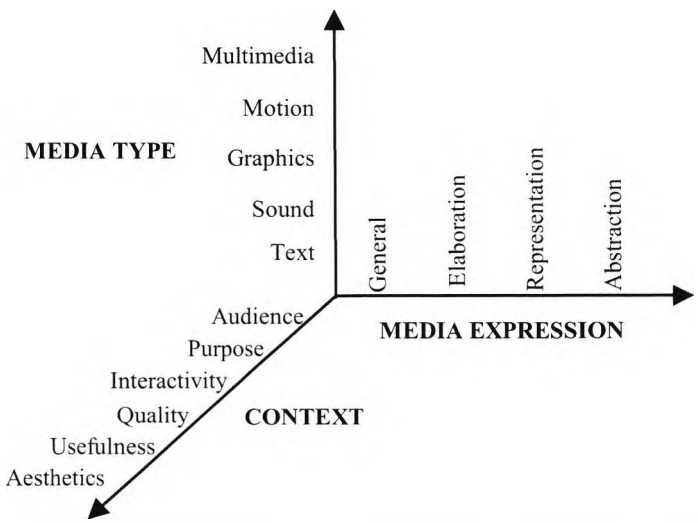


Figure 2.2: Multimedia Taxonomy (Heller and Martin, 1999a)

The taxonomy contains 120 cells, each of which is intended to address specific issues of media design or questions for evaluation. Based on that information, evaluators from different backgrounds are required to generate 'evaluation protocols' by asking

questions in relevant cells of the taxonomy. For instance, using the attributes of the Quality category, the evaluator could ask a number of questions, as shown in Table 2.2.

The Taxonomy is based on theoretical work and observations of its practical application. As it is three-dimensional it considers the interaction between different media types, their expression and certain contextual factors. Although the taxonomy considers a great deal of MM issues, as it is at the moment it is not complete, and not every cell is populated with relevant questions. The Taxonomy has been tested in practical use, based on which it has been refined (Heller and Martin, 1999c).

Table 2.2: Example Questions in QUALITY Dimension (Heller and Martin, 1999c)

Media Type / Media Expression	Text	Sound	Graphics	Motion	Multimedia
General	Word choice was clear/unclear? Grammar was correct?	Sound quality was unclear to clear? Sound volume was too loud to just right?	Quality of image was low to high? Size of image (too large/small to just right)?	Quality of motion was jerky to smooth?	Media integration (uncoordinated to coordinated)?

2.3.3.5 Multimedia Cognitive Walkthrough (Faraday and Sutcliffe, 1997)

A very limited amount of work has been done to develop usability evaluation methods specifically designed to validate the usability of multimedia presentations at a cognitive level. In their work, Faraday and Sutcliffe (1997) have developed a cognitive walkthrough method for evaluating MM presentations. The proposed method is to be used by MM designers. This cognitive walkthrough approach is intended to review how effective multimedia presentations are in supporting users' cognitive processes, such as attention to, perception and comprehension of information presented concurrently in a variety of media. In their analysis of the user interface, evaluators should follow three steps, which are based on a model of the cognitive processes of comprehension of multimedia presentations created by the authors. The steps are described below.

Step 1: Evaluating Attentional Design

Aims to explore which media will be attended to and in what order, thus seeks to reveal potential problems in sequencing and timing of media. An 'attentional graph' is introduced as a means of analysing the sequence of attention to each

medium in the presentation, and a set of guidelines are provided to evaluate its effectiveness. Example attentional graph is given in Figure 2.3.

Step 2: Evaluating Contact Points:

The purpose is to test how well visual and verbal media are combined together. This is achieved by adding 'contact points' to the attentional graph, which represent the points at which visual and verbal media co-refer. Guidelines are provided to validate the effectiveness of the contact points.

Step 3: Evaluating Media Selection and Combination

Validates how the content maps onto information types, and whether the media used in the presentation are suitable for representing the information. Media selection guidelines suggest what types of media should be used to represent different information types.

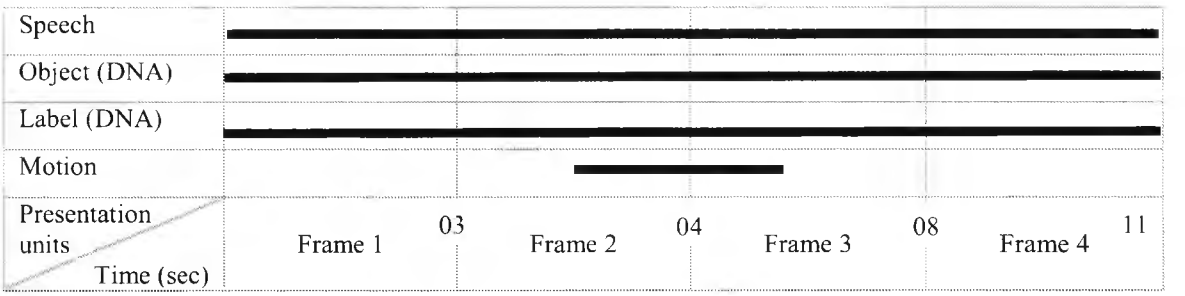


Figure 2.3: An Extract of an Attentional Graph (Faraday and Sutcliffe, 1997)

This method aims to reveal aspects of the design which can cause cognitive problems to the users, such as distracting their attention from important information. The framework emphasises on cognitive aspects of using MM presentations, and contextual factors, such as user domain knowledge, are not considered when predicting potential problems. The proposed guidelines do not deal with aspects of user interaction with and navigation within the system, e.g. navigation within multimedia information, or user control of dynamic media. The approach also does not allow for the evaluation of high-level issues, such as user attitude to media, media biases, aesthetics. The method is based on empirical studies the authors have conducted of human cognitive processes, but has not been validated itself.

2.3.4 Combined Approaches for the Evaluation of IMM

A limited number of researchers have developed more systematic methodologies for evaluating IMM systems. They incorporate user and expert data gathering techniques to be employed either together or at different stages of the development life cycle.

2.3.4.1 IMM Evaluation Methodology (Barker and King, 1993)

A methodology for evaluating IMM was developed by Barker and King (1993) as part of the ILDIC (Integration of Learning Design in Interactive Compact Disc) project. The aim of the methodology was to evaluate a number of IMM products in order to extract pointers of good quality designs to enable the building of well-designed software. The method is also suitable for informal, small scale evaluation, which could focus on innovation, locate usability problems, and could provide results applicable to the design and development of educational software. The approach is category-based, and is comprised of 12 categories, which embody essential principles of good instructional design. These evaluation categories are:

- Engagement,
- Interactivity,
- Tailorability,
- Appropriateness of media mix,
- Mode and style of interaction,
- Quality of interaction,
- Quality of end-user interfaces,
- Learning styles,
- Monitoring and assessment techniques,
- Built-in intelligence,
- Adequacy of ancillary learning support tools,
- Suitability of single user/group/distributed use.

The authors suggest a checklist of 90 questions distributed amongst all categories, which encompass the instructional design features of each evaluation category. A suggested question to assess a system's interactivity is: 'What level of interaction does the product support – passive, trivial, active, reactive, excessively active?' Sample questions to assess the appropriateness of the multimedia mix include 'Do the

multimedia techniques complement one another or compete for the user's attention?' and 'Is the multimedia mix on screens used responsibly or does over-crowding of multimedia objects reduce the necessary effectiveness of some media?'.

The authors suggest two methods of assessment. The first one involves different types of expert, such as media and content specialists, reviewing aspects of the design and answering only the questions from the checklist that they consider relevant. The second method involves user evaluations using expert and novice users. Expert users could be asked to perform several tasks using the system, and then complete the checklist. The second user approach involves one-to-one evaluations with novice users, during which an evaluator can collect comments and verbal observations from the users, and then complete a simplified version of the checklist.

The evaluation questions are compiled based on learning theories and practical experiences of IMM evaluation. They encompass a significant amount of learning, instructional, and user interface issues. Some of the questions include relationships between related design issues. As the questions are high-level, regarding the entire application rather than specific aspects of it, the method relies on the expertise of the evaluators in assessing particular aspects of the software being evaluated.

2.3.4.2 A Formative Evaluation Program for IMM (Kennedy, 1999; Kennedy, Petrovic and Keppell, 1998)

A systematic method for formative evaluation of IMM is proposed by Kennedy (1999), to meet the needs of projects within the Faculty of Medicine at the University of Melbourne. In their work Kennedy, Petrovic and Keppell (1998) present a set of usability evaluation criteria, based on theories of learning and on evaluation criteria which researchers in the department have used in past projects. The evaluation criteria aim to evaluate the design and development of educational software, and cover aspects of MM instruction and medical education. The criteria are classified into three domains, which are briefly described below:

- *Instructional and Conceptual Design.*

  This domain is made up of five categories, including Introductory Objectives, Navigation, Interactivity, Sequencing, and Consistency between learning

objectives and content of instruction. An example of a navigation criterion is: "the system should usually allow students to 'exit', go 'forward', go 'back', go to the 'main menu'".

- *Interface and Graphic Design.*

The criteria in this domain cover aspects of the user interface, such as colour, frames, and media types. Each of these interface components are evaluated by asking specific questions covering issues relating to usability, consistency, clarity, structure, relevance, and usefulness. For example, the authors suggest that 'thought should be given to the integration of these media'.

- *User Attitudes and Affect.*

In this domain the researchers propose asking students a number of questions which aim to determine users' perception of IMM at both general and specific level. These questions cover issues such as effectiveness, degree of interest, appeal, and degree of engagement.

The evaluation criteria consist of high-level instructional and user interface principles, it is not evident how media resources should be selected or effectively integrated, and what the impact of multimedia on users' cognitive processes is likely to be. Although the authors suggest that the most important information should be given prominence, they do not specify how users' attention could be shifted to such information.

The proposed evaluation methodology consists of four stages of evaluation to be integrated with four stages of the development cycle, as follows:

- *Stage 1: Story Board.*

The evaluation consists of internal reviews of instructional and conceptual design, and interface and graphic design. The reviews include written reports and verbal discussions, and are carried out by content and graphic design experts and instructional designers.

- *Stage 2: Graphic Design.*

Specified as stage 1.

- *Stage 3: Alpha Version.*

  In this stage, reviews are conducted by content experts, graphic design experts, instructional designers, who are not associated with the project, and users.

- *Stage 4: Beta Version.*

  In this final stage, users' attitudes and affect is evaluated by means of questionnaires, focus groups and observations.

There are two main groups of participants suggested: experts (including content, graphic design, and educational) and potential users. The authors explain that some criteria could not be reliably evaluated by users while others could not be reliably reflected by experts.

The proposed methodology for usability evaluation is very comprehensive, including a variety of techniques and considering a number of different perspectives. However, as the questionnaires used in the reviews are based on the evaluation criteria discussed above, the feedback produced is likely to be at a very high level. According to the methodology, user evaluations should be performed at a later stage of development. Thus, considerable modifications may not be easy to implement, and would probably be at a greater cost. Although, the methodology and the evaluation criteria defined have been tested in practice, there is no evidence that the effectiveness of the evaluation results has been formally investigated.

2.4 Reviews of the Effectiveness of Evaluation Methods for Instructional Software

As evaluation techniques such as those discussed in the previous section proliferate, designers need to know how to choose between methods for their particular project, and researchers want to know which ones need further development and in what areas to concentrate their research efforts. These practitioners require information about the effectiveness of the results produced using such EMs, especially in comparison to one another. Evidence of how good these methods are at uncovering usability and learning

problems, what kinds of problems they are best suited to identify, and how they compare in cost benefit, is particularly important.

There have been very few attempts to critically review and compare existing evaluation methods for instructional software. Four such reviews are presented below.

- A review of various evaluation forms used to evaluate instructional software was performed by Schueckler and Shuell (1989). The authors compare nineteen evaluation forms based on the number and the nature of the evaluation criteria they use. They also discuss how appropriate and comprehensive each set of criteria is for assessing the effective use of instructional software. The authors point out that the evaluation criteria used by such methods focus more on the technical characteristics of software, than on the pedagogical effectiveness, as fewer than half of the techniques contain any criteria regarding such issues. Thus, the authors conclude that the biggest limitation of the forms is that they do not address principles of effective teaching and learning in an adequate manner. Another critique of the techniques is that the overall ratings of the quality of software are usually based on the subjective impression of the evaluators rather than the evaluation criteria. Therefore, it is concluded that such judgements are general and have limited usefulness for improving the quality of instructional software.

- Based on thirty journal articles, Reiser and Kegelmann (1994) review key features of both expert and user evaluation methods for instructional software. Review criteria include who is responsible for rating the software, the nature of the evaluation process proposed by each technique, and what design features are evaluated. The authors found that although several of the EMs call for the collection of learner attitude data, very few require the evaluators to examine how much students learn from a particular application. The evaluation techniques are also criticised for having poor reliability, as the majority of them require evaluators to make subjective judgements about a variety of factors. Reiser and Kegelmann (1994) then suggest ways of improving the evaluation techniques. They suggest that the learning effectiveness predictions can be improved by involving users more in the evaluation process in order to measure student knowledge before and after using the

software, and to collect attitude data about the learners' perception of the instructional software.

- Squires and McDougall (1996) review and critique checklist-based evaluation methods. The authors point out a number of problems with checklist approaches including that it is difficult to indicate the relative weighting of checklist questions, and that they do not allow for different instructional strategies. The authors conclude that such problems are symptomatic of the failure of such techniques to adopt a situated perspective of the use of educational software, as aspects of the educational context in which the software will be used are not considered. Furthermore, Squires and McDougall (1996) point out that checklists typically focus on the software application as an object of evaluation in its own right, rather than evaluate its use by typical learners. Another limitation found is that the diversity and complexity of the target learners is not reflected in the design of checklists. The authors propose a more situated approach to the evaluation of instructional software.

- Finally, Tergan (1998) also examined the strengths and limitations of checklist approaches. The author firstly summarises some key features of checklist-based approaches, such as the scope of their questions and the type of data gathered. Then, the author critically examines these methods by pointing out that two central problems with checklist-based evaluations are the unknown reliability and theoretical validity of their criteria. In particular it is pointed out that items belonging to the same category may be rated differently by different reviewers, jeopardising the reliability of the results. Furthermore, as the criteria are not typically founded on empirical evidence, the validity of the predictions made remain unknown. Another important limitation discussed is the methods' shortcomings for assessing learning effectiveness. Tergan (1998) emphasises that it is due to the single-dimensionality of the questions regarding the effectiveness of the software for enhancing learning, disregarding the complexity of the interaction between the design features and learner characteristics. Furthermore, the inadequacy of checklist questions is viewed as a primary reason for the limited predictive power of evaluation checklists. To improve the validity of evaluations regarding the learning effectiveness of education software, the author proposes that evaluation criteria need to be based in empirically validated models of instructional design and theoretical frameworks of learning.

Finally, Tergan (1998) points out that lack of guidance for tailoring the evaluation criteria to the individual interests of a project can potentially make the method difficult to use.

Although the above reviews provide useful information about the strengths and limitations of different evaluation techniques, none of the authors have carried out user tests to substantiate the conclusions reached. Thus, the conclusions made about the methods' validity and reliability are speculative in nature. Empirical evidence is very important for establishing the validity and the reliability of EEMs, as they can provide useful indicators regarding improving the credibility of these methods. The reviews also do not give concrete details of the cost benefit of the evaluation methods.

Furthermore, the improvements suggested in the reviews are only hypothetical, as none of the researchers specifies how such recommendations can be built into the design of evaluation methods, nor how implementing them will improve the performance of EMs. Most importantly, the suggestions made are not substantiated with empirical evidence illustrating that implementing such features will in fact improve the performance or the outcomes produced by the evaluation methods.

In order to gain insight into how effectively EMs perform when applied to the evaluation of IMM, empirical evidence of their performance is required. Appropriate effectiveness criteria need to be applied based on the empirical evidence. Based on this evidence, researchers can then propose adequate measures to improve the performance of such EMs.

Empirical measures of effectiveness have been defined and applied to measuring the effectiveness of traditional UEMs, such as Heuristic Evaluation and GOMS. A number of empirical investigations into their validity are discussed in the following sub-section, together with an account of the aspects studied and the measurement criteria used.

2.5 Empirical Studies of the Effectiveness of Traditional UEMs

Traditional UEMs such as Heuristic Evaluation (Nielsen, 1993), GOMS (Card, Moran, and Newell, 1983) and Cognitive Walkthrough (Wharton, Rieman, Lewis, and Polson, 1994), have been subjected to empirical tests by a number of practitioners and researchers in the past few years. The empirical studies have produced a significant amount of data regarding the strengths and limitations of these evaluation methods according to different aspects, and have proposed ways of improving their effectiveness. The following paragraphs review the main studies conducted, emphasising on the aspects measured.

- As early as 1991, a group of HCI practitioners including Jeffries, Miller, Wharton, and Uyeda compared empirically Heuristic Evaluation, Usability Testing, Guidelines and Cognitive Walkthrough methods (Jeffries et al., 1991). The authors studied a limited number of factors, including the *number* and the *severity of problems* found by the various methods, as well as the *cost-benefit* of each technique when applied to the evaluation of a visual interface for the UNIX operating system. Based on these quantitative results, the authors revealed the comparative advantages and disadvantages of each method, and made suggestions for improving these techniques.

- Karat, Cammpel and Fiegal (1992) investigated the relative effectiveness of empirical Usability Testing and individual and team Heuristic Walkthroughs in identifying usability problems in two graphical user interface office applications. The results produced by each method were compared based on the *number, the type* and *severity of the problems* uncovered by each evaluation technique. The authors also analysed the proportion of the problem areas that were common across methods and those that were unique for each method in order to establish when each method is most effective. Finally, the *cost-effectiveness* of the three methods was studied and compared, based on the ratio of the total time to prepare, administer and analyse the data to the total number of problems identified.

- Nielsen and Phillips (1993) compared the ability of User Testing, Heuristic Evaluation and GOMS to estimate users' performance of tasks when using two alternative user interface designs of a database query system. The aim of the study was to establish the benefit of applying each type of UEM for comparing alternative

designs. To achieve that the authors studied the *cost/benefit* of each method. They proposed measuring the costs of using each method by calculating the number of people involved and the time they spent, and multiply that by their salary rates. The methods' benefits, on the other hand, were calculated by studying each method's estimations of the time needed for users to perform certain tasks with each alternative design, and then compare the estimates to actual user performance in the field, and measure the standard deviation of the estimates. According to the authors such analysis would reveal which design is likely to bring savings of time in executing tasks, which can then be converted into monetary terms to calculate potential savings. Based on the *cost/benefit* analysis the authors concluded that performance estimates from both Heuristic Evaluation and GOMS were highly variable compared to actual user performance in the field. Nonetheless, the authors overall conclusion was that there was still value of applying usability inspection methods for benchmarking of user interface designs. However, as the authors themselves acknowledge, measuring UEMs' *cost/benefit* in such a way is very difficult as there is no way of quantifying the value of having time-on-task estimates of a certain accuracy.

- Dutt, Johnson and Johnson (1994) assessed Heuristic Evaluation and Cognitive Walkthrough for the evaluation of a recruitment system for handling job vacancies and job applications. The aim of their study was to assess the effect of using each UEM at different stages of the software development process. The methods were compared based on the *number*, *nature* and *severity of the problems* identified, the *time it took to use* each technique and their *ability to generate requirements for re-designing* the application being evaluated. Although the authors emphasise that a usability inspection method should be able to generate requirements for effective re-design of the application, they do not provide a precise description or indicators of how to establish whether this ability is present or not. Therefore, the conclusions regarding the UEMs' ability to generate re-design solutions are inconclusive. Furthermore, some suggestions for improvements to both UEMs are proposed, however they tend to be of general nature and their effect on the UEMs' performance is not explained.

- An experimental study comparing Heuristic Evaluation and Usability Testing for the evaluation of an information retrieval interface was performed by Doubleday et al. (1997). The authors first compared the *number* and *severity of the problems* predicted to those experienced by users. After that, both problem sets were compared to establish commonalities and differences in the *number* and *types of problems* predicted and observed. The cost of evaluation was measured in terms of the time taken to apply each method and to analyse the results produced, which was juxtaposed to the significance of the outcomes in order to assess the *cost-effectiveness* of each method. However, the procedure for establishing the *cost-effectiveness* was not clearly defined.

- Zhang, Basili and Shneiderman (1998) also studied the Heuristic Evaluation and it was compared to Perspective-Based Usability Inspection. The authors compared the *number* and *the type of problems* predicted by each usability evaluator. The scope of the study was limited as its aim was to validate the Perspective-Based Usability Inspection method developed by the authors by only comparing the number and types of problems predicted by the new method to those predicted by Heuristic Evaluation. Thus, the study does not provide an independent and thorough assessment of the evaluation techniques.

- More recently, Woolrych and Cockton (2000) performed a detailed analysis of the *number, type* and *severity* of the problems predicted using Heuristic Evaluation, by comparing its results to the problems found during user testing. The aim of the study was to reveal what kind of problems Heuristic Evaluation can reliably predict, and those that cannot be identified using this method.

- Methodological empirical measures of UEMs' effectiveness were formulated by Bastien and Scapin (1995), who define three important characteristics that should be considered when comparing evaluation techniques, namely: *validity, thoroughness* and *reliability*. Using the three criteria, Bastien and Scapin (1995) compared the results produced using a set of ergonomic criteria to those produced using evaluators' expertise, in an attempt to establish how useful ergonomic criteria are for the evaluation of user interfaces. The authors define each of the three criteria as follows:

Validity

Validity is considered to be the evaluators' ability to focus on specific, predefined aspects of the interface design. The *validity* was measured by counting the number of problems predicted by both groups of evaluators. The authors assume that all problems predicted by the experts were legitimate usability problems.

Thoroughness

A method is considered thorough if it allows the examination of as wide a range of interface aspects as possible. The *thoroughness* was established by comparing the proportion of problems uncovered using ergonomic criteria to those identified using evaluators' expertise.

Reliability

Finally, the authors define that a UEM is reliable if evaluations using it provide the same results under the same conditions. The *reliability* was assessed based on the number of problems found in common across all individual evaluators. Hertzum and Jacobsen (1999) also studied empirically the same aspect however they called it *robustness*. The authors examined the robustness of predictions made using the Cognitive Walkthrough method by measuring the extent to which evaluators detected the same problems in a user interface.

The proposed measures give numeric values of an EM's effectiveness without qualitatively explaining the factors that have contributed to the method's performance. Therefore, they are useful for comparative analysis of EMs' effectiveness, however reveal little information based on which the design of EMs can be improved.

- Lavery and Cockton (1995, 1996) assessed the quality of two usability inspection methods, namely a form of Heuristic Evaluation for the assessment of software visualisations developed by the authors and Task Analysis. The methods were applied to two software visualisations for a programming environment. To assess the methods' quality the authors studied their *effectiveness* and *validity*. *Effectiveness* was defined as the proportion of problems predicted to those revealed during user testing, whereas *validity* was measured by the number of false alarms, which is the

number of problems predicted by expert evaluators but not experienced by users. The authors emphasise that simply counting the number of problems identified by expert evaluations is not a legitimate measure of their *validity*. Instead, to validate expert predictions they need to be compared against a known set of real problems in the design, usually those experienced by typical users performing typical tasks. Determining whether usability problems are valid or not based on whether they have been observed by users or not is an important step towards validating usability problems, as many researchers and usability practitioners in the past have assumed that all problems predicted by evaluators are valid by default. A number of other authors, such as Bailey, Allan and Raiello (1992), and Bailey (1999), also determined the *validity* of predicted problems by establishing whether the usability problems were actually experienced by users during user tests or not. Even in studies of more general expert-based human error prediction methods, their *validity* was measured by comparing the errors identified by experts with those observed by users (Stanton and Stevenage, 1998).

- Numeric ratios for measuring of the *validity, thoroughness* and *reliability* of UEMs were defined by Sears (1997) in the context of traditional inspection-based evaluation techniques. The author applied these criteria to assess the effectiveness of five techniques: Heuristic Evaluation, Cognitive Walkthrough, Usability Walkthrough, Ergonomic Criteria, and Heuristic Walkthroughs. Sears (1997) suggested that:

  Validity

  As shown below, the *validity* is measured as a ratio of the number of 'real' usability problems identified to all issues predicted as usability problems. 'Real' usability problems are considered those that were identified by users during user tests.

$$\text{Validity} = \frac{\text{\# Real Problems Found}}{\text{\# Issues Identified as Problems}}$$

  This approach to validating predicted usability problems is very similar to the one used in Lavery and Cockton (1995) and Bailey (1999).

Thoroughness

To measure the *thoroughness* the ratio of real problems that are identified to the number of problems that exist in the system is proposed by Sears (1997). However, the denominator is very difficult to calculate.

$$\text{Thoroughness} = \frac{\text{\# Real Problems Found}}{\text{\# Real Problems that Exist in the System}}$$

A similar measure of methods' thoroughness was proposed by Kirwan (1992) and Stanton and Stevenage (1998), who called it *comprehensiveness*. According to these authors *comprehensiveness* is concerned with the breadth of coverage of the technique, i.e. it measures whether the method identifies all errors that exist or only a specific subset.

Reliability

Finally, Sears (1997) proposes that *reliability* be measured by the ratio of the standard deviation of the number of problems found to the average number of problems found across all evaluators.

$$\text{Reliability} = 1 - \frac{\text{Stdev (\# Real Problems Found)}}{\text{Average (\# Real Problems Found)}}$$

The ratios proposed by Sears improve and standardise the measures proposed by Bastien and Scapin (1995), and they provide a useful way for comparing methods' performance. The outcomes of such comparisons are however quantitative, in the form of numeric values. For example, in his study, Sears found that the validity of the Heuristic Evaluation was 0.69 for groups of five evaluators, whereas that of the Cognitive Walkthrough method was 0.91. Such ratios do not provide details of which aspects of EMs limit their performance, and based on such results it is, therefore, hard to suggest any measures to enhance methods' effectiveness.

- John and Marks (1997) conducted a comprehensive assessment of the effectiveness of five traditional UEMs: Claims Analysis, Cognitive Walkthrough, GOMS, Heuristic Evaluation, and User Action Notation. The authors assess the *predictive*

power, persuasive power and *design-change effectiveness* of each evaluation method using 'an effectiveness tree', which is given in Figure 2.4.

Starting from left to right on Figure 2.4, the *predictive power* is measured first by comparing the predictions made by experts to results of user tests in order to establish whether a predicted problem has been observed by users or not. This measure is very similar to the one proposed by Lavery and Cockton (1995), Sears (1997), and Bailey (1999) for calculating methods' *validity*. The methods' *persuasive power* is then measured by establishing whether the problems identified by the UEMs led to changes in the system or not. Finally, the *design-change effectiveness* is established by testing the new version of the system with users, after changes have been made based on the experts' predictions, in order to verify whether improvement has been made in terms of whether the problems have been rectified or not. John and Marks (1997) believe that the most desirable outcome of expert evaluations is the very top one, #01 in Figure 2.4. Therefore, according to the authors expert evaluations are most effective when the predicted problems will be observed in real-world use if not fixed, if their results provide motivation for the development team to change the design, and the changes implemented actually rectify the problems.

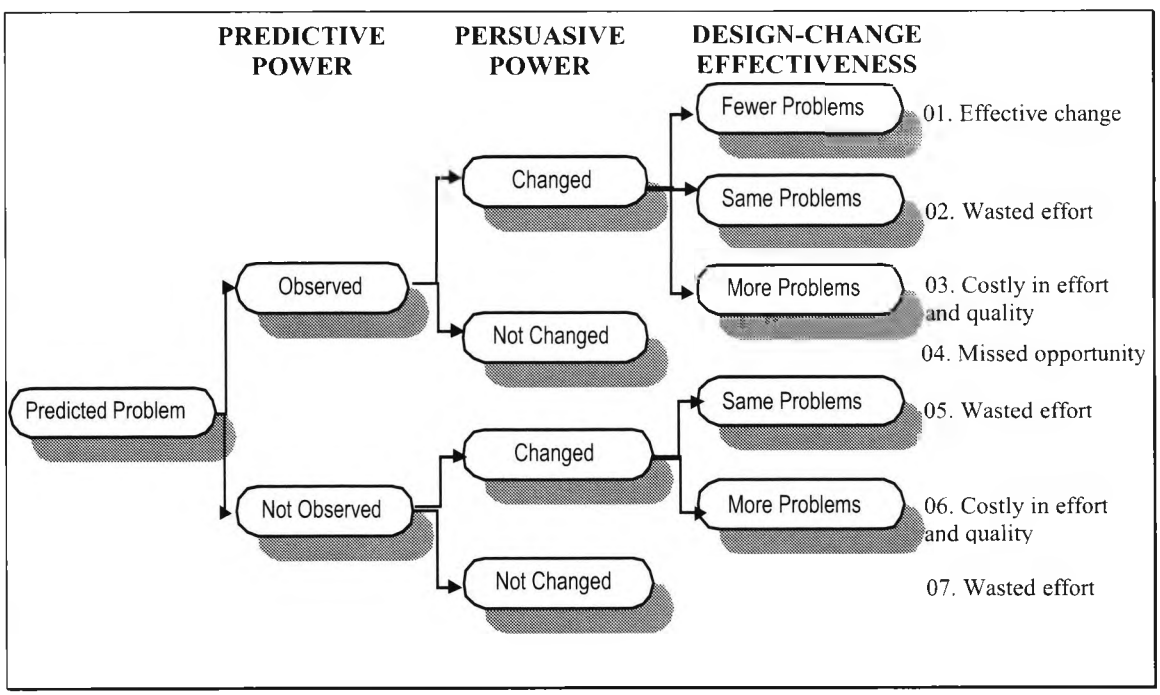


Figure 2.4: UEM Effectiveness Tree (by John and Marks, 1997)

This is a comprehensive approach to assessing the effectiveness of UEMs as it involves re-assessing the design changes made as a result of usability predictions in order to verify their effectiveness. It thus tracks the effect of usability evaluations on the quality of the software through the iterative cycle of design and re-design. The approach, however, provides limited insight into other important parameters of UEM's performance, such as reliability and cost-effectiveness. Performing such tests is also time-consuming and relies on the availability of evaluators, developers and users.

2.6 Towards More Effective Expert Evaluation Methods for IMM

Instructional Multimedia software continues to be extensively used to support or be the sole provider of teaching and training, extending the concepts of learner interaction and engagement in authentic learning environments. The effectiveness of such software in supporting learners' knowledge construction and skill development processes is of primary importance, as the frustration involved in using poorly designed IMM programs can be particularly detrimental for the learning performance of students. As redesigning artwork and recreating video and audio footage can be very expensive, as early in the design as possible it needs to be ensured that different media will deliver the content effectively and will support users in their learning tasks.

Although traditional UEMs can be used to evaluate the design of IMM user interfaces, they do not support the assessment of their learning effectiveness. There are only a limited number of EEMs specifically intended for the evaluation of IMM. None of these methods, however, have been empirically validated or established as effective in providing adequate support in usability evaluation during formative stages of development. Their ability to predict valid usability and learning problems also has not been proven. EEMs do not directly involve target users. Instead, they rely on the expertise and subjective judgement of experts to predict potential user problems. These problems may or may not be experienced by real users, they also may not cover all problems which users will encounter. Without empirical studies to establish the validity and thoroughness of such methods, the true value of performing expert evaluations will remain unknown. Furthermore, to improve the effectiveness of existing EEMs and to ensure that new ones built will be effective, a conceptual approach informed by such

empirical studies is required. Therefore, thorough empirical studies are required to validate EEMs for IMM in a number of dimensions.

The following chapters describe the thesis research into assessing and improving the effectiveness of Expert Evaluation Methods for IMM. Next chapter defines a framework for measuring the effectiveness of such evaluation methods.

Chapter 3

A Framework for Measuring the Effectiveness of Expert Evaluation Methods for IMM

Chapter 3

A Framework for Measuring the Effectiveness of Expert Evaluation Methods for IMM

3.1 Introduction

The effectiveness of evaluation methods is complex to determine, as it depends on a number of interdependent criteria. In order to establish how effective a formative evaluation method is, a comprehensive set of effectiveness measurement criteria is needed. However, there is no evidence in the literature that such set has been developed for the expert evaluation of IMM to date.

Therefore, to fulfil Objective 1 of the PhD research:

> *Objective 1:* To define a framework for measuring the effectiveness of formative expert evaluation methods for IMM.

an effectiveness framework was developed from a synthesis of literature on the effectiveness of traditional usability evaluation methods, as well as material on the effectiveness of existing evaluation methods for Computer-Based Learning software. The aim of the framework is to provide a systematic way of assessing the overall effectiveness of EEMs for IMM, which can serve as a uniform basis for comparison of the relative effectiveness of such EMs. In such a way the strengths and limitations of EEMs can be highlighted in respect of their ability to support the formative evaluation of IMM applications.

This chapter first presents a summary of the effectiveness criteria used in previous studies assessing the effectiveness of traditional UEMs, which were discussed in Chapter 2. After that, in Section 3.3 the approach adopted to formulating the

Effectiveness Framework is presented. Finally, a detailed description of each effectiveness criterion is given in Section 3.4.

3.2 Summary of Previously-Used Effectiveness Criteria

Although all studies presented in Chapter 2 investigate the performance of various usability evaluation methods, some are more comprehensive than others. All studies, however, reveal different strengths and limitations of the techniques investigated using various measures of effectiveness.

A total of 14 criteria of the effectiveness of traditional UEM's were used in the studies reviewed in Chapter 2. As some of them duplicate one another, such duplicating criteria were combined together. For example, the *validity* and *the number of problems* predicted were combined, as the latter assumes that only valid problems were counted. The *problem types* measure was incorporated in the *thoroughness* by most definitions, therefore they were considered as one measure. Finally, the *time to apply* a method was combined with its *cost-effectiveness*, as it is considered to be part of this measure. As a result, ten distinct measures of effectiveness of UEMs were identified, which are shown in Figure 3.1 together with the number of studies in which each criterion was used.

# of occurrence	Effectiveness Criterion
13	*Validity*
9	*Thoroughness*
5	*Reliability*
4	*Cost-Effectiveness*
3	*Theoretical validity*
1	*Tailorability to context, Ability to assess instructional efficacy, Design-change Effectiveness, Persuasive Power, Ability to generate re-design requirements*

Figure 3.1: Effectiveness Measurement Criteria Previously Used and the Number of Studies in Which They Were Used

As can be seen from Figure 3.1, the most studied criterion of UEM effectiveness is its *validity*, which was used in 13 of the 15 studies reviewed. The validity measures not only how many usability problems are revealed during usability evaluations but whether they are valid ones, and therefore it is of primary importance for determining how effectively a UEM performs. The method's *thoroughness*, *reliability*, *cost-effectiveness* and the *severity of the problems predicted* are also important measures to consider when assessing and comparing UEMs, and they have been assessed in a number of the studies. From Figure 3.1 it can also be seen that the remaining six criteria do not appear to be as widely used since all but one have only been studied once. There are different reasons to explain that, which do not diminish the importance of these criteria. One reason why the *design-change effectiveness* is not investigated often, for example, could be because it is difficult to study as it involves implementing changes in the design of the application as a result of its evaluation, and then performing follow-on empirical tests with users. Such a procedure takes a considerable amount of time and resource to complete. Two other less 'popular' measures are *tailorability to context* and *ability to assess instructional efficacy* They are actually specific for the evaluation of educational software, and are very important for establishing the quality of the evaluation results produced while assessing such software. Therefore, the less-studied criteria should not be overlooked when assessing the effectiveness of EEMs for IMM.

3.3 Creating an EEM Effectiveness Framework

In order to compile a framework for measuring the effectiveness of EEMs for IMM the suitability and feasibility of the above criteria were assessed. Based on this analysis, it was established that some of the criteria, particularly those for measuring the effectiveness of traditional UEMs, cannot be applied directly for measuring the outcomes of evaluations of IMM, as they do not cater for the specific characteristics of the IMM evaluation process. Therefore, it was necessary to adapt these criteria in order to address the individual characteristics of the evaluations of IMM applications.

In particular, Sears' (1997) measures of *validity* and *thoroughness,* and Bastien and Scapin's (1995) measure of *reliability* were adapted to reflect the unique nature of the IMM evaluation process.

- *Validity*

The concept of *validity* was modified to incorporate the ability of a method to assess the instructional efficacy of IMM applications (Tergan, 1998). As user tests may not be sufficient to validate content and instructional approach problems, it was decided that instructional experts need to be involved to validate such problems. Otherwise, a number of valid content problems may be classified as invalid. The measure of *validity* was also extended to calculate the number of false alarms predicted (Lavery and Cockton, 1995; John and Marks, 1997).

- *Thoroughness*

A comprehensive measure of *thoroughness* is also proposed, that assesses three aspects: *actual thoroughness* (as proposed by Sears, 1997), *comparative thoroughness* (as used by Reiser and Kegelmann, 1994), and *incompleteness*. All provide quantitative and qualitative assessment of EEMs in terms of the number and nature of problems predicted either in comparison to those predicted using other EEMs (comparative thoroughness) or to those experienced by users (actual thoroughness and incompleteness).

- *Reliability*

Similarly, the measure of *reliability* was refined to provide qualitative information about the inconsistency of the predictions made by different experts using the same EEM. For the purpose, factors internal and external to the EEM are proposed that can influence the reliability of predictions. In this way, more concrete indicators of the factors that enhance and inhibit the effectiveness of EEMs for IMM can be obtained, which can then be used to generate more effective EEMs for IMM.

The measures of *theoretical validity of evaluation criteria* and *tailorability to context* were both refined by defining precise ways of measuring each one tailored to the unique conditions of the IMM evaluation. Finally, Jeffries' et al. (1991) measure of *cost effectiveness* was extended to reflect not only the severity of the valid problems predicted but also the cost of predicting false alarms.

Three of the criteria discussed in Section 3.2 were also considered important, but not feasible to test in the time available for conducting this thesis research. Therefore, they were not included in the effectiveness framework proposed in this chapter. These criteria are *persuasive power*, *design-change effectiveness* and *ability to generate re-design requirements*. As mentioned in the previous section, in order to test these measures re-design solutions of identified usability problems need to be defined and implemented. After that further usability tests with users will need to be conducted to establish where improvement has been achieved in terms of reducing the number of usability problems remaining in the user interface. Sufficient resources to conduct such software development were not available for this research.

Finally, five criteria were defined which are also considered important, but which have not been proposed in any of the studies reviewed in Chapter 2. The motivation for these criteria was provided by literature on structured usability problem reports (Lavery et al., 1997), effectiveness studies of general error prediction methods (Shackel, 1990; Stanton and Stevenage, 1998), and material characterising the evaluation process of Computer-Based Learning software (Saroyan and Geis, 1988). These measures are *diagnostic power*, *learnability* and *ease of use* of a method, *method specialisation* to evaluators from different professional backgrounds, and *quality of documentation*. Each criterion reveals further information about the effectiveness and usefulness of EEMs, and thus can enable instructional design practitioners to make better choices about what EEMs they could use.

3.4 EEM Effectiveness Framework

The proposed EEM effectiveness framework contains eleven effectiveness criteria, which are divided into *primary* and *secondary*. As can be seen from Figure 3.2, the primary category comprises six main criteria, each of which directly measures the effectiveness of the evaluation results produced using EEMs. In particular, they measure how valid, reliable and thorough the evaluation results are, how well specified the usability problems are, and whether the evaluation criteria included in the EEMs are founded in theoretical research within the field. As highlighted before, the most important of these criteria is the *validity* of evaluation methods.

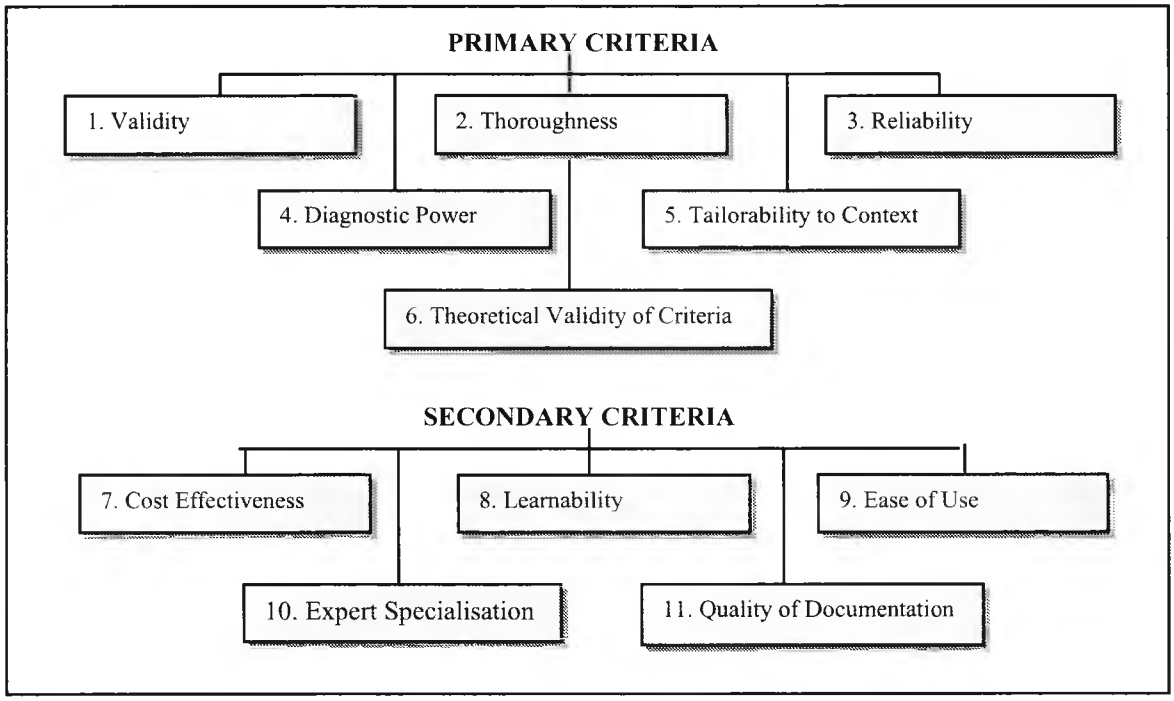


Figure 3.2: Expert Evaluation Method Effectiveness Framework

The secondary category, incorporates five criteria, which do not directly measure the outcome of the evaluation, rather they are concerned with the usefulness of EEMs. As can be seen from Figure 3.2, these criteria assess how cost-effective, easy to learn and easy to use a method is, together with whether it is targeted at different specialists who will apply it, and how well documented it is.

Although presented separately, the criteria are not independent from each other. For instance, the *validity* of the predicted problems is believed to be dependent on the *theoretical* and research *validity* of the evaluation criteria used in the EEM (Bailey, 1999; Tergan, 1998). The *validity* of the evaluation results is also influenced by the ability of the evaluators to consider context-specific issues, such as the target learners.

In the sections below, each effectiveness criterion is described in terms of its importance and the method of measuring it. Some measures are quantitative, i.e. they provide numerical values of EEMs' effectiveness, whereas others are qualitative measures describing desirable characteristics of effective EEMs for IMM. Some criteria, such as *thoroughness*, combine both quantitative and qualitative measures to provide a more comprehensive analysis of the particular aspect of the EEM it measures.

3.4.1 Validity

The most important measure of the effectiveness of an EEM is the extent to which the evaluation method fulfils the underlying purpose of evaluation. The primary aim of formative expert usability evaluation is to predict as many *valid* usability problems as early as possible in the development of a software product. The objective is to provide designers with enough evidence on which to make informed judgements regarding how effective the user interface design is in achieving the intended outcomes. From these judgements, actions can be taken which will result in redesigning the product in a way which improves its quality.

Sears (1997) postulates that a UEM is valid if the problems identified using this method are important for the achievement of the user tasks, and will affect users and their performance if they are not rectified. When evaluating IMM this means that evaluators should identify issues as problems only if they would impact learners, including not only their interaction with the interface but their learning performance as well.

Validity is traditionally measured by the ratio of 'real' problems that are predicted by experts to the total number of problems predicted by them, where 'real' problems are considered to be the ones experienced by users during user tests (John and Marks, 1997; and Sears, 1997). Although this is a sound approach for certain types of applications, it is not considered adequate when the evaluation of educational software is concerned. The main reason is that for the evaluation of the effectiveness of such software, subject-matter experts and instructional specialists are involved, as well as user interface specialists (Saroyan and Geis, 1988; Heller, 1996; Heller and Martin, 1999b). Such experts are able to uncover problems regarding the quality of the content and the effectiveness of the underlying instructional approach, which can affect users' knowledge construction processes and their learning performance. As users do not have the necessary knowledge to identify such issues, it is considered that independent instructional specialists, rather than users, should determine whether such problems are valid or not. Thus, a two-step approach to validating usability evaluations is proposed – firstly validation of all predicted problems against user test data and secondly validation of problems not experienced by instructional experts.

During Step 1 all predicted problems will be compared to all user problems identified in user tests, as suggested by John and Marks (1997) and Sears (1997). The predicted problems can also be compared to learning problems encountered during post-exposure knowledge tests. During Step 2 all problems not experienced by users will be validated by independent instructional judges. The rationale behind Step 2 is that instructional problems identified by expert evaluators often relate to the process of learning, and if compared only to the final learning performance of users (measured during post-exposure knowledge tests), such problems may be wrongly classified as invalid. This could distort the measure of *validity*. Therefore, independent instructional experts can be asked to validate whether any problems not experienced by the users are valid instructional issues. In this way, any valid learning problems that have been classified as false alarms during Step 1 can be reinstated as valid. The rating of instructional experts can however introduce bias in the validity results. To reduce this effect, two or more such experts need to rate predicted problems independently and their results then need to be compared.

Two aspects of the validity of the expert predictions need to be studied:

- *Actual validity*: measures the ability of the EEM to predict valid usability problems.
- *Invalidity*: measures the proportion of false alarms predicted during the expert evaluations.

The *actual validity* of an EEM can be determined using the following formula:

$$\text{Validity} = \frac{\text{\# of valid problems predicted}}{\text{Total \# of predicted problems}} \times 100$$

where # of valid problems predicted = # of problems experienced by users + # of problems not experienced by users but considered valid by independent judges

As the percentage signifies what proportion of the total number of predicted problems are valid, the higher the percentage, the higher the validity of the method.

On the other hand, all issues which are predicted by experts but are not experienced by users and are not considered valid by the instructional judges can be classified as *false alarms*. Therefore, the *invalidity* of the predictions made using an EEM can be established by calculating the proportion of false alarms identified by evaluators when applying the EEM. As false alarms are not likely to have an effect on the users' interaction with the application or their learning performance, they need to be identified and separated from those issues which will have such an effect, as any effort spent rectifying such problems will be wasted. Such false alarms also need to be excluded when measuring the validity of an EEM, as they distort the measure.

3.4.2 Thoroughness

Evaluations are thorough if they result in examining as many aspects of the application as possible (Sears, 1997). Three ways of measuring the thoroughness of expert evaluations of IMM are proposed. The first is a comparative measure of the design areas covered across different EEMs, and the last two assess how comprehensive expert predictions are in terms of the proportion and type of user problems successfully predicted and those missed out. The first measure can be used in comparative studies of the effectiveness of two or more EEMs, even if user test data is not available. The last two measures require user test data, but can be used in assessing the thoroughness of individual EEMs.

i) The first approach proposed measures the *comparative thoroughness* of UEMs by comparing the types of the problems identified by the different evaluation techniques to establish the comprehensiveness of the focus of each technique in comparison to the others. This approach was used by Reiser and Kegelmann (1994).

ii) The second approach assesses the *actual thoroughness* of evaluations by establishing the ratio of the problems identified by expert evaluations over the real problems that exist in an application, as proposed by Sears (1997). Since the denominator is difficult to establish, it is suggested to calculate the percentage of the user problems which have successfully been predicted by the experts, using the following formula:

$$\text{Actual Thoroughness} = \frac{\text{Total \# of user problems predicted}}{\text{Total \# of user problems found}} \times 100$$

The higher the percentage, the greater the thoroughness is, as that will signify that the method is able to predict more of the valid user problems that exist in the application. A qualitative analysis of the nature and severity of the user problems predicted also needs to be conducted in order to determine the scope of the EEM, or what types of real problems could be predicted applying the EEM.

iii) The final approach is inverse of the previous one, as it aims to establish the *incompleteness* of expert predictions. It first quantifies the proportion of actual user problems missed out by the evaluators, using the formula presented below:

$$\text{Incompleteness} = \frac{\text{Total \# of missed out problems}}{\text{Total \# of user problems found}} \times 100$$

It then determines the type and the severity of the user problems which evaluators could not predict using the EEM. In such a way, it can be identified what types of problems are outside the remit of the EEM. Such information can guide usability and instructional design practitioners as to what complementary evaluation methods need to be employed for a thorough analysis of IMM designs. It can also point researchers towards which aspects of existing EEMs need to be extended in order to provide more complete analysis of IMM designs.

Although method thoroughness bears similarities with method validity, the former measures not only the number of valid usability problems predicted, but also what type they are when using the last two approaches. The last approach also reveals the number of important valid problems that have been missed out by the expert evaluations.

3.4.3 Reliability

Evaluations are reliable if they consistently provide the same results under the same conditions (Sears, 1997). To establish the reliability, we need to determine whether different evaluators tend to find similar numbers of problems and corresponding problems when applying evaluation techniques. When IMM software is being evaluated, typically experts from different professional backgrounds are involved in the evaluation, including subject-matter experts and usability specialists. In such cases, differences between the problems different types of specialist identify are expected and desirable. Therefore, it is suggested to perform the analysis of similarities and consistency of problem sets separately for each type of expert to achieve an accurate measure of reliability.

The following two measures of method reliability are proposed:

i) Firstly, to compare the number of problems individual evaluators have identified.

ii) Secondly, to examine specific problems identified by the evaluators to identify the number of corresponding problems predicted using the same EEM, as was used by Bastien and Scapin (1995) and Hertzum and Jacobsen (1999). Qualitative analysis can then be performed to reveal potential reasons for inter-evaluator inconsistencies. For example, if significant differences or contradictions between the reviews of evaluators are found, this may mean that the evaluation criteria are ambiguous or that the evaluators have applied them in different ways because of their own expertise. The reliability of the evaluations is thus undermined, and this will pose difficulties to the development team in their effort to improve the design of the application.

3.4.4 Diagnostic Power

As defined by Lavery, Cockton and Atkinson (1997), a usability problem can be expressed in terms of a cause (usually an intrinsic design feature), an undesirable effect (usually an effect on the users' behaviour or performance), and its context. A thorough analysis of usability problems should therefore reveal the cause-and-effect relationship

between intrinsic design features and their outcomes in order for designers to be able to make the right judgements as to how critical the outcomes are for the target learners and what would be the best redesign solution. Thus, an EEM should be able to point evaluators to intrinsic design features that could potentially cause a particular problem to occur in specific interaction contexts. An EEM also needs to support evaluators in relating these features to the effects they are likely to have on target users and the achievement of their tasks. The ability of an EEM to support evaluators in determining problem cause-and-effect relationships was called *diagnostic power.* Accurately diagnosing the cause and user effect of a problem is the first step towards successfully redesigning user interfaces. Hence, this measure is important.

In order to determine the diagnostic power of a method, a two step approach is proposed, which is described below:

i) Firstly, the problem definitions need to be analysed to establish how well each problem has been specified. Predicted problems can be categorised as being either *fully-specified* or *under-specified.* A fully-specified problem describes the cause of the problem, and its effect on user's behaviour or performance. An under-specified problem definition can either include the cause without specifying the potential effect the problem may have on users, or it may state that there is a user problem without relating it to a particular design feature which may cause it.

ii) After that, it needs to be established whether it was the EEM that influenced how fully each problem was specified, and what other criteria may have contributed to that. The evaluation criteria, which prompted the identification of the problems, need to be analysed to determine whether they provide an explicit indication of the specified causes and/or effects. In such an analysis, other criteria that may have influenced how fully a problem has been specified need to be taken into account. Such criteria may include the expert's prior evaluation experience and the time available for conducting evaluations.

3.4.5 Tailorability to Context

Expert evaluations typically occur separately from the context of use, making context sensitive evaluation highly problematic. Expert EMs should thus provide evaluators with ways of mentally projecting how the target user population may use the application in the intended learning environment. Squires and McDougall (1996) define that an effective predictive evaluation tool should generate issues and questions specifically tuned to the perceived use of an application by the intended users. If no explicit way of tailoring the criteria is provided, it is left to the evaluators' initiative to consider how the application under consideration may be used by the target users, and make judgements as to the likely problems they may encounter. Furthermore, Tergan (1998) identifies that the lack of context-dependent evaluation criteria (e.g. regarding the particular subject matter content, the pedagogical strategy and the educational setting) limits the number of valid learning problems predicted, which can jeopardise the validity of the predictions made.

The following three categories are proposed to measure the EEM's *tailorability to context*: direct context consideration, indirect context consideration and no context consideration.

i) *Direct context consideration* means that evaluation criteria, such as guidelines and checklist questions, make an explicit reference to particular aspects of the learning context, and indicate what the likely implications of certain design features for the potential usage of the application may be.

ii) *Indirect context consideration* includes prompting the experts to look into factors in the learning context, but leaving it to their expertise to judge what implications they may have for learners and the achievement of their goals.

iii) *No context consideration* signifies that the evaluation criteria do not refer to any aspects of the context in which the design would be used. In such cases evaluators may judge the quality of the design without taking into account who will be using it and under what circumstances.

Direct consideration is the most preferable, as it indicates that evaluators are given more guidance as to how certain design aspects could influence users' interaction and potentially their behaviour and learning performance using the user interface.

Information about what contextual aspects have been considered during evaluations and whether the EEM has prompted evaluators to consider them can be gathered in a questionnaire or interview with the evaluators after their evaluation experience.

3.4.6 Theoretical Validity of Evaluation Criteria

Tergan (1998) and Stanton and Stevenage (1998) argue that evaluation criteria must be grounded in empirical evidence and relevant theoretical work in order to be valid. Thus, it needs to be ensured that there is a relationship between evaluation criteria and empirical qualitative and quantitative data regarding the instructional effectiveness and the usability of IMM programs (Reeves and Harmon, 1994). If there is no evidence about the empirical validity of the evaluation criteria, there is a danger that the conclusions reached about the effectiveness of the IMM application could be subjective and speculative (Bailey, 1999).

To establish the theoretical validity of EEMs we need to determine whether in formulating the criteria the authors have resorted to one or more of the following: cognitive theories, pedagogical theories, empirical studies, or case studies of the use and effectiveness of IMM. Conclusions will be drawn based on the information provided by the authors regarding the formation of their evaluation criteria.

3.4.7 Cost Effectiveness

As IMM software is typically developed under a number of practical constraints, such as limited budget to create artwork and tight release deadlines, cost-effectiveness is a particularly important factor in selecting an EEM to be used in a project. To establish the cost-effectiveness of a method, the costs of applying it need to be calculated first, including the resources required, such as the time needed and the number of evaluators required. After that the costs have to be juxtaposed to the potential benefits of the evaluation results. A cost-benefit ratio is proposed by Jeffries et al. (1991), which calculates the ratio of the time spent using each technique to the severity of the problems identified by experts. The formula proposed by Jeffries at al. (1991) was

modified in order to reflect the number of false alarms identified by the evaluators using the EEMs. The new formula is shown below:

$$\text{Cost-benefit ratio} = \frac{\text{Sum of severity scores of all valid problems predicted} - \text{Number of false alarms}}{\text{Total evaluation time}}$$

For that purpose, predicted problems are first rated according to a severity schema. The benefit from each evaluation is then measured by calculating the sum of the severity scores of the valid problems predicted by the experts. The benefit is adjusted by deducting the number of false alarms from the sum of the severity scores. The adjustment value, however, does not have an additional weighting associated with it in the way the valid problems are weighted according to their severity. Had a weighting been added to the cost of the false alarms identified, the bigger the number of false alarms, the lower the cost-benefit ratio of an EEM would have been. The adjustment is necessary as extra time is being spent to identify false alarms. More importantly, if the false alarms remain in the problems set, actions can be taken and the design of IMM can be modified based on such false predictions. This can result in wasting further time in re-design without making any improvement on the quality of the IMM, with a danger of introducing new problems. As an adjustment the number of false alarms is deducted from the sum of severity scores of all valid predicted problems. Finally, the total number of person-hours spent conducting the evaluations can then be taken to calculate the denominator. The higher the value of the ratio the more cost effective the EEM is.

3.4.8 Learnability

Shackel (1990) defines method learnability as the amount of training required to achieve a reasonable level of efficient and effective error prediction. Stanton and Stevenage (1998) propose two ways of measuring learnability: familiarisation time and subjective assessment of the ease of acquisition of the technique by the evaluators. The average time taken by the evaluators to familiarise with the EEM will be calculated first. The subjective assessment of how easy to learn an EEM can then be measured in post-evaluation interviews or questionnaires with evaluators.

3.4.9 Ease of Use

This criterion assesses how easy to use the evaluation technique is for those applying it in a particular evaluation situation. The ease of use of an EEM depends on how complex it is and how adequately and comprehensively it is documented.

It is proposed to establish the ease of use by asking evaluators for their opinion on how easy or difficult they found applying the method and its procedure. Such information can be gathered in post-evaluation interviews or questionnaires with evaluators.

3.4.10 Expert Specialisation

Unlike traditional UEMs, evaluation methods for instructional software typically require a range of experts to provide their feedback regarding various aspects of the usability and instructional effectiveness of the software. Saroyan and Geis (1988) distinguish between three main types of experts usually required for the evaluation of instructional software, namely subject-matter experts, instructional designers and presentation experts. *Subject-matter experts* typically have a firm grasp of the basic concepts of their particular domain, including knowledge about the rules, goals and principles of the field. During the evaluation, such experts can comment on content accuracy, comprehensiveness, or semantic and syntactic ambiguities. The role of *the instructional designer* is usually to ensure that all of the components of the instructional approach employed have been implemented in the software. Finally, *presentation experts* make contributions in user interface and human-computer interaction areas, such as graphics design, screen layout, navigation, and modality of the interaction.

Saroyan and Geis (1988) argue the need to specialise evaluation criteria within the realm of expertise of the specialists to be involved. If all types of expert are required to use the same evaluation criteria, there is a danger that experts may attempt to comment on areas they are not familiar with, which can jeopardise the validity and reliability of the predictions if the evaluator's judgements are inaccurate or contradict each other.

Whether a method is adequately tailored towards different specialists and whether it takes into account the various perspectives such specialists bring into the evaluation can be established by analysing the method to check whether explicit guidelines are included for different experts regarding which evaluation criteria to use. Furthermore,

evaluators can be asked for their opinion regarding the relevance of evaluation criteria to their realm of expertise, whether they are asked to comment on concepts that are outside their expertise, and whether any terminology unfamiliar to them is used.

3.4.11 Quality of Documentation

A well-documented EEM should provide a detailed and unambiguous description of its principles and evaluation criteria. In order to be of use to designers, a procedure for evaluation needs to be explicitly specified, guiding the evaluators in analysing user interfaces. In particular, instructions as to what steps to follow in reviewing the interface and how to use the evaluation criteria should be included. Furthermore, criteria-based evaluation methods are usually constructed as instruments, which are appropriate for application to a variety of software for a variety of educational purposes. When an individual product is to be evaluated, however, only a selection of the criteria may apply. Therefore, a well-documented method should include a clear procedure for selecting which evaluation criteria are relevant to the particular application being evaluated. If there is no procedure to select appropriate criteria, evaluators have to decide themselves on the relevance of criteria to their individual intentions, which may be a tedious experience depending on the number of evaluation criteria included (Tergan, 1998).

The quality of the documentation can be assessed by asking evaluators to comment on the adequacy and sufficiency of the method documentation, how understandable it is, and whether a specific procedure for analysing user interfaces is provided and how adequate it is for the evaluation of the particular application. Finally, evaluators can be asked how they selected the relevant evaluation criteria and whether the method guided them in their selection.

3.5 Conclusion

The Effectiveness Framework, presented in this chapter, presents a thorough methodological approach for assessing the effectiveness of EEMs for the evaluation of IMM. It contains eleven effectiveness criteria for measuring various aspects of a method's performance and usefulness to achieve a comprehensive assessment of the merit of an EEM. It extends the previous efforts made in the area, as such a systematic

approach to the effectiveness of EEMs for instructional software has not been proposed or used before, which was discussed in Section 2.4.

The framework provided a basis for empirically assessing the effectiveness of three existing EEMs for evaluating IMM software. The empirical study, which used the Effectiveness Framework, is presented in the following chapter. As EEMs for IMM proliferate, the framework can be applied by method developers and usability practitioners to establish EEMs' effectiveness, and highlight their advantages and disadvantages.

Chapter 4

Assessing the Effectiveness of Existing Expert Evaluation Methods for IMM

Chapter 4

Assessing the Effectiveness of Existing Expert Evaluation Methods for IMM

4.1 Introduction

A comprehensive empirical study was carried out to achieve Objective 2 of this thesis:

> *Objective 2:* To measure empirically the effectiveness of existing EEMs for IMM using the effectiveness framework created in Objective 1.

The aim of the study was to assess and compare the effectiveness of existing EEMs for IMM. As argued in Chapter 2, little empirical evidence exists of the effectiveness of expert evaluation techniques for IMM, especially in comparison to one another. Previous studies mainly summarise key features of such methods, and highlight certain strengths and potential limitations, without providing empirical evidence of methods' evaluation performance. Therefore, an empirical study of three existing EEMs for IMM was undertaken to obtain more comprehensive information regarding how effective and efficient such methods are at predicting valid usability and learning effectiveness problems. The study investigated both methods' performance and usefulness, focusing on important aspects such as validity, thoroughness, reliability, and cost-effectiveness, as outlined in the Effectiveness Framework described in the previous chapter.

In order to gain insight into the actual occurrence of the problems predicted by the experts in real use of the application, usability tests with representative users were conducted (John and Marks, 1997; Sears, 1997). In this way, it was possible to establish how valid and thorough expert predictions are. This was achieved by comparing the predicted and the experienced problem sets in order to establish how successful experts were at predicting real user problems. An overview of the process of assessing the effectiveness of the EEMs is illustrated in Figure 4.1.

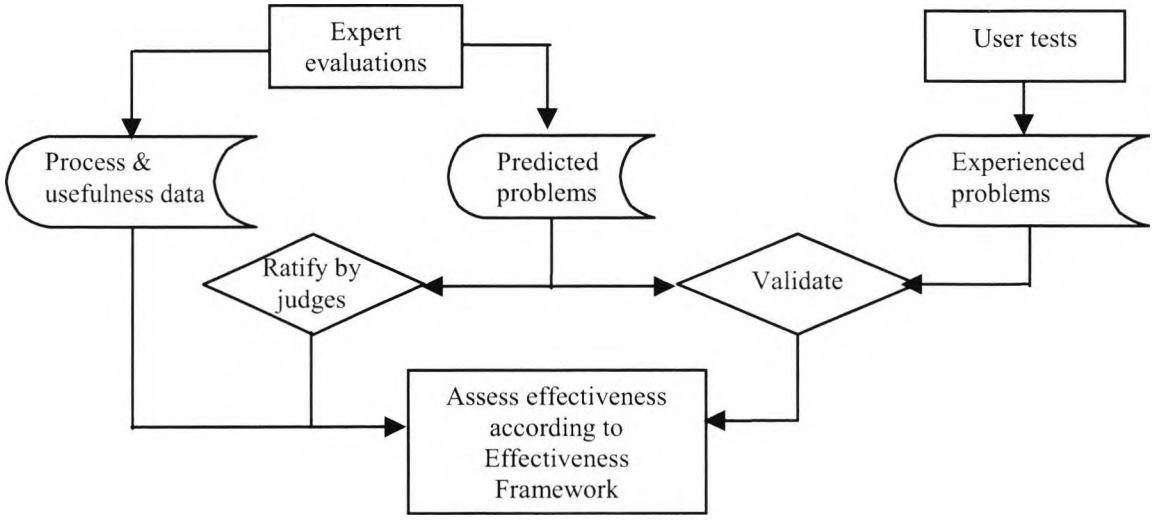


Figure 4.1: Overview of EEM Effectiveness Assessment

4.2 Methodological Approach of Expert Evaluations

This section describes the method used to conduct the expert evaluation studies.

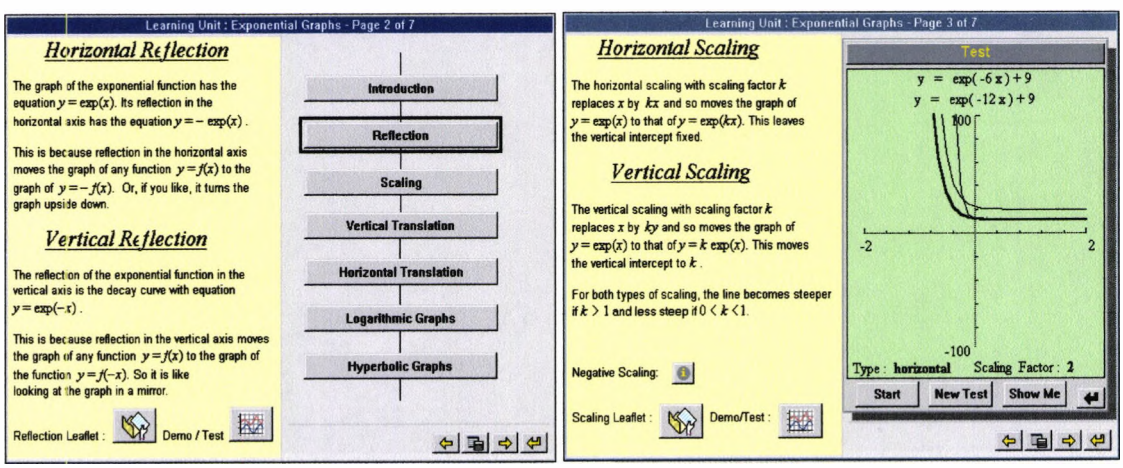
4.2.1 The Expert Evaluation Methods

From the EEMs suitable for the evaluation of IMM, described in Chapter 2, three methods were selected to be assessed and compared. They were the *Multimedia Cognitive Walkthrough* (Faraday and Sutcliffe, 1997), the *Interactive Multimedia Checklist* (Barker and King, 1993) and the *Multimedia Taxonomy* (Heller and Martin, 1999a, 1999b, 1999c). Copies of the methods' documentation can be found in Appendix 4.1. These methods were selected as they are of different nature – the first is a taxonomy-based approach, the second is a checklist and the third is a cognitive walkthrough. Although each method relies on experts to review design aspects of IMM applications, they vary in the way evaluators' judgements are derived and the evaluation criteria provided to make these judgements. In particular, the *Multimedia Cognitive Walkthrough* (MMCW) encourages evaluators to examine multimedia presentations in three consecutive steps. In each step the evaluators need to check the conformance of certain aspects of the MMUI to a set of guidelines. On the other hand, the *Interactive Multimedia Checklist* (IMMC) consists of a set of questions divided into twelve categories, such as engagement, interactivity and appropriateness of the media combination. In reviewing the quality of IMM applications, different types of experts are required to answer those questions which they consider relevant to the application

under consideration. Finally, the *Multimedia Taxonomy* (MMT) differs from the other two in that it does not provide explicit evaluation criteria. Instead, it represents a categorisation schema of multimedia design issues presented in three separate, but interrelated dimensions, and the evaluators are expected to firstly generate evaluation questions for the intersections of the three dimensions relevant to the particular software being evaluated, and then examine the multimedia application using these questions.

4.2.2 The IMM Software

A commercially available multimedia learning environment for studying and practising Mathematics at university level was chosen to be evaluated. The software comprises a number of mathematical modules, reference material, assessments and resource tools. It was considered that only one topic of a module could be evaluated due to the limited time the experts were available for. The topic chosen covers the principles of exponential graphs, their functions and the three types of transformation – reflection, scaling and translation. Two typical screens from the learning tool are presented in Figure 4.2. A series of 23 screens present the maths content in textual, graphical and animation formats.



(a) (b)

Figure 4.2: Sample Screens from the Maths Learning Tool (NAG ©)

Once a topic is selected, the learner may navigate within the topic either linearly - forward to the next or backward to the previous screen using a set of navigation buttons - or non-linearly - select a sub-topic of interest from a navigation menu presented on the right-hand side (shown on Figure 4.2 (a)).

In order to present the maths concepts, a textual explanation is presented on the left-hand side, which is supported by graphs usually presented on the right-hand side, as can be seen on Figure 4.2 (b). The learner can also obtain further information regarding the sub-topic from a 'leaflet', which is a short document that deals with a specific mathematical concept. The application adopts a tutorial style of teaching, where the learners are first presented with a textual explanation of the maths concepts and then they are able to assess and refine their knowledge of the concepts learnt by performing tests for each concept.

4.2.3 The Experts

Evaluators, with expertise recommended by the authors of each method, took part in the study. In particular, ten specialists (8 male and 2 female) with the following professional backgrounds took part in the expert evaluations:

Table 4.1: Experts Involved Using Each EEM

Method	Evaluators Required
MM Cognitive Walkthrough	1 multimedia designer and 1 usability specialist with multimedia design experience
Interactive MM Checklist	2 multimedia designers and 2 subject-matter experts
MM Taxonomy	2 multimedia designers and 2 subject-matter experts

When recruiting evaluators, effort was made to ensure that only qualified and experienced evaluators took part in the study. That was necessary to ensure the quality of the results and to reduce the likelihood of evaluators identifying *bogus* problems (Cockton and Woolrych, 2001) due to the evaluators' inexperience, which would reduce the measure of validity of the EEM (see Section 3.4.1). Thus, multimedia designers (MMDs) with a minimum of 2 and as much as 9 years of experience in developing and evaluating multimedia software were recruited. The usability specialist had 5 years of usability engineering experience of multimedia applications. Similarly, all subject-matter experts (SMEs) had advance knowledge in Maths, and between 6 and 35 years of experience teaching Maths to university students. The evaluators' qualifications and prior experience were gathered in a questionnaire, which is given in Appendix 4.2.

Recruiting only experienced evaluators limited to a large extent the number of evaluators that could take part. Although more professionals were contacted, only ten

agreed to participate, meaning that each method could only be applied by a small number of evaluators. Furthermore, only two specialists were recruited to apply the Multimedia Cognitive Walkthrough because its authors only recommend it to be used by MMDs. Subject-matter experts are not recommended as the method concentrates on low-level multimedia design issues.

Based on a mathematical formula for predicting the number of usability problems found, Nielsen (1994) demonstrates that up to fifteen experts are needed to identify as much as 90% of the total real usability problems that exist in a piece of software. The authors also found that there is an accumulative effect and that the bigger the number of experts, the larger the number of new problems identified. From these findings it appears that involving two to four experts in an evaluation will impact the thoroughness of expert predictions. In particular, Nielsen (1994) found that four evaluators find around 70% of the real usability problems that exist, and two experts find just over 55%. Therefore, if more evaluators could have been recruited to use each EEM the bigger the *thoroughness* of their predictions might have been (see Section 3.4.2). However, the benefits of having more evaluators need to be weighed against the costs. Woolrych and Cockton (2002) found that the bigger the number of experts the higher the number of false alarms, hence the lower the *validity* of their predictions. Given these findings and the practical constraints in recruiting experts, having up to four evaluators to apply each method was considered an acceptable number. However, having up to ten adequately qualified evaluators using each method would have achieved more reliable results, particularly in terms of the EEMs' thoroughness.

4.2.4 Protocol of Expert Evaluations

4.2.4.1 Familiarisation with EEMs

None of the evaluators had any prior knowledge of any of the three evaluation methods used in this study. Prior to the evaluation, each expert was given all materials provided by the authors of the method they were going to use to familiarise themselves with. The evaluators took as much time as they needed to learn the method, and they were asked to record how long this took. The evaluators did not have any knowledge of the other two methods. In this way, no bias from any of the other two techniques could occur, thus preserving the internal validity of the evaluation studies (Gray and Salzman, 1998).

4.2.4.2 Expert Evaluations

At the start of the evaluation session, the evaluators were presented with a context description, giving a short description of the Math software, the target user group and the context in which it is intended to be used (see Appendix 4.3). The evaluators were then given 15 minutes to familiarise themselves with the IMM application. After that, each expert individually evaluated all 23 screens of the Exponential Graphs section in order to uncover potential usability and learning problems. The evaluators were asked to apply the methods as prescribed by their authors, and no additional instructions were given to them as to how to conduct the evaluations (see Appendix 4.4 for evaluator task description). During the evaluations, the experts were asked to write problem reports, providing a description of each problem and specifying the evaluation criterion that prompted them to identify that problem. The latter was necessary to differentiate the problems identified using the EEMs from those predicted using the experts' own judgement. They were free to record any additional problems they considered had limited the usability or the learning effectiveness of the IMM application. Each evaluator worked individually in the same office under the same conditions. Therefore, environmental factors were kept the same for each usability evaluation, which was important for maintaining the external validity of the study (Gray and Salzman, 1998).

4.2.4.3 Post-Evaluation Interviews and Questionnaires

To be able to assess each EEM according to all criteria in the Effectiveness Framework, as specified in Chapter 3, a significant amount of information, in addition to that contained in the evaluators' problem reports, was required. Firstly, it was necessary to collect data about the evaluators' experience and attitude towards key aspects of the EEM they used, which would help assess the method's usefulness. Secondly, it was necessary to gather information about various aspects of how the expert evaluations were conducted in order to gain understanding of the factors that influenced the methods' performance. As can be seen from the effectiveness criteria definitions given in Chapter 3, most criteria are multi-causal, i.e. methods' performance would depend on multiple factors or variables, which could either be independent or interrelated.

It was decided that a combination of two data gathering techniques was most appropriate to collect data required (Oppenheim, 1992). Firstly, questionnaires were used to gather the overall experience and impression of the evaluators, as they are

economical in terms of time and resources. Semi-structured interviews were then used to gather evaluators' opinions regarding specific aspects of the individual EEMs and other variables contributing to the effectiveness of the usability evaluations.

The post-evaluation questionnaire (see Appendix 4.5) was standard for all methods and aimed to gather evaluators' experience with, and attitude towards, important features of the method they used. The questions were structured into the following three sections:

- *General experience with the evaluation method*: gathered information about how easy to learn, how easy to use, and how well documented each method was found.
- *The process to evaluation*: gathered procedural information regarding the approach to evaluation adopted by the evaluator, the adequacy of the instructions provided for reviewing the user interface and those for applying the evaluation criteria.
- *Evaluation criteria*: collected information on how clearly and sufficiently defined the evaluator found the evaluation criteria provided in the EEM.

Semi-structured interviews were also administered with each evaluator to gather further information on their experience using the method. Separate interview questions were designed for each method to collect detailed information regarding specific features of the EEM. For example, the interview questions regarding the Multimedia Taxonomy covered issues regarding how well-defined the three dimensions of the taxonomy are, how easy the evaluator found generating new questions, and which context specific aspects the method encouraged them to consider, if any. All interview questionnaires are attached in Appendix 4.6.

The questionnaire and the semi-structured interviews were administered immediately after the evaluators had completed the evaluations.

4.3 Methodological Approach of User Tests

This section describes the method used to conduct the user tests.

4.3.1 The Users

Four students (3 male and 1 female), undertaking a foundation course in Mathematics at City University, were involved as they represented typical users of the Maths application. Ages ranged from 18 to 33 years, with a mean of 24 years. The students were required to use the software as part of their tutorials, and the material covered in the software was included in their course curriculum. Thus, as the students were aware that they would be assessed at the end of the course, they were motivated to use the software and learn the concepts presented. Draper et al. (1996) emphasise the importance of involving subjects with the right motivation to learn in learner tests of educational software, which was ensured in the learner tests conducted in this study. Rewarding each student with £10 for their time may have further motivated them.

Experiments, such as the one conducted by Virzi (1992), have shown that only a few participants are needed in user testing to identify problems and even fewer participants to identify severe problems. Nielsen and Landauer (1993) also recommend using up to five users in a usability test, as they would be sufficient to discover 85% of the existing usability problems. Although the authors suggest that at least 15 users are required to discover all usability problems in a design, the results of a cost-benefit analysis of user testing they performed showed that the optimal number of users is between three and five users per single test. Therefore, for this study, four users were considered optimal to identify a considerable proportion of the existing usability problems, within the time available. Furthermore, the four students represented typical users of the Maths application, as they were attending a course in Mathematics requiring the use of the software during tutorials and for revision. However, as Woolrych and Cockton (2001) point out, the chances of getting the right users depends on the distribution of the individual differences between test users, the tool under test and the tasks performed during testing. Therefore, although care was taken to select four representative users, the small sample selected may not have represented the variety of skills and abilities that students who are likely to use the software typically possess. Although that was not possible in this study, it is recommended that minimum 15 users should be asked to evaluate IMM software to produce more reliable results.

Henderson et al. (1995) recommend a multi-method strategy to user testing as the most effective in generating the most reliable data. Therefore, in order to collect as much data as possible regarding the users' interaction with the application and the problems they experienced, three different data collection techniques were employed: one-to-one user tests, post-evaluation interviews, and *before* and *after* knowledge tests.

4.3.2 Protocol of One-to-one User Tests

As the learning objectives of the Exponential Graphs section of Maths application are to teach the basic concepts of these graphs, the learners were given four tasks to perform which consisted of learning about the principles of exponential graphs and exploring the three different types of their transformation (see Appendix 4.7). For example, the users asked to find out what exponential functions are, where they are used and how their functions and graphs are represented. The tasks were carefully designed to ensure that the users covered the same 23 screens and every single function within these screens that was evaluated by the experts. In this way the predicted problems could be validated or rejected by the user's experience. As part of each task, the learners were required to familiarise themselves with the Maths concepts and carry out tests in order to check and refine their knowledge of the material. As the students had not used the software prior to the experiment, they were given 15 minutes to explore and familiarise themselves with the application. After that they were asked to perform the tasks in the order of their choice, and no time limits were given for individual tasks or overall. The subjects were asked to provide concurrent think aloud protocol while performing each task, as defined by Ericsson and Simon (1984). The sessions were video recorded in order to capture users' interaction with the application. The users' verbal protocols were also captured on an audio-tape to ensure higher quality of the voice recording. The researcher was present during the evaluation sessions observing users' interactions and prompting them for clarification of their actions.

4.3.3 Post-Evaluation Interviews

User surveys are recommended by Preece at al. (1994) and Shneiderman (1998) to complement usability tests, as they are an inexpensive and acceptable form of data gathering. Furthermore, Shneiderman (1998) highlights that interviews with individual users can be productive because the interviewer can pursue specific issues of concern, and can lead to specific, constructive suggestions. Henderson et al. (1995) also suggest

that semi-formal funnelled interviews are a better method than attribute-style questionnaires for examining the usability of a piece of software with users. The authors also highlight that the time between software evaluation and interview should be minimised to avoid interference effects due to memory loss.

Based on these recommendations, immediately after the users had completed the tasks to their satisfaction, which took between 40 and 70 minutes, retrospective interviews were carried out with each user to gather their subjective impressions about different aspects of the user interface design. The users were also asked to clarify any points about the interaction sessions, such as reasons for actions carried out, which had not been given in the verbal protocol. The interview questionnaire (see Appendix 4.8) contained two general questions aiming to elicit users' likes and dislikes about the application, and 11 categories of specific questions eliciting users' opinion of specific aspects of the user interface. The categories covered a wide range of design aspects including media design, user interaction, navigation, Maths content, and learning support. In total 58 questions were included. The questions were systematically formulated to validate the problems predicted in the expert evaluations. For example, the users were asked to comment on how engaging the material was and how easy to navigate they found the application, which were two areas experts found problem in. This aimed to ensure that any false alarms are correctly identified against users' experience with the IMM. The majority of the questions were formulated as multi-choice, to ensure uniformity of the answers. The answer choices ranged from the least favourable to the most favourable on a scale of 1 to 5, and also included a 'Do not know' option not to force the learners to provide an opinion if they had not formed one (Oppenheim, 1992). This aimed to ensure the truthfulness of the responses. The questionnaire was administered as an interview with each individual user, in order to seek further explanation and detail from them, as well as examples of problems encountered and design recommendations.

4.3.4 Before and After Knowledge Tests

In order to establish how much knowledge the learners gained as a result of using the IMM application, knowledge tests were administered before and after each session. Such tests are highly recommended by Draper et al. (1994) for assessing the instructional effectiveness of education software.

The pre-exposure knowledge tests (see Appendix 4.9) were conducted immediately before each user test, and aimed to establish learners' prior knowledge of the subject matter covered in the application. It was found that all learners were domain novices as they had no prior knowledge of exponential functions and their graphs. However, they were all familiar with the principles of linear functions and their graphs, whereas two of them also had some knowledge of the principles of reflection.

After the post-evaluation interview, each learner was given a post-exposure knowledge test (see Appendix 4.10) to examine the knowledge s/he gained while working with the application. The material covered by the learners was divided into 20 knowledge propositions, of which the learners were expected to have a reasonable level of comprehension. The propositions are shown in Table 4.2. The propositions correspond to the learning objectives of the Exponential Graphs section, and each knowledge proposition was tested in the post-exposure knowledge test. In order to establish whether effective learning had taken place, the questions were designed to test not only learners' recall of information, but also transfer and application of information. To ensure that the test questions would accurately test learners' understanding of the material, a lecturer in Mathematics at City University was consulted, who provided sample exam questions on exponential graphs and also reviewed and approved the post-exposure knowledge test.

Table 4.2: Propositions Tested in Post-Exposure Knowledge Tests

No	Knowledge Proposition
1	There are two ways of representing exponential functions: $y=e^x$ and y=exp(x).
2	There are three fundamental exponential functions: e^x, 2^x and 10^x , and depending on their base factor, the corresponding graphs are more or less steep.
3	During *horizontal reflection* the equation of y=exp(x) is transformed to y=-exp(x).
4	During *horizontal reflection* the graph of y=exp(x) is reflected in the horizontal axis or turned upside down.
5	During *vertical reflection* the equation of y=exp(x) is transformed to y=exp(-x).
6	During *vertical reflection* the graph of y=exp(x) is reflected in the vertical axis or represents a mirror image.
7	During *horizontal scaling* the equation of y=exp(x) is transformed to y=exp(kx).
8	During *horizontal scaling* the graph of y=exp(x) is moved, but the vertical intercept remains fixed. If k>1 line becomes steeper, if 0<k<1 line becomes less steep.
9	During *vertical scaling* the equation of y=exp(x) is transformed to y=k exp(x).
10	During *vertical scaling* the graph of y=exp(x) is moved and the vertical intercept is equal to k. If k>1 line becomes steeper, if 0<k<1 line becomes less steep.
11	*Horizontal scaling with negative factor* (i.e. -k) is a composition of horizontal scaling with factor k and vertical reflection. Therefore y=exp(x) is transformed to y=exp(-kx).

Table 4.2: Propositions Tested in Post-Exposure Knowledge Tests (continued)

No	Knowledge Proposition
12	During *horizontal scaling with negative factor* the graph of y=exp(x) is horizontally scaled and reflected in the vertical axis at the same time.
13	*Vertical scaling with negative factor* (i.e. –k) is a combination of vertical scaling and horizontal reflection. Therefore y=exp(x) is transformed to y=-k exp(x).
14	During *vertical scaling with negative factor* the graph of y=exp(x) is vertically scaled and reflected in the horizontal axis at the same time.
15	During *horizontal translation* the equation of y=exp(x) is transformed to y=exp(x+k).
16	During *horizontal translation* the graph of y=exp(x) is moved to the left if k>0, and moved to the right if k<0.
17	During *vertical translation* the equation of y=exp(x) is transformed to y=exp(x) + k.
18	During *vertical translation* the graph of y=exp(x) is move up with k units if k>0, and moved down with k units is k<0.
19	In *horizontal translation* the distance between the two graphs is equal to the value of k.
20	In *vertical translation* the distance between the two graphs is equal to the value of k.

4.4 Method of Analysis of the Expert and User Evaluation Data

This section describes the method which was used to analyse the expert and the user data, in order to inform the assessment of the effectiveness of the three EEMs against the criteria set in Chapter 3. In particular, the procedures for refining and validating the expert problem sets are defined, together with those for extracting and analysing user problems. The procedures follow those suggested in Chapter 3.

4.4.1 Analysis of Expert Evaluation Data

4.4.1.1 Refining of Expert Problem Sets

In order to achieve an accurate measure of the methods' *validity* the original problem sets needed to be refined to exclude any repetitions made by individual evaluators and also any bogus problems (Cockton and Woolrych, 2001). The first could potentially inflate the value of *validity*, whereas the second would deflate it if not removed from the problem sets. Duplicated problems were recorded as they are an important aspect of a method's *reliability*.

It was also essential to make an accurate distinction between the problems that were identified as a result of applying each EEM and those that were identified due to the evaluator's own expertise. This was necessary in order to establish with certainty the

EEM-derived problems and to achieve causal construct validity (Gray and Salzman, 1998), i.e. ensuring that only EEM-derived problems are being manipulated.

Therefore, the original problem descriptions (PDs) generated by the evaluators were refined to exclude any repetitions made by evaluators using the same method. Any bogus problems (Cockton and Woolrych, 2001) due to the evaluator's unfamiliarity with the rest of the application were also excluded from the predicted problem sets. These included issues regarding facilities that exist, but the evaluator did not encounter while reviewing the application. In this way, the number of unique problems identified using each EEM was established.

Each unique problem set was divided into problems whose detection evaluators attributed directly to the EEM, and those discovered based on the experts' own judgement. The distinction between the two was made based on the information provided in the problem reports, where the evaluators explicitly indicated which guideline, which checklist question or which cell of the Multimedia Taxonomy the problem was prompted by. The PDs for which no such information was given were assumed to be based solely on the evaluator's expertise. The number of EEM-derived problems was then counted.

4.4.1.2 Assigning Severity to Expert Problems

It is important to differentiate between predicted usability and learning problems that would have a severe impact on the achievement of the user tasks and those which are minor, cosmetic or would have no impact on user's performance, i.e. false alarms. Rating the severity of the predicted problems was also necessary to establish the *cost effectiveness* of each EEM, as defined in Section 3.4.7. Therefore, the predicted problems were analysed to establish their severity.

To reduce the likelihood of rater bias while rating of the problem severity, two judges were asked to assign severity to the predicted problems independently from each other. In particular, one instructional expert and one instructional designer, both familiar with the design of the Maths software, were asked to review the predicted problems. For the purpose, Nielsen's (1999) severity rating schema was modified to accommodate aspects of the usability and learning effectiveness of IMM applications. The five-grade rating

scale is presented in Table 4.3. When differences between the ratings of the two judges occurred, the ratings were averaged. If the average was not a whole number, it was rounded to the more favourable category.

Table 4.3: Problem Severity Rating Schema

Grade	Description
0	**Invalid problem** This is not a valid problem, either because the problem statement is not true or because the problem statement, although true, should not be classed as a usability problem.
1	**Cosmetic problem only** The users will be able to overcome this problem easily. The problem needs not be fixed unless extra time is available to developers.
2	**Minor usability problem** The users will be able to continue with a pause for thought; or it is likely to cause a minor obstruction of the learning process, which will not affect learners' learning outcomes; or it causes a minor dislike with users. Fixing this problem should be given low priority.
3	**Medium usability problem** The users could continue after pausing for a significant time or trying alternatives successfully; or the learners are able to comprehend the material, but with difficulty; or it causes significant dislike or decreases learners' motivation. Fairly important to fix this problem, thus should be given medium priority.
4	**Major usability problem** The users cannot continue without external help; or the learners will not be able to comprehend the material or acquire the skill being taught without external help. Imperative to fix this problem, thus should be given highest priority.

4.4.1.3 Categorisation of Expert Problems

To assess the *comparative thoroughness* of the expert evaluations, as defined in Section 3.4, it was necessary to categorise the predicted problems depending on the type of design area they are concerned with. Each predicted problem was analysed to identify what design area it covers, e.g. navigation, media design, user control. Generic problem categories were then identified from the data, refined and then defined. The categorisation schema that was devised is presented in Table 4.4. It contains nineteen problem categories that accommodate the diversity of the problems identified across all EEMs. The schema defines typical usability problems that can be identified during usability tests of IMM, and therefore can be used by practitioners to categorise problems identified during such tests.

Table 4.4: Problem Categorisation Schema

Category Name	Category Description
Attentional design	Issues regarding perception and attention to individual media. Includes techniques for making information salient and directing user's attention to important information.
Assessment	Aspects relating to monitoring and assessment of learner's knowledge and performance.
Subject-matter content	This generic category encompasses issues which relate to the maths content, such as the structure of material, its completeness and comprehensiveness.
User engagement	Includes problems about the capability of the software to engage the users and involve them because it includes features which they could find motivating and enjoyable.
Feedback to user	Comments on the amount and quality of response the system provides to the users as a result of their actions.
Learning goals	Aspects related to the main learning aims and objectives of each stage of the presentation.
Media design	Issues regarding design and formatting of individual media resources, such as text and graphics.
Media integration	Issues regarding establishing cross-references between various media resources simultaneously representing information, as well as time synchronisation between dynamic media.
Misleading functionality	Features that could lead users to believe that they afford certain actions, but actually perform different actions.
Missing functionality	Features or functions that are not currently available in the application, but experts believed they were required.
Mode of interaction	Comments on the mode of interaction, e.g. the use of keyboard and mouse, both discretely or in parallel.
Navigation	Issues regarding traversing within and across the different parts of the application, such as advancing forward, moving backward. It also includes issues about the clarity of navigation buttons and menus.
Presentation	Generic category covering a broad range of aspects regarding presenting information in the software, which includes conformity to standards. It differs from media design category in the way that it goes beyond design of individual media.
Screen layout	Spatial arrangement of presentation objects on the screen.
Tailorability	Ability of the application to be personalised to user's preferences and be adapted to their needs and level of competence.
Task instructions	Directions to users regarding how to perform operations, such as perform a test or request more information.
User control	Comments concerned with the nature of the control given to the users in their interaction with the system.
User interaction	Issues assessing the quality of the users' interaction with the system, including levels and style of interaction, user input and feedback provided.
User support & scaffolding	Includes techniques of supporting users in performing tasks, or help learners with achieving their learning goals.

In this study, the schema was used to categorise and compare the problems predicted using the three EEMs. The results from this analysis are presented in Section 4.5.

4.4.1.4 Establishing the Reliability of Expert Predictions

As discussed in Chapter 3, in order to establish the *reliability* of an EEM an analysis was conducted to determine whether different evaluators tend to find similar numbers of problems and corresponding problems when using the same method. Because in two of the methods used in this study – Interactive Multimedia Checklist and Multimedia Taxonomy – experts from two different professional backgrounds were involved, differences in the numbers and types of problems are expected and desired. Thus, the analysis of similarities of problem sets was based primarily on pairs of experts from the same area of expertise, e.g. MMDs or subject-matter experts, and not across specialist fields. The results from this analysis are given in Section 4.5.

4.4.1.5 Establishing the Diagnostic Power of each EEM

To establish how fully diagnosed a predicted problem is, each PD was analysed in order to identify whether the problem cause and its effect on the users had been specified or not.

The evaluators were instructed to identify and record problems as guided by each EEM, and no explicit format was set for specifying problems. Thus, it was left to the experts' discretion to decide on the content and level of detail of their PDs. The reason for this was to preserve the authenticity of the results by not constraining evaluators to specify aspects that were not prompted by the method they were using, as explained earlier. Therefore, the format and the detail of the PDs varied across problem reports. Some descriptions were more detailed than others, some also included redesign suggestions, whereas others did not.

In order to describe the degree to which a problem has been specified, the following four categories were defined:

- FS – fully specified problem, indicating the problem cause and its likely effect on users,
- US-NE – under-specified, where the likely effect on users has not been predicted,

- US-NC – under-specified, where the design feature responsible for the problem has been omitted,
- RS – redesign suggestion only without specifying what the usability problem is.

4.4.1.6 Analysis of Expert Interviews and Post-Evaluation Questionnaires

In order to be able to make conclusions regarding the usefulness of each EEM, e.g. its ease of use or expert specialisation, the evaluators' feedback given during the interviews and the post-evaluation questionnaires were analysed.

The interview data gathered from the experts was first transcribed. Sample transcripts are given in Appendix 4.11. Together with the post-evaluation questionnaires, this data was then analysed using a coding system, following the principles outlined by Oppenheim (1992). The coding system contained seventeen data codes, each of which reflected a variable influencing the effectiveness of EEMs, as described in Section 3.4. Example codes include: *UndCr* - understandability of evaluation criteria, and *ContAspMet* - contextual aspects considered during the evaluation which were prompted by the EEM. The questionnaire and interview transcripts were annotated using the relevant codes, and were then used to analyse and formulate conclusions regarding different aspects of the EEMs' usefulness.

4.4.2 Analysis of User Evaluation Data

4.4.2.1 Analysis of User Test Data

The videotape data containing the users' interaction with the IMM and the post-evaluation interview data had to be analysed to extract user problems. In order to reduce the effect of evaluator bias (Jacobsen and John, 1998) in the process of usability problem extraction, it is recommended that two evaluators analyse the user data independently (Lavery, Cockton and Atkinson, 1997). However, a second evaluator was not available, and therefore the researcher decided to create a set of problem extraction criteria to enable more controlled problem extraction. An attempt was made to apply the problem extraction criteria defined by John and Marks (1997), however only three applied to the data from the user tests of the IMM. Therefore, the researcher formulated six additional criteria to accommodate the data from the user tests. Table 4.5 presents the nine criteria that were created for the purpose of user problem extraction. The first three were

adopted from John and Marks (1997), and the last six were defined specifically for this study.

Table 4.5: Criteria for extracting user interaction problems

No	Criterion Description
1	The user articulated a goal but cannot succeed in achieving it without external help from the experimenter.
2	The user articulated a goal and tried different alternatives before achieving it.
3	The user articulated a goal, tried different things and explicitly gives up.
4	The user reports a difficulty in understanding part of the material.
5	The user expresses confusion while trying to achieve a task.
6	The user expresses a dislike regarding a design feature.
7	The user experiences problems with identifying the goal of the presentation.
8	The user omits a segment of the presentation, which is considered important for the completion of their task.
9	The user proposes a design suggestion.

The videotape data and the post-evaluation interview data were then carefully analysed using the above criteria to extract user problems. A problem was identified if the user verbally expressed a difficulty, e.g. the student said that they could not comprehend a task instruction, or when they expressed a goal which they wanted to achieve but their behaviour showed that they could not achieve it, e.g. the student wanted to change the scale of a graph but after a few attempts they still were unable to do so. The results of the problem extraction are given in Section 4.6.

4.4.2.2 Analysis of Knowledge Tests

In order to establish how effective the IMM is in teaching students the concepts of exponential graphs, the pre-exposure knowledge test results were compared with those from the post-exposure tests. Student answers to the post-exposure tests were then analysed in detail to identify which knowledge propositions (see Table 4.2) students had difficulties comprehending. These comprehension difficulties were recorded.

4.4.2.3 Assigning Severity to User Problems

The usability and learning problems encountered by the users varied in the way they affected their performance and behaviour. Therefore, it was important to identify which problems were critical, and which just caused minor obstruction of their tasks. It was also essential to establish how many of the really critical user problems experts were able to predict and how many they missed out. This analysis informed the measures of the *actual thoroughness* and the *incompleteness* of each EEM.

The two independent judges who rated the severity of the predicted problems, were asked to assign severity ratings to the problems experienced by the users, using the severity rating schema presented in Table 4.3. Where inconsistency between the two ratings was observed, the scores were again averaged. In the cases where the average was not a whole number the scores were again rounded to the more favourable rating.

4.4.2.4 Categorisation of User Problems

A variety of problems were identified by the users, and to accommodate them sixteen problem categories were used. Twelve categories are the same as those used to categorise the predicted problems, however it was necessary to formulate four additional categories, which are defined in Table 4.6.

Table 4.6: Additional Problem Categories

Category Name	Category Description
Affordance	A difficulty identifying how to proceed with the interaction, as the student is unsure what part of the presentation affords the actions they would like to perform.
Comprehension	A difficulty with understanding a certain part of the Maths content.
Language	A problem with an English language expression.
Missed interaction	A situation when the user did not engage in interaction or visit a section of the presentation, which is considered important for achieving their tasks.

4.4.3 Validation of Predicted Problems

The procedure for analysing the *validity* of the problems predicted using the EEMs is presented in this section. Two aspects of the effectiveness of the expert predictions were studied, as defined in Section 3.4:

- *Actual validity*: measures the ability of the EEMs to predict valid user problems.
- *Invalidity*: measures the proportion of false alarms predicted during the expert evaluations.

To be able to draw conclusions regarding the two aspects of *validity*, each predicted problem was validated using the two-step validation procedure outlined in Chapter 3: firstly against the users test results and secondly by independent judges. The method of each step is illustrated in Figure 4.3 first and then described in detail.

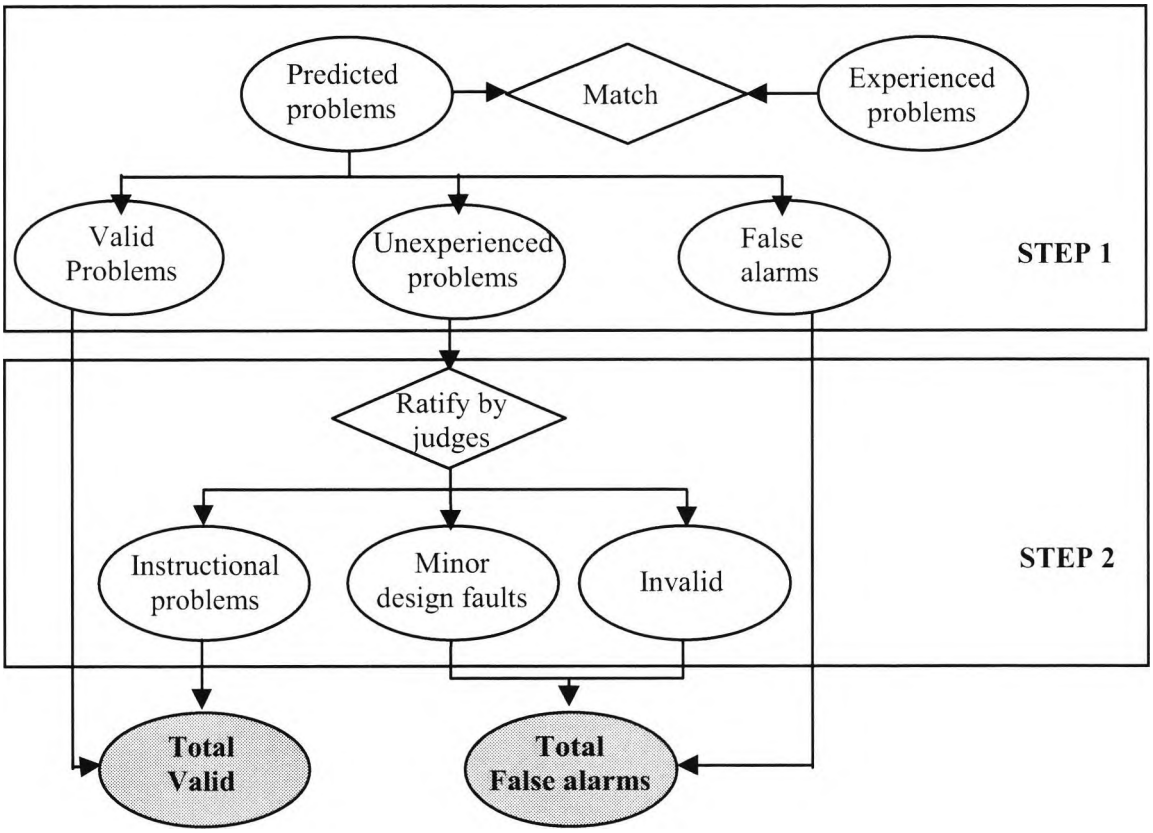


Figure 4.3: Two-Step Validation of Predicted Problems

4.4.3.1 Step 1: Validation against User Test Results

During this step all predicted problems were compared to all user problems experienced during the user tests. The predicted problems were also compared to the comprehension problems identified from the post-exposure knowledge tests in an attempt to validate any instructional problems predicted by the experts. The latter was introduced specially to address the nature of expert evaluations of IMM. This approach to assessing the effectiveness of EEMs is new and has not been used before.

In validating the problems predicted by the experts, it was important to identify which problems could potentially have an effect on target users and which would not have an effect at all. To achieve that, the predicted problem set had to be matched against the user problem set. Heller (1996) and Gray and Salzman (1998), however, report differences between problem definitions of expert and user evaluations due to the different nature of measuring performance. Gray and Salzman (1998) explain that whereas user evaluation methods measure user performance directly, EEMs examine the interface or aspects of the interaction and infer usability problems indirectly.

Therefore, the outcomes of both types of evaluation cannot be viewed as equivalent, and care should be taken in matching the results, in order to preserve the *effect construct validity* of the results (Gray and Salzman, 1998).

To regulate problem matching, Lavery, Cockton and Atkinson (1997) suggest establishing matching rules. The authors also propose using a uniform problem report format to describe predicted and experienced problems to ease problem matching.

In this study, six problem matching rules were defined, which are given in Table 4.7.

Table 4.7: Rules for Matching Predicted and Experienced Problems

No	Problem Matching Rule Description
1	Two problems can be matched if both describe the same fault with the same design feature, although it may be observed in a different part of the software.
2	Two problems can be matched if both describe the same breakdown of user interaction.
3	Two problems can be matched if both describe the same user behaviour.
4	Two problems can be matched if one describes a breakdown and the other describes an effect on users caused by the same design feature.
5	Two problems can be matched if observed user behaviour or performance is considered to be a result of a given design fault specified by the expert. The user performance could include comprehension difficulties.
6	Two problems can be matched if both describe the same or a very similar redesign suggestion.

A problem report format was not used either in the expert or in the user study. As a result, the expert and user problem descriptions varied in their granularity and level of abstraction. Because of this difference, the degree of certainty of each match varies. To overcome this uncertainty, the strength of each match was rated on a scale of 1 to 4, where 1 is a very weak match and 4 is a very strong match, as described in Table 4.8.

Table 4.8: Scale of the Degree of Certainty of Problem Matches

Grade	Description
4	If the predicted problem explicitly specifies a user effect and at least one user experiences this effect; or if all components in the predicted problem are fully described in the user problem report, with or without redesign suggestions.
3	If one PD describes a subset of the other (not counting redesign suggestions), without any contradictions. Also if the predicted one states a cause and the user behaviour or performance could be considered due to this cause (matching rule 5). If one is more general (e.g. refers to navigation buttons) and the other more precise (e.g. specifies particular navigation buttons); or if there is a slight difference in the degree of the effect specified by both problem descriptions (e.g. not interesting and not very interesting).
2	If only a small section of one PD can be linked to a small section of another PD; or if both problems are concerned with the same design feature, however different but related faults are specified with that feature or the faults are at a different level of abstraction.
1	If the level of granularity of both PDs is very different but problems can be considered to be linked. Usually at least one of the PDs does not specify a precise feature or user behaviour.

As a result of the problem matching, three categories of predicted problems were distinguished:

- *Valid* predicted problems are those that were experienced by the users.
- *False alarms* are those that the users rejected, as the users explicitly stated during the post-evaluation interviews that they did not experience certain problems which the experts said they would.
- *Unexperienced* problems are those that were not experienced by the users.

4.4.3.2 Step 2: Ratification by Independent Instructional Experts

The evaluators, particularly the subject-matter experts, identified a set of problems related to the pedagogical effectiveness of the IMM and its ability to effectively represent the Maths concepts. Some of these problems relate to the process of learning, and if compared only to the final learning performance of the students (demonstrated in the post-session knowledge test), they may wrongly be classified as false alarms because users could not be expected to recognise them as problems. This could distort the measure of *validity* of the expert predictions. Therefore, the instructional experts who assigned severity to the predicted problems were asked to ratify them at the same time. Using the severity schema presented in Table 4.3, they identified the major and

medium instructional problems, and minor and cosmetic design faults, as well as rated some predicted problems as invalid.

Their ratings were applied only to the problems not experienced by the users as there was not sufficient user data to reject these problems as false alarms. In this way, any valid instructional problems that have been classified as unexperienced could be identified and classified as valid.

At the end of both validation steps, each predicted problem was classified as valid or false alarm according to the following two definitions (also given in Chapter 3):

- *Valid* if it was experienced by the users and if it was classified as a major or medium instructional problem by the instructional judges although not experienced by the users.
- *False alarm* if it was not experienced by the users and not ratified as a major or medium instructional problem by the instructional judges.

4.4.4 Reverse Problem Matching

To establish the *actual thoroughness* of the EEMs a second problem matching, reverse to the one described in Section 4.4.3.1, was performed. Every user problem was compared to every expert problem to establish whether any aspect of the user problem was predicted by the experts. The same approach to matching the problem sets was applied as the one described earlier in this section, using the same matching rules and match certainty scale.

4.5 Results from the Expert Evaluations

This section presents the results from the expert evaluations in terms of the total number of problems predicted using each EEM, together with their severity ratings and categorisation.

4.5.1 Predicted Problem Sets

A total of 201 unique problems were predicted by the experts across all EEMs. The problems were then divided into EEM-derived and those attributed to the experts' own judgement. The total number of PDs generated during the expert evaluations is shown in Table 4.9, and a full list is given in Appendixes 4.12.1 and 4.12.2.

Table 4.9: Problems Predicted by the Experts Using Each Evaluation Method

EEM	Total number of problems predicted	Number using EEM	% using EEM	Average per expert using EEM	Number from own judgement	% from own judgement	Average per expert from own judgement
MMT	79	56	71%	14	23	29%	5.75
MMCW	34	19	56%	9.5	15	44%	7.5
IMMC	88	70	80%	17.5	18	20%	4.5

As can be seen from the data in Table 4.9, in total 145 problems were predicted using the EEMs. The evaluators using the Interactive Multimedia Checklist identified the most problems attributed to the EEM. The experts using the Multimedia Taxonomy identified fewer problems. Finally, in the case of the Multimedia Cognitive Walkthrough the least number of problems were identified. This could be due to the fact that only two evaluators applied the MMCW, however Table 4.9 also shows that on average these experts identified the least number of problems using the EEM. Although all evaluators specified problems not covered by the EEM they were using, the ones using the MMCW identified the highest proportion of issues outside the scope of the method. This could be due to the narrower scope of the cognitive walkthrough in comparison to the other two methods.

The analysis of predicted problems that follows covers only the problem sets identified using the three EEMs and not those identified through their own experience.

4.5.2 Severity of the Predicted Problems

As seen above, the evaluators identified a large number of problems, and thus it was essential to differentiate between the ones that could have a critical impact on the achievement of the user tasks and those which are minor and cosmetic or not valid usability or instructional problems. During software development, severity ratings can also help the development team decide which problems need to be addressed first and

how much resource to allocate to fix the most serious faults. As explained earlier, in this study, severity ratings were used as part of the method's *cost effectiveness* analysis.

The two instructional judges rated the severity of all predicted problems and detailed results from their rating are given in Appendix 4.12.1. Table 4.10 shows a summary of the number of major, medium, minor and cosmetic problems for each EEM.

Table 4.10: Number of Predicted Problems per Severity Category as Classified by the Independent Judges

Rating / EEM	# Invalid Problems	%	# Cosmetic	%	# Minor	%	# Medium	%	# Major	%
MMT	2	3%	13	23%	25	45%	11	20%	5	9%
MMCW	1	6%	6	32%	7	36%	5	26%	0	0%
IMMC	30	43%	10	14%	26	37%	4	6%	0	0%

The data in Table 4.10 shows that the experts applying the MMT uncovered a few major problems, but they also found a significant amount of cosmetic and minor problems. The experts implementing the IMMC identified predominantly minor and cosmetic issues, and a small number of medium ones. Finally, with the aid of the MMCW a similar number of medium, minor and cosmetic issues were uncovered.

4.5.3 Types of Predicted Problems

As explained in Section 4.4, each predicted problem was also allocated into a problem category that was the closest to the design fault it is concerned with. The bar chart in Figure 4.4 shows how the problems discovered using each EEM get distributed across the problem categories.

From the bar chart, presented in Figure 4.4, it can be seen that the evaluators using the Multimedia Taxonomy identified problems in the most number of categories – in 16 out of 19 (or 84%). Quite a few problems are concerned with content and media design issues, which can be explained by particular emphasis given to these design areas in the taxonomy, especially to individual media resources. The Interactive Multimedia Checklist aided the identification of faults in a similar number of categories – 15 out of the 19 (or 79%). The extensive coverage could be attributed to the diverse nature and the large number of checklist questions included in the method. Finally, the evaluators using the Multimedia Cognitive Walkthrough identified problems in just over a third of

the categories – 7 out of 19 (or 37%). This is not a surprising result since the evaluation technique specialises in cognitive aspects of multimedia presentations. Thus, the majority of the issues uncovered fall into related categories, such as media integration and attentional design. The only slightly surprising result is that although media selection guidelines are included in the method, no such problems were identified. A possible explanation could be that both evaluators thought that the mathematical content does not fit in adequately within any of the information types described in the method. The MMCW distinguishes between six different types of information: physical, spatial, action, role, procedure, and causal. Its authors then provide a set of 'evaluatory guidelines' that explain what medium or combination of media is best suited to represent each type of information. As the evaluators who applied this technique could not fit the Maths concepts into any of the above six categories they had problems using the evaluation guidelines to validate the adequacy of the media selection.

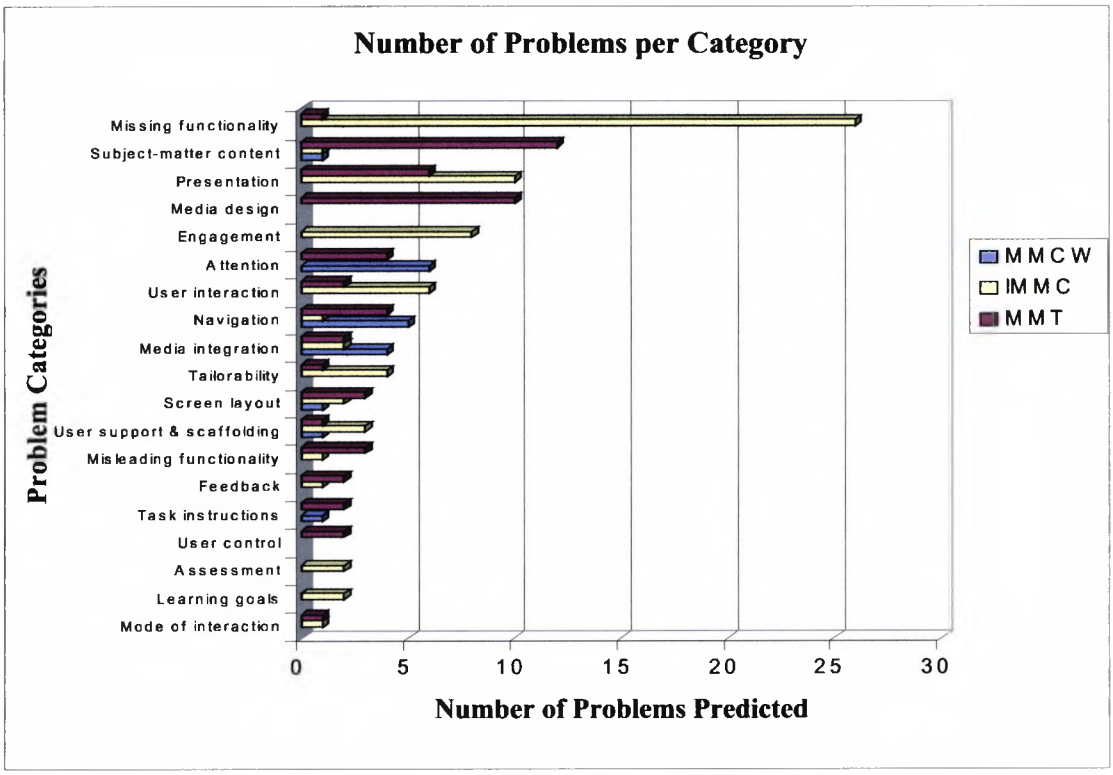


Figure 4.4: Number of Problems Predicted per Problem Category

4.5.4 Reliability of Expert Predictions

4.5.4.1 Multimedia Cognitive Walkthrough

The two MMDs using the MMCW identified respectively 6 and 16 problems, i.e. one of them found nearly three times more faults using the same evaluation guidelines. Furthermore, only 3 problems were identified by both evaluators. This inconsistency could be attributed to the fact that one designer believed that only 30% of the guidelines were relevant to the application, which could explain the small number of problems identified by this evaluator – only 6. Whereas, the other MMD reported that about 70% of the guidelines were relevant to the application. Thus, they used more guidelines and attributed more problems to the EEM. This suggests that some guidelines are specified in a way which allows different experts to interpret them differently. A vivid example of the difference in interpreting the method's guidelines is that two of the same problems identified were based on different guidelines. For instance, both evaluators stated that the graph lines are not revealed gradually, but too quickly for the viewer to read, however each one of them used a different attention guideline to identify the problem.

4.5.4.2 Interactive Multimedia Checklist

Out of the 70 PDs uncovered by the experts using the IMMC 5 were found by all 4 reviewers (7%), 14 problems by 3 of the reviewers (20%), and 19 by 2 of them (27%). This indicates that 38 out of the 70 problems were predicted by more than one evaluator, which accounts for 54% of the total number. However, in terms of the number of problems predicted by each set of specialists the consistency is not so well established. The two MMDs found 28 and 48 problems respectively. The one who identified the larger number of problems had significantly greater experience with developing and evaluating IMM applications than the other designer. The MMDs agreed on their answers to 19 of the checklist questions. Although this represents a fair proportion, this also indicates that some of the evaluators' responses contradicted each another. On the other hand, the subject-matter experts found 23 and 34 PDs respectively, where the difference is less significant. They identified 17 problems in common, which is greater than the MMDs in terms of proportion of the total number of problems predicted. However, it was again observed that some of the responses given by the SMEs contradicted each other.

4.5.4.3 Multimedia Taxonomy

Finally, the two MMDs using the MMT identified 9 and 16 problems respectively. Although they found similar types of problems, e.g. media design, user interaction, navigation, they did not uncover any corresponding problems. The subject-matter experts identified 5 and 18 problems respectively. Although both experts knew the material quite well, they did not identify any corresponding errors with the Maths content or the method of instruction. In particular, one of the SMEs detected errors with the use of mathematical conventions, factual errors with the material, as well as some issues about how logically the content is organised and how comprehensive it is. The other SME, who was quite familiar with the learning needs of learners, found more high-level issues regarding how the material presented in the application helps the development of the necessary skills needed by these learners for their professional development. The MMT is a technique which gives evaluators freedom to comment on aspects of the application they consider important, and as a consequence the problem sets are quite diverse. Therefore, in terms of the similarities of the problem sets only 2 duplicate problems were found across all four problem reports. They were identified by a MMD and a SME, which shows the randomness of the duplication. This leads to the conclusion that expert bias and selective data gathering are strong in this evaluation technique, which can undermine the reliability of the results produced.

4.5.5 Diagnosis of Predicted Problems

Table 4.11 below reveals a summary of the results in terms of the number and the percentage of problems in each of the four diagnostic categories defined in Section 4.4. These categories are fully specified (FS) problem, under specified problem – no effect specified (US-NE), under specified problem – no cause specified (US-NC), and redesign suggestion (RS). Detailed results from the diagnosis analysis of each PD is given in the last column of Appendix 4.12.1.

Table 4.11: Diagnosed vs. Under-Specified Predicted Problems

Diagnostic Category / EEM	FS	% FS	US-NE	% US-NE	US-NC	% US-NC	RS	RS %
MMT	15	27%	38	68%	0	0%	3	5%
MMCW	6	32%	11	58%	2	10%	0	0%
IMMC	2	3%	53	76%	13	18%	2	3%

As can be seen from Table 4.11 the evaluators using the MMCW fully specified the largest proportion of PDs, which is still one third of all problems identified using the method. The experts who used the MMT fully diagnosed a similar proportion of the problems predicted, or just under a third. On the other hand, the evaluators who applied the IMMC fully specified only 3% of all errors. The data in the fourth and fifth columns of Table 4.11 show that the vast majority of the predicted problems do not describe the likely effect the problems may have on users. Similar results have been observed by a number of researchers, who have concluded that expert evaluations tend to reveal problem causes without correlating them to likely effects on users, which is attributed to the fact that experts usually concentrate on analysing specific design features (Gray and Salzman, 1998). Tergan (1998) also believes that a main disadvantage of EEMs for instructional software is their limited power to predict potential learner behaviour or performance.

In real software development situations, it would be sufficient to have a fully defined set of problems for the effective redesign of applications. However, for the purpose of this research, it was necessary to analyse more deeply the reasons behind the diagnostic power of each EEM, and therefore it was necessary to distinguish between diagnosis made based on the EEMs' evaluation criteria and that due to evaluators' own judgement. The evaluation criteria which prompted the generation of the fully specified problems were analysed to establish whether they provide any indication of the problem causes or the effects specified. In order to determine whether there is correlation between the methods' evaluation criteria and the components specified in the PDs, the following rule was used:

> *Correlation exists if the evaluation criterion, which was specified to have prompted the problem, contains explicit and unambiguous reference to design feature/s likely to cause a usability problem and/or the likely effect of such feature/s on users' behaviour, performance, motivation or satisfaction.*

In the case of the MMCW it was found that 3 of the 6 fully specified problems could be attributed to the attentional guidelines to which they relate. These guidelines explicitly specify how users' viewing order is controlled and what objects the users are likely to

focus on first in a multimedia presentation. In the case of the IMMC evaluation, the two fully specified problems could not be attributed to the method, as in both cases the relevant questions simply ask for generic design features to be reviewed. Finally, in the case of the MMT, the problem definitions depend entirely on the expertise of each evaluator.

4.6 Results from the User Tests

This section presents the results from the one-to-one user tests, post-exposure interviews with the users and from the comprehension tests.

4.6.1 User Problem Set

As a result of the video analysis conducted, it was found that the learners encountered 21, 22, 23 and 28 problems respectively, or 94 usability problems in total. A significant number of the problems were duplicating, thus in total 51 unique user problems were found to have been experienced during the users' interaction with the IMM. A complete list of the experienced problems is given in Appendix 4.13, whereas Table 4.12 presents the number of problems found using each problem identification criterion

Table 4.12: Number of user problems per problem identification criterion

Criterion Description	No. of Problems
The user articulated a goal but cannot succeed in achieving it without external help from the experimenter.	4
The user articulated a goal and tried different alternatives before achieving it.	3
The user articulated a goal, tried different things and explicitly gives up.	8
The user reports a difficulty in understanding part of the material.	9
The user expresses confusion while trying to achieve a task.	8
The user expresses a dislike regarding a design feature.	7
The user experiences problems with identifying the goal of the presentation.	2
The user omits a segment of the presentation, which is considered important for the completion of their task.	5
The user proposes a design suggestion.	5

As a result of the analysis of the post-exposure knowledge tests it was found that the students experienced a number of comprehension problems. The results from the post-exposure knowledge tests are presented in Table 4.13 against each knowledge proposition, as defined in Table 4.2. The 20 knowledge propositions are shown in the second column of Table 4.13, and the number of learners who demonstrated

understanding of the respective proposition is given in the final column. The results show that the learners managed to grasp the concepts of horizontal and vertical reflection (propositions 3, 4 and 6) and horizontal translation (propositions 15 and 16). However, they had particular problems understanding the principles of scaling (propositions 7-14), as well as some principles of translation, especially the method of calculating the distance in horizontal translation (proposition 19). As can be seen from the last column of Table 4.13, the learners experienced difficulties comprehending 13 knowledge propositions in total, i.e. at least two of the learners did not grasp the essence of the proposition.

As a result of all types of user tests it was found that the users encountered 51 usability problems and 13 comprehension ones, i.e. 64 problems in total.

Table 4.13: Comprehension Test Results

No	Knowledge Proposition	No. of learners with correct answers
1	There are two ways of representing exponential functions: $y=e^x$ and y=exp(x).	4
2	There are three fundamental exponential functions: e^x, 2^x and 10^x , and depending on their base factor, the corresponding graphs are more or less steep.	2
3	During *horizontal reflection* the equation of y=exp(x) is transformed to y=-exp(x).	4
4	During *horizontal reflection* the graph of y=exp(x) is reflected in the horizontal axis or turned upside down.	3
5	During *vertical reflection* the equation of y=exp(x) is transformed to y=exp(-x).	2
6	During *vertical reflection* the graph of y=exp(x) is reflected in the vertical axis or represents a mirror image.	3
7	During *horizontal scaling* the equation of y=exp(x) is transformed to y=exp(kx).	2
8	During *horizontal scaling* the graph of y=exp(x) is moved, but the vertical intercept remains fixed. If k>1 line becomes steeper, if 0<k<1 line becomes less steep.	2
9	During *vertical scaling* the equation of y=exp(x) is transformed to y=k exp(x).	1
10	During *vertical scaling* the graph of y=exp(x) is moved and the vertical intercept is equal to k. If k>1 line becomes steeper, if 0<k<1 line becomes less steep.	1
11	*Horizontal scaling with negative factor* (i.e. -k) is a composition of horizontal scaling with factor k and vertical reflection. Therefore y=exp(x) is transformed to y=exp(-kx).	0
12	During *horizontal scaling with negative factor* the graph of y=exp(x) is horizontally scaled and reflected in the vertical axis at the same time.	1

Table 4.13: Comprehension Test Results (continued)

No	Knowledge Proposition	No. of learners with correct answers
13	*Vertical scaling with negative factor* (i.e. –k) is a combination of vertical scaling and horizontal reflection. Therefore y=exp(x) is transformed to y=-k exp(x).	2
14	During *vertical scaling with negative factor* the graph of y=exp(x) is vertically scaled and reflected in the horizontal axis at the same time.	1
15	During *horizontal translation* the equation of y=exp(x) is transformed to y=exp(x+k).	3
16	During *horizontal translation* the graph of y=exp(x) is moved to the left if k>0, and moved to the right if k<0.	3
17	During *vertical translation* the equation of y=exp(x) is transformed to y=exp(x) + k.	2
18	During *vertical translation* the graph of y=exp(x) is move up with k units if k>0, and moved down with k units is k<0.	2
19	In *horizontal translation* the distance between the two graphs is equal to the value of k.	1
20	In *vertical translation* the distance between the two graphs is equal to the value of k.	4

4.6.2 Severity of User Problems

As a result of the severity rating performed by the instructional judges, all user problems were distributed as shown in Table 4.14.

Table 4.14: Number of Experienced Problems per Severity Category

Severity Category	Number of User Problems	%
Major	11	17%
Medium	20	31%
Minor	26	41%
Cosmetic	7	11%
Total	**64**	

The figures presented above indicate that the majority of the problems experienced by the users are of medium or minor severity, however it was found that nearly a fifth of the problems are major, hindering their comprehension and the achievement of their learning tasks. The users also identified a small number of cosmetic issues.

4.6.3 Types of User Problems

The results from the categorisation of the user problems are shown in Figure 4.5. From the bar chart it can be seen that the users encountered predominantly comprehension and other content related problems, together with some missed interaction and affordance difficulties. These problems directly relate to the learning tasks they were asked to perform and to how the application supports or hinders their achievement. The users also identified six media design issues, reporting aspects regarding the design of the icons or the quality of the graphics.

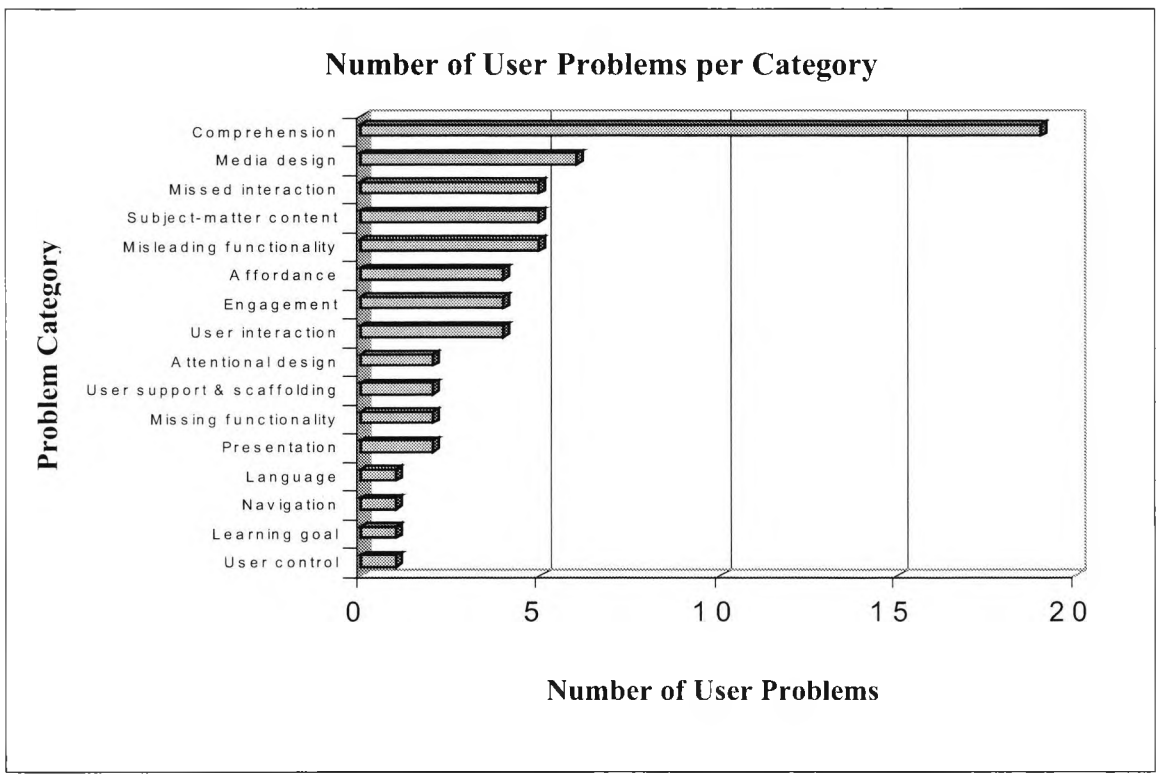


Figure 4.5: Experienced Problems by Category

4.7 Validation of the Predicted Problem Sets

4.7.1 Results from Validation Step 1

As explained in Section 4.4.3, in order to identify whether the predicted problems were actually experienced by the users, every predicted problem was compared to every user problem to establish whether they match according to the matching rules described and how strong the match is. The results of the problem matching are given in the last but one column in the table in Appendix 4.12.

As can be seen in Figure 4.6, it was found that 52 of the 145 problems predicted using the EEMs were actually experienced by the users, or 36%. The predicted problems were also compared to the answers given by the users during the post-evaluation interviews. It was found that certain user experiences contradicted some of the predicted problems. For examples, two evaluators predicted that users may feel lost when navigating within the application, whereas all users reported that it was easy for them to find their way through the application and never felt lost. Based on this analysis, 14 predicted problems were classified as false alarms. The remaining 79 predicted problems were neither experienced by the users nor rejected by them during the post-evaluation interviews. The unexperienced and the rejected problems account to 64% of the total number of predicted problems, which represents a quite significant proportion.

4.7.2 Results from Validation Step 2

The results from the severity rating and validation of the predicted problems by the two independent judges were given in Table 4.10. The results presented in Table 4.10 were compared to the 79 unexperienced problems to identify whether any of them are valid instructional problems. As can be seen from Figure 4.6 the judges classed 10 unexperienced predicted problems as major or medium instructional problems. The judges also classified 40 unexperienced predicted problems as minor or cosmetic design problems. Finally, the judges also rated 29 unexperienced predicted problems as invalid.

4.7.2.1 Instructional Problems

This set of problems was classified as major and medium by the independent judges. They include pedagogical and learning design issues, which fall into the following three categories:

- Content: many predicted problems concerned the accuracy and completeness of the Maths content and the notation used. Such problems were identified by the SMEs. For instance, two SMEs who evaluated the EMM application identified a mistake in one of the equations of Vertical Scaling.
- Assessment: issues regarding the adequacy of different monitoring and assessment techniques were also identified by the SMEs.
- Intelligent tutoring facilities: issues concerned with whether different expert system facilities are required to support learners.

Some of these instructional problems can potentially point to learning difficulties. Therefore, these problems were reinstated as valid, as due to their nature user tests cannot be expected to identify them.

4.7.2.2 Minor Design Faults

These issues are concerned with the design of navigation buttons, and the quality of the text and graphics presentation. The users may have noticed some of these issues, however they did not report them. Such minor and cosmetic issues were identified predominantly by the MMDs using all EEMs. They account for 28% of the total number of problems predicted using the EEMs, which is a fair proportion. Despite the fact that the two judges identified these problems as valid, they were classified as false alarms as there is no evidence that any of these issues had an impact on the users' interaction and the achievement of their tasks. However, had more users been involved, a greater number of such problems could have been validated. Therefore, the number of false alarms given below may be inflated.

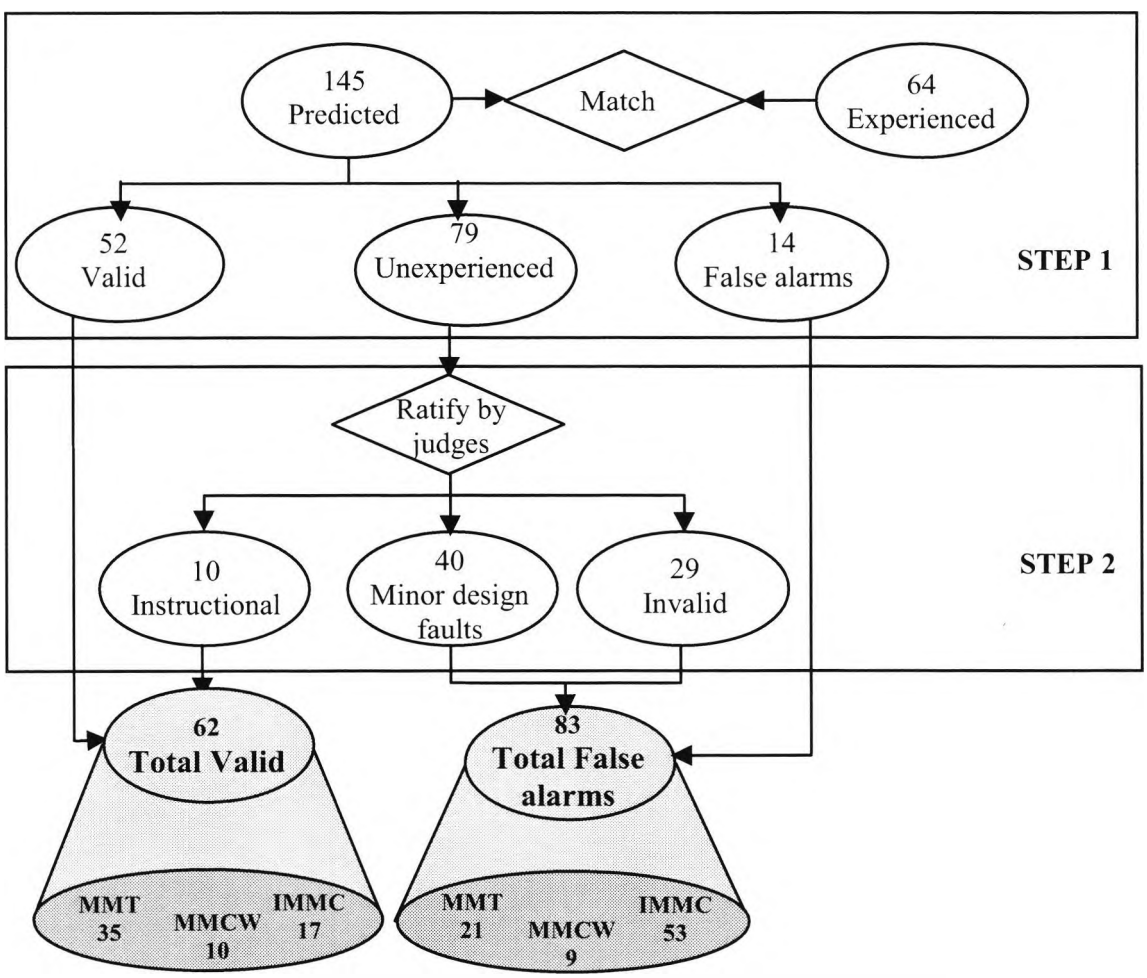


Figure 4.6: Validity Calculations of Predicted Problem Sets

4.7.3 Final Validation Results

As a result of the two validations, it was found that in total 62 of the problems predicted by the experts using the EEMs are valid, or 43%, as shown at the bottom of Figure 4.6. Therefore, it was found that over half of the predicted problems are false alarms, namely 83 are false alarms, or 57%.

The particular intermediate and final validity calculations for each EEM are presented in Table 4.15.

Table 4.15: Intermediate and Final Validity Calculations for each EEM

	MMT	**MMCW**	**IMMC**
Total problems predicted	56	19	70
Total problems experienced by users	27	9	16
Total false alarms by user data	*7*	*3*	*4*
Total instructional problems	8	1	1
Total false alarms by independent judges	*14*	*6*	*49*
Grand total valid problems	**35 (63%)**	**10 (53%)**	**17 (24%)**
Grand total false alarms	***21 (37%)***	***9 (47%)***	***53 (76%)***

As can be seen in Table 4.15, 35 of the 56 (or 63%) problems identified using the MMT, and 10 of the 19 (or 53%) problems predicted using the MMCW are valid. A considerably smaller proportion of the issues discovered using the IMMC were classed as valid, namely 17 of the 70 problems, or 24%. Therefore, between 37% and as much as 76% of the problems predicted using the EEMs were found to be false alarms, which is a considerable proportion.

It is important to identify the nature of the false alarms identified by the EEMs, given their high proportion. Therefore, an analysis of the false alarms predicted using each EEM is presented in the next sub-section. The nature of the correctly predicted problems will then be discussed in the following section.

4.7.4 Analysis of False Alarms

Figure 4.6 shows that 14 issues the experts identified as problematic but were rejected by the user test data, particularly based on the evidence given in the post-evaluation user interviews, which were designed to validate the predicted problems. Most such false alarms were due to the experts making wrong assumptions about users' sense of orientation within the application and the information presented, their control over the application and preferences regarding customisation of program settings. Figure 4.6 also shows that further 29 issues predicted by the experts were classified as invalid and 40 classified as minor issues by the independent judges. The following paragraphs discuss the potential reasons why such false predictions were made.

4.7.4.1 Analysis of False Alarms Predicted Using the MMCW

Nine issues, or 47% of the total number of problems predicted using the MMCW, were classified as false alarms. The evaluator who identified all of these issues had some multimedia design and testing experience, but he had not used any formal evaluation methods before and did not have sound cognitive psychology knowledge. This evaluator seems to have interpreted the guidelines in a general sense, rather than in their specific meaning within the context of cognitive processing of MM presentations. For instance, he applied the term 'focus' in its broad sense rather than to signify directing user's attention to different audio and visual media. The same evaluator also did not seem to be able to determine easily which guidelines were applicable to the IMM software and tried to apply most of them. As a result, he generated nearly three times as many problems than the more experienced evaluator (16 and 6 respectively), however 25% of them were classified as false alarms. This evidence suggests that evaluators with limited cognitive psychology knowledge and limited usability evaluation experience require more training in the method to be able to apply it effectively. An example false alarm is that the MM designer suggested that "there is never a direct indication of where to advance to next". However, none of the users reported that they had difficulties deciding where to progress next. On the contrary, when asked during the post-evaluation interview all users reported that they found it easy to progress to the next page. As can be seen from Figure 4.7 the application is designed to provide linear navigation within the section, and most students worked through the presentation in the suggested order.

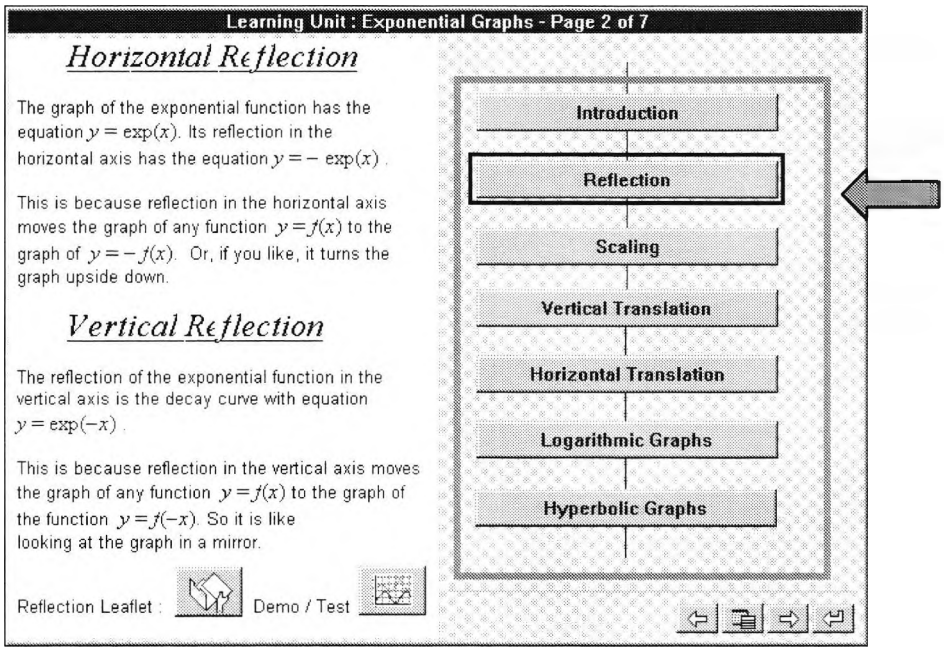


Figure 4.7: Sample False Alarm predicted by the Experts

4.7.4.2 Analysis of False Alarms Predicted Using the IMMC

As shown in Figure 4.6, the largest number of false alarms were found amongst the usability problems predicted during the IMMC evaluations, namely 53 or 76% of the total number of problems predicted. A major factor found to influence the validity of the predictions was the nature of the checklist questions. Two different kinds of questions can be distinguished, which differ in the level of scientific knowledge and the level of subjective judgement required to answer them. One kind of question requires the evaluators to make predictions regarding how an aspect of the IMM will influence the users, their motivation, attitude and behaviour. An example of such question is: "can the user identify with the goals and objectives and build their own personal plan for achievement?". Such questions require pedagogical knowledge and a considerable amount of subjective judgement on the part of the evaluators, who did not constitute typical users and thus would have had difficulties predicting users' attitude and behaviour with accuracy. The data shows that 75% of the false alarms predicted using this method were incurred when answering this type of question. For instance, the SMEs presupposed that learners' attention and concentration could not be maintained consistently, which was not the case with the students.

Another set of questions simply asks the evaluators to check for the presence of a good design feature, e.g. "does the product monitor user performance?". Such questions

require a simple 'yes' or 'no' reply, and the level of subjectivity and the level of specialised knowledge required to answer them are low. Some of the replies to such questions were classed as false alarms by the independent judges. This occurred because some good design features not present in the application were considered irrelevant or inapplicable to the IMM application. For example, some of the experts identified that the software does not offer features of an expert system. As the evaluators who took part in the study were not familiar with the design rationale of the Maths software, they could not make such decisions. Thus, this method is more suitable for use by evaluators who have sufficient knowledge of the objectives and the design rationale of the application to be evaluated.

4.7.4.3 Analysis of False Alarms Predicted Using the MMT

Finally, 21 problems or 37% of the total number of problems uncovered during the MMT evaluations were classed as false alarms. Evaluators using the MMT are given freedom to comment on aspects of the multimedia interface they feel are appropriate and to provide their own judgement. It was found that all false alarms were generated by the least experienced MMD. For example, the evaluator suggested that it is difficult to feel in control of the application, which was not experienced by the users. The same MM designer also identified the false alarm depicted in Figure 4.7. These results propose that the validity of the predictions made using the Multimedia Taxonomy greatly depends on the evaluators' expertise and experience in usability evaluation. Thus, the MMT is more suited for use by more experienced evaluators.

4.8 Actual Thoroughness of Predicted Problem Sets

The *comparative thoroughness* of the three EEMs was examined in Section 4.5. This section discusses the *actual thoroughness* of the expert predictions, as defined in Chapter 3, measuring the number and nature of the user problems predicted using each EEM. It is also essential to identify any user problems the experts were unable to predict. This will provide an indication of the *incompleteness* of the expert predictions. Two aspects of such unidentified user problems will be discussed, namely their severity and their nature.

The results of the problem matching are depicted in Figure 4.8. As can be seen from Figure 4.8, exactly half of the learner problems were predicted by the experts, whereas the other half eluded the evaluators' attention. It can also be seen that of the 11 major problems the experts only managed to predict 5 or 45%. Which means that 55% of the really critical learner problems remained uncovered by the expert evaluations. On the other hand, the experts were more successful at predicting the medium severity ones, and less successful at identifying the minor and cosmetic problems. Desurvire (1994) reports similar findings when user test data was compared with results from usability inspections.

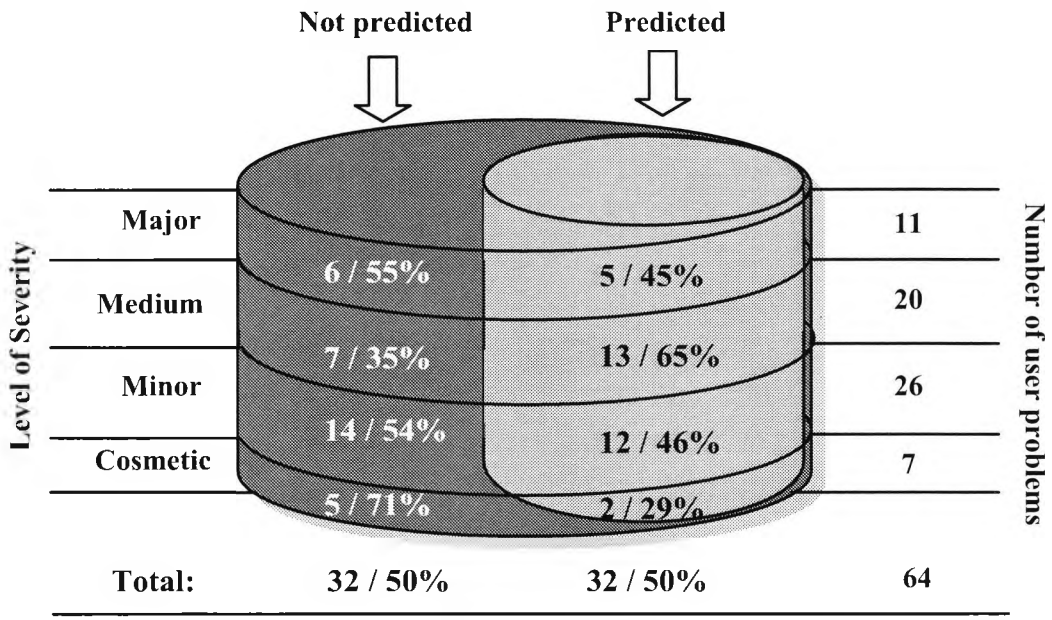


Figure 4.8: Learner Problems Predicted and Unidentified by the Experts

4.8.1 Nature of Correctly Predicted User Problems

Figure 4.9 shows the number of user problems predicted by each EEM. Based on the figures presented in Figure 4.8, the *actual thoroughness* of each EEM was calculated, which is as follows:

- MMT = 34.38%
- MMCW = 14.06%
- IMMC = 14.06%

These figures reveal that all the expert evaluations appear to be patchy in predicting real usability and comprehension problems. Out of the three methods, the MMT predictions covered more of the user test and comprehension problems than the others, but still missed out 66% of them. The IMMC and the MMCW seem to be equally poor at

predicting problems the user encountered, as 86% of the user problems were not identified by the experts using these techniques. The following is a discussion of the type of user problems which the experts were able to predict, followed by an analysis of the nature of the missed out problems.

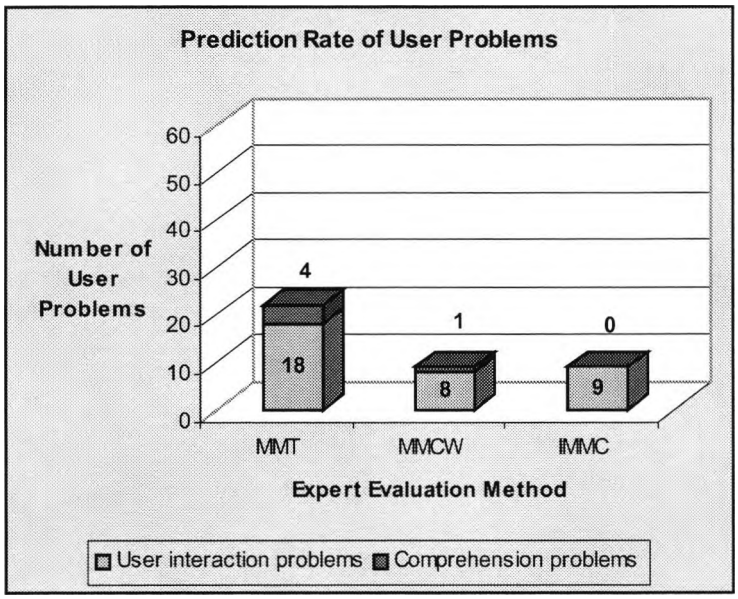


Figure 4.9: Coverage of User Problems

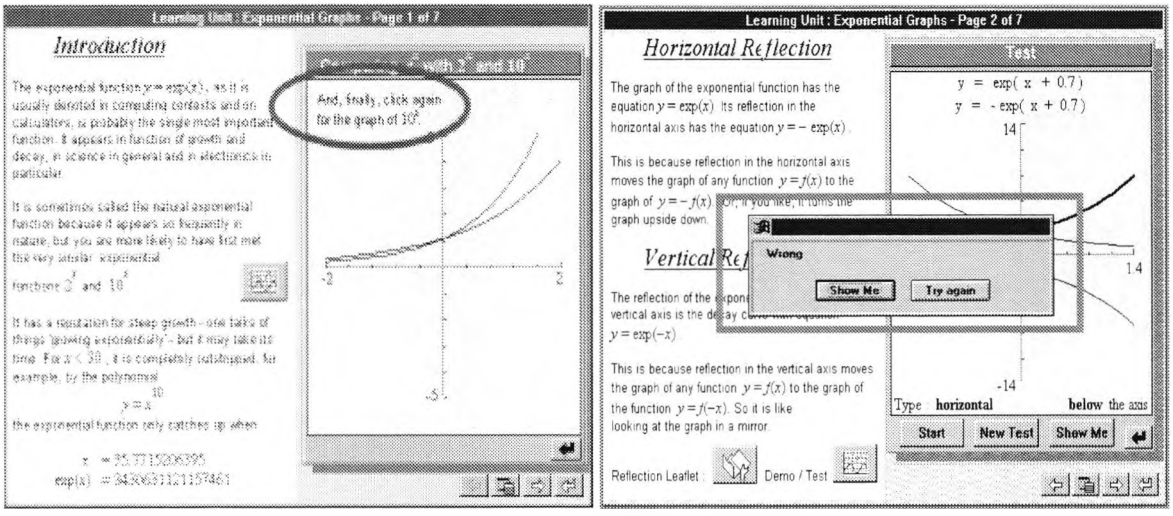
4.8.1.1 Nature and Severity of the Problems Predicted Using the MMT

The experts who used the MMT predicted problems in 13 of the 16 categories of user problems, including issues related to *affordance*, *media design*, *comprehension* of the Maths content and *navigation*. The evaluators uncovered all 5 major problems that the user encountered, together with 8 medium, 5 minor and 3 cosmetic issues.

One area where both experts and users identified usability problems is *affordance*, which encompasses difficulties relating to users not being able to identify which part of the presentation affords certain actions or what action a particular button affords. An example of such a problem is shown in Figure 4.10 (a), which illustrates that after reading the instruction circled the students had difficulty identifying where to click for the graph of 10^x. The MMDs also focused on problems with the *design and appearance of the media resources* used, such as the design of the graphics, graph lines, quality of the icons and the pop-up message boxes. This kind of issue was also identified by the users. Finally, the MMDs also identified some problems with the *navigation* within the application, which did cause confusion to the users.

4.8.1.2 Nature and Severity of the Problem Predicted Using the IMMC

With the use of the IMMC 5 medium and 4 cosmetic user problems were identified. These problems fall into 7 different categories, including some issues regarding *user engagement*, *learning goal*, *scaffolding, screen layout* and *media design*. For example, the experts predicted that the IMM application is not very challenging, which the learners agreed with. The MMDs also identified a problem with the placement of pop-up windows, illustrated in Figure 4.10 (b), which the users found to obstruct the execution of the tests.



(a) An example of an *affordance* problem (b) An example of a *screen layout* problem

Figure 4.10: Sample Screens from the Maths Application Illustrating Typical User Problems Successfully Predicted by the Experts

4.8.1.3 Nature and Severity of the Problem Predicted Using the MMCW

Finally, during the MM cognitive walkthroughs 1 major, 6 medium and 2 minor user problems were predicted, which fall into 7 different categories. The evaluators were able to identify issues regarding users' *attention* to important information, *screen layout*, *affordance,* and the Maths *content*. For instance, the experts predicted the problems depicted in Figure 4.10. They further spotted some problems with synchronising time-varying media resources, such as animated text which changes too quickly for the users to read.

4.8.2 Nature of Unidentified User Problems

Amongst the user problems that the experts failed to identify are two thirds of the comprehension difficulties (namely 9 out of 13, or 69%) and just under half of the usability problems (namely 23 out of 51, or 45%) the students experienced during their interaction with the IMM application. Therefore, the experts had difficulty identifying potential comprehension problems, but they were more successful at predicting usability problems which the learners encountered. These results support findings by other researchers, such as Tergan (1998), who reveals that one of the major limitations of checklist-based evaluation techniques is their limited power to assess the instructional effectiveness of educational software. A detailed analysis of the reasons why experts failed to identify the learner comprehension problems is given in Chapter 5.

In particular, the problems which the experts could not predict fall into five categories: *learning support, comprehension, missed interaction, misleading functionality, and user dialogue.*

4.8.2.1 Learning Support

Learning support problems deal with how much explanation of the material the students required. This greatly depends on the students' prior knowledge. Most students requested more help with Scaling and Translation, especially Scaling, since they had no previous knowledge of these concepts. Although before the evaluation sessions the experts were told to assume no or little prior knowledge of the subject matter on the part of the students, none of the experts could envisage where students may need further explanation of the material. Furthermore, none of the evaluation methods explicitly asked the evaluators to consider students' prior knowledge in order to identify such issues.

4.8.2.2 Comprehension

The *comprehension* problem category describes which parts of the material the students had problems understanding. Although the experts identified some areas of the material which could potentially cause such difficulties to students, they missed out a significant number of them. One factor found influencing the comprehension of the three types of transformation is the varying complexity of the Maths material. The higher the complexity of the material the greater the cognitive task requirements were on the

students. Reflection was found to be the simplest, the principles of Translation were slightly more complex, and those of Scaling were the most complex of the three. The comprehension test results showed that all students grasped the concepts of Reflection, the majority of them got the Translation right as well, however most of them experienced difficulties with understanding Scaling. None of the EEMs suggests that the complexity of the material or the cognitive task requirements should be considered, and none of them correlates these aspects to how media resources could be used and designed to represent complex concepts in order to enable students to comprehend them more easily.

4.8.2.3 Missed Interaction

*Missed interaction*s are situations where the users did not perform an interaction that is considered important for achieving their learning tasks. One such situation arose on the Horizontal Reflection screen, illustrated in Figure 4.11, where a student skipped the test on Reflection, which would have helped them reflect on what they have learned about it. Such situations occurred predominantly because the learner's attention was not explicitly drawn to important parts of the presentation. As can be seen from Figure 4.11, the icon to start the test is placed at the bottom right-hand corner of the main presentation screen where the learner is not likely to look very often.

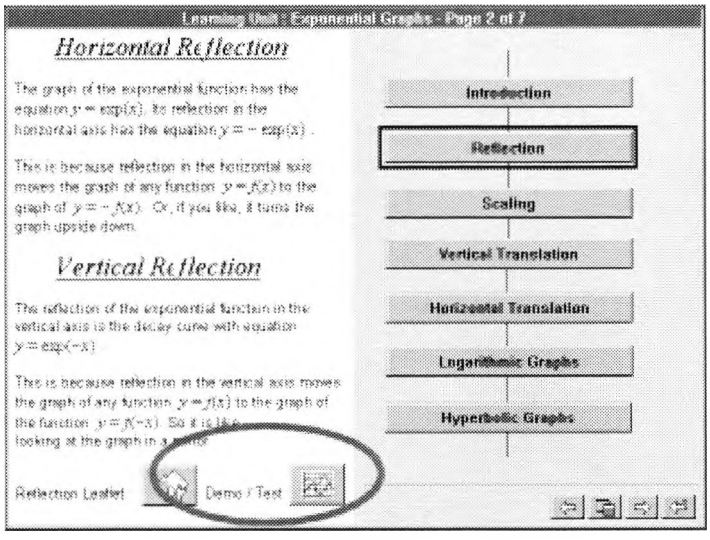


Figure 4.11: An Example of a *Missed Interaction* Not Identified by the Experts

4.8.2.4 Misleading Functionality

This category describes difficulties that the users experienced with identifying what functions are exhibited by certain user interface components. In these cases the users

assumed functionality which was different to the actual one exhibited by the design components. For instance, most users were confused between the function of two buttons: 'start' and 'demo'. They assumed that by pressing on the 'demo' button more graphs will be displayed, whereas they were expected to first click on the 'start' button to begin a new demonstration.

4.8.2.5 User Dialogue

Finally, *user dialogue* encompasses a set of difficulties with user input and output. An example of such a difficulty is "did not initially understand a prompt to input a parameter while performing a test".

4.9 Discussion of the Effectiveness of the Three EEMs

The following sections present how the three EEMs compare in each of the eleven effectiveness criteria defined in the effectiveness framework, in Chapter 3. Particular emphasis will be given on the results regarding the most important measures of the methods' effectiveness: their validity, thoroughness, reliability, and cost effectiveness. The implications of the results for the usability evaluation process of IMM will also be discussed. A summary of the results for each criterion across the three EEMs will then be provided in Table 4.18.

4.9.1 Validity

The validity is the most important measure of an EEM's effectiveness, as it reveals the ability of the method to predict real user problems that will impact users, their behaviour, their ability to achieve the learning goals, and their satisfaction. Such information can then inform instructional software designers in making decisions to effectively rectify design faults and enhance the quality of the software design before it is released.

The problem sets predicted by the experts using the three EEMs investigated were validated against the results from the user tests, as well as by the two independent instructional judges. The validity ratios for each EEM are:

MMT:	63% valid problems	37% false alarms
MMCW:	53% valid problems	47% false alarms
IMMC:	24% valid problems	76% false alarms

The ratios suggest that the MMT seems to be the most valid EEM, despite the fact that quite a few false alarms were identified. The MMCW also showed poorer validity of the predictions compared to the MMT as only 53% of the problems predicted were valid. The IMMC proved to be the least valid of the three EEMs, as only a quarter of the problem predicted were valid.

On the other hand, between 37% and as much as 76% of the problems predicted using the three EEMs were false alarms, identifying issues which had no impact on the users and the achievement of their learning tasks. This includes 28% of all expert predictions which were classified as minor design issues by the independent judges, which also did not influence the users' interaction. These figures undermine the usefulness and credibility of the predictions made, and can give a cause for concern in usability practitioners. As comprehensive validations of predicted problems are not possible during real software development, there is a danger that developers could spend valuable time and effort redesigning software in response to such expert evaluations, without making any improvement to the usability and with a risk of introducing new problems. To improve the validity of expert evaluation methods, understanding of the reasons for generating false alarms is needed and measures for improvement should be taken.

4.9.2 Thoroughness

Thoroughness is also a very important indicator of an EEM's effectiveness, as it measures how comprehensive expert predictions are.

In terms of the *comparative thoroughness* of the EEMs, the evaluations made with the MMCW appear to be quite narrow in scope compared to those of the MMT and IMMC, as depicted in Figure 4.4. The MMT and the IMMC seem to be quite comprehensive, identifying problems in 84% and 79% of the design areas respectively. On the other hand, the results showed that the MMCW could only identify problems in 37% of the design areas covered by all methods. As discussed in Section 4.5.3, the comparative

thoroughness of the techniques is directly related to the scope of the evaluation criteria included in each method. In the case of the MMCW, it needs to be acknowledged that two evaluators rather than four applied the EEM, which would have also reduced its comparative and actual thoroughness.

The ratios of the *actual thoroughness* of the three EEMs were calculated as:

- MMT = 34.38%
- MMCW = 14.06%
- IMMC = 14.06%

The ratios reveal the proportion of the real user problems predicted by each EEM. These figures are alarming as they indicate that all expert predictions were quite patchy in identifying real usability and learning problems. As revealed in Section 4.8, the experts failed to predict nearly half of the most critical user problems, including 69% of the comprehension problems experienced by the learners. This evidence suggests that usability practitioners cannot rely solely on expert predictions to identify all usability problems, and should employ a combination of expert and user testing for a through investigation of the usability of IMM designs, particularly for assessing their instructional effectiveness.

The poor thoroughness of the expert predictions can be characterised by the fact that the experts emphasised aspects which are not so pertinent yet missed out important ones. The methods predominantly focused on design features and did not link these features to likely user behaviour and performance outcomes. Thus, further empirical investigations are needed regarding how to relate intrinsic design features to user performance and behavioural outcomes. The poor prediction rate of learner comprehension problems is another disturbing pattern found. This may be due to the fact that the complexity of the learner knowledge construction processes and the context of learning were not adequately addressed. In order to improve EEMs' thoroughness, further research is required into how the design of MM influences users' interaction and especially their learning processes. Effective evaluation techniques then need to be developed that embody such theories and empirical work in order to improve the EEMs' ability to predict what the likely user behaviour and performance with IMM could be and where difficulties could occur.

4.9.3 Reliability

Reliability is another essential measure of an EEM's effectiveness, as it indicates whether the same evaluation results can be produced under the same conditions, i.e. whether different evaluators tend to find similar number of problems and corresponding problems. Table 4.16 summarises the reliability findings regarding each EEM.

Table 4.16: Reliability Findings for each EEM

Usability Evaluation Method	Number of problems per evaluator		Number of corresponding problems
MMCW	MMD1	6	
	MMD2	16	3
IMMC	MMD1	28	
	MMD2	48	19
	SME1	23	
	SME2	34	17
MMT	MMD1	9	
	MMD2	16	0
	SME1	5	
	SME2	18	0

The results show that the MMT appears to have the poorest reliability of the three EEMs, both in terms of numbers of problems predicted and the similarities of the problem sets. In this study no contradictions were found between the evaluator predictions, but also no similarities were discovered between the predictions made between each type of evaluators. Therefore, depending on the number of evaluators a long list of problems could be produced, which may be difficult to organise and prioritise.

The MMCW showed some similarities between the problems identified by the experts, although one of them discovered nearly three times as many problems as the other. Finally, the IMMC appears to have greater reliable than the other two EEMs as 54% of all problems were uncovered by two or more evaluators. This figure shows that potentially the greatest consistency was observed in the IMMC problem reports. However, some contradictions were found between the problems predicted by the two MMDs and the two subject-matter experts who applied the IMMC. Prior experience with designing IMM and the diversity of their professional background were both found to be responsible for the inconsistency.

4.9.4 Diagnostic Power

The diagnostic power measures how comprehensively each usability problem has been specified in terms of whether its cause/s and the likely undesirable effects on the users' behaviour and performance have been defined.

The data presented in Section 4.5.5 showed that the diagnostic power of all three methods is low. In particular, 32% of the problems predicted using the MMCW and 27% of those identified using the MMT were fully specified, whereas only 3% of the errors defined using the IMMC were fully specified. It was found that the majority of the predicted problems do not specify the likely effect a design fault may have on the target users. The analysis conducted revealed that neither the IMMC nor the MMT explicitly support experts in predicting such effects. Only in the case of the MMCW are experts supported by some of the guidelines in considering the cognitive implications of design features on users, but it remains to the discretion of the evaluator to specify them or not.

A major implication of the failure to specify the likely effect a problem may have on the target users is that software designers and developers would not be able to easily assign severity to the problems and prioritise which ones need to be addressed urgently. An implication of under-diagnosing the cause of the problem, which was observed to a small extent in the cases of the MMCW and the IMMC predictions, is that software developers would not be able to determine what redesign solutions will effectively rectify the problems identified. Both limitations ultimately have an effect on the quality of the user interfaces developed.

4.9.5 Tailorability to Context

As described in Chapter 3, an effective expert evaluation method should generate issues tailored to the potential context of use of the application. Such tailorability has a profound effect on the validity of expert predictions. In order to measure the extent to which the three EEMs are effective in achieving this, in the post-evaluation questionnaire the experts were asked to specify what contextual factors they considered and which were prompted by the method. The experts were also asked to specify

whether the EEM encouraged them to modify the existing evaluation criteria to match the particular context of use.

The evaluators who applied the MMCW said that the method prompted them to consider two contextual factors: cognitive task requirements and aspects of the Maths content. In particular, the cognitive aspects of attending to information presented in different media concurrently were considered. Their effect on users' focus was explicitly specified in the attentional guidelines provided, thus it can be classified under *direct* context consideration. This led to the identification of valid attentional design problems, some of which were experienced by the users.

In the course of evaluating the application using the IMMC, all four experts stated that they considered several contextual factors, including users' prior knowledge in Maths, their learning styles, aspects of the target learning environment and the subject-matter, and also how students learn with the multimedia environment. The last one was considered only by the SME and the instructional specialist. These factors were prompted by the checklist questions, however as the questions do not explicitly refer to possible implications of the design for the users, the consideration of the context-dependent factors is *indirect* in the case of the IMMC. This applies only to a subset of the checklist questions. However, the majority of the questions do not support tailorability either directly or indirectly.

While using the MMT, the experts specified that they were partly led by the taxonomy to consider certain user characteristics, such as the users' prior knowledge in Maths, age and computer literacy. Furthermore, the method suggested that evaluators, such as SMEs, should consider the subject matter content, however it does not recommend the consideration of any particular aspects of the content or how effectively it was presented using the different media resources. Therefore, it is left to the initiative of the expert to decide what aspects are relevant to the context of use of the application under consideration. The method also invites the evaluators to generate evaluation questions tuned to the perceived context of use of the software via the creation of 'evaluation protocols' including questions relevant to the application under consideration. However, none of the reviewers created such context-tailored protocols explicitly, as they reported that the procedure was tedious.

4.9.6 Theoretical Validity of Evaluation Criteria

It was argued in Chapter 3 that evaluation criteria must be grounded in empirical evidence and relevant theoretical work to be valid, as this will impact the validity of the predictions made using such criteria. From the documents provided by the authors of the EEMs, it was found that the guidelines included in the MMCW were based on empirical tests performed by the authors and a cognitive framework of MM (Faraday and Sutcliffe, 1997). Theoretical research into the pertinent aspects of MMUI design had guided the formulation of the MMT (Heller and Martin, 1999c). Finally, Barker and King (1993) state that the IMMC was based on both learning theory and practical IMM development experience. Therefore, the predictions made using the three EEMs were made based on theoretically founded or empirically tested evaluation criteria, which supports their validity. This suggests that the poor validity of the expert predictions is due to factors not wholly related to the validity of the evaluation criteria themselves.

4.9.7 Cost Effectiveness

Cost effectiveness is a particularly important factor in selecting an EEM to be adopted in a particular project. When deciding on an EEM to apply, software developers need to look at a number of factors, including the resources required, such as the time needed and the number of evaluators required. They also have to estimate the potential benefits of the evaluation for their projects.

As described in Chapter 3, when measuring the cost effectiveness of an EEM, the severity scores, representing the number of significant problems identified, need to be taken together with the number of false alarms, and juxtaposed with the time spent using each evaluation technique. The severity scores were calculated by multiplying the number of valid problems identified in each severity category by the corresponding severity grade (ranging from 1 to 4). The severity grades given in Table 4.10 were used. For example, 5 medium severity problems were predicted using the MMCW, therefore 15 (5 x 3) was added to the severity score of the MMCW. The same was done for all valid minor and cosmetic problems predicted using the method. The results of the cost-benefit estimations are given in Table 4.17.

Table 4.17: Cost-benefit Ratio Calculations

	MMT	MMCW	IMMC
Total severity score	116	28	62
Number of false alarms	21	9	53
Total evaluation time	4.4 person hours	2.4 person hours	3 person hours
Number of evaluators	4	2	4
Cost-benefit ratio value	**21.59**	**7.92**	**3.00**

The figures in the table show that the Multimedia Taxonomy appears to be the most cost-effective evaluation method of the three. This indicates that using this EEM the highest number of critical problems have been identified with the least number of false alarms in the time taken by the experts to evaluate the Maths application. The Multimedia Cognitive Walkthrough came second, with a considerably lower value of the cost-benefit ratio. This was mainly due to the smaller number of significant problems identified by the evaluators. The cognitive walkthrough also took on average the longest time. Finally, the Interactive Multimedia Checklist proved to be the least cost-effective EEM than the other two techniques, This suggests that although the evaluators took the shortest time on average to identify a considerable amount of significant usability problems, they also incurred a considerable number of false alarms.

4.9.8 EEM Learnability

In terms of familiarisation time, the IMMC took the shortest time to learn, and it was found that this EEM does not require initial training to be used, as two of the experts did not read the paper or the checklist before conducting the evaluations. The other two evaluators took on average 45 minutes to read the paper. The MMCW requires more extensive training than the IMMC, as it took the evaluators 50 minutes on average to familiarise with the method. Finally, the MMT took the longest time on average - 55 minutes to familiarise with.

In the post-evaluation questionnaires and interviews the evaluators were asked how difficult they found the method to grasp and how complex different aspects of the EEMs were to them. Two of the evaluators who used the MMT thought they were difficult to grasp, and the other two thought they were relatively easy to comprehend. One of the evaluators who applied the MMCW thought the concepts were relatively easy to grasp however the other evaluator had difficulties understanding the cognitive model included

in the EEM. Finally, all the evaluators found it very easy to grasp the concepts of the IMMC.

Therefore, the IMMC took the shortest time to learn and was the easiest to grasp. The MMT was also relatively easy to learn, whereas the MMCW not only took the longest to learn was quite complex and hard to grasp.

4.9.9 Ease of Use

When asked during the post-evaluation interviews, the evaluators who applied the Multimedia Cognitive Walkthrough stated that it was not very easy to use. Due to the interactive nature of the application they found it hard to draw attentional graphs (part of the output expected from the MMCW), as the EEM assumes linear presentation of information and control over the time and duration for which dynamic media are displayed. Because no attentional graphs were drawn, this made it difficult for the evaluators to follow the first two steps of evaluation suggested by the method. Thus, the experts had to devise their own way of using the guidelines. The third step – evaluation of the media selection – was also not easy to apply as the evaluators felt that the information types specified by the authors do not adequately reflect the Maths content.

All but one of the evaluators who used the IMMC found it easy to apply. The evaluator who experienced difficulties had some problems deciding which questions were relevant to the application and which were not. Although explicit guidance as to how to use the checklist questionnaire was not provided, all reviewers were happy to adopt their own approach to evaluation, and did not feel that any specific instructions were required.

Finally, two of the specialists using the Multimedia Taxonomy found it easy to use, however they did not try to create evaluation protocols. The other two evaluators, who attempted initially to use the Taxonomy as suggested by its authors, found it very hard because insufficient guidance was provided about how to create evaluation protocols and how to use them for the inspection of the user interface. All evaluators also had difficulties understanding certain parts of the Taxonomy, particularly the relationships between its dimensions. Also no information was provided about how to determine which cells of the Taxonomy are relevant to the application under consideration,

however only one evaluator said that they would prefer to have selection guidance given to them.

4.9.10 Expert Specialisation

This criterion is relevant mainly to the IMMC and the MMT, where specialists from different professional backgrounds were involved. Neither of these methods provides any explicit specialisation of evaluation criteria depending on the expertise of the evaluators to be involved. In the case of the IMMC, all experts are asked to use the same list of questions, and they have to decide which ones they can answer. In the study, both types of expert however experienced difficulties understanding certain specialist terms used in the questions. For example, the SME had problems understanding the meaning of some HCI-related terms, such as end-user, mode of interaction, visual and spatial access to information. He also found the wording of other questions ambiguous. Furthermore, the majority of the experts seem to have interpreted the various learning styles (e.g. focused, browsing and shallow) differently, as precise definitions in the context of CAL applications were not provided. Such misunderstandings and differences in interpreting key terms could have had an effect on the reliability of the evaluations produced. Another consequence was that the evaluators using the IMMC were supposed to select only the questions which they felt they were able to answer. However, sometimes the SMEs attempted to answer questions regarding the mode or style of interaction, and the quality of the user interface, which sometimes contradicted the answers given by the MMDs.

The MMT does not provide explicit specialisation in terms of how evaluators from different backgrounds should use the Taxonomy. However, as the context dimension is divided into six distinct categories: audience, discipline, interactivity, quality, usefulness and aesthetics, the SMEs mainly provided comments regarding the content and audience implications of the design features. On the other hand, the MMDs assessed the interactivity, quality and aesthetics, and all of them considered the target learners. One of the SMEs had some software engineering experience, thus he provided some comments in other areas as well. On the other hand, one of the MMDs had knowledge about the Maths content and he made some comments on that too.

These results show that the above two methods, in their present form, cannot be used easily by evaluators from different professional backgrounds. If not improved, developers and evaluators may be reluctant to adopt such methods in their usability inspection practices.

4.9.11 Quality of Documentation

In the post-evaluation interviews and questionnaires, the evaluators were asked to comment on the adequacy and sufficiency of the documentation given to them. In particular, they were queried on how understandable, easy to follow and sufficient the documentation was, and whether specific steps for analysing user interfaces were provided and how adequate they were for the evaluation of the Maths application. Before the findings are presented, it needs to be acknowledged that all documentation given to the reviewers consisted of academic journal or conference publications made by the method authors, rather than practical instructions about the methods. Although these documents are not explicitly intended for use in evaluation, they were used as none of the method authors could provide any further documentation.

The two evaluators using the MMCW found the method's principles and guidelines clearly defined. However, one of them thought that the documentation was insufficient, and both of them found the instructions on how to use the cognitive walkthrough were not adequate or suitable for the evaluation of the Maths application. In particular, the evaluators found that no sufficient guidelines were provided as to how to implement the three steps of evaluation, hence the two experts agreed that appropriate instructions need to be provided to ensure more effective evaluation using the method.

The evaluators using the IMMC found the method well documented. In fact two of them who did not read the paper provided could apply the checklist with the same speed and ease as the other two who had read the paper. Although explicit guidance as to how to use the checklist questionnaire was not provided, all reviewers were happy to adopt an approach to evaluation they considered appropriate, and did not feel that any specific instructions were required for more effective evaluation. From the long list of questions, 90 in total, the reviewers had to answer only the ones relevant to the application. As no guidance was provided, they did that 'intuitively' as one of them expressed it. All four

evaluators agreed that some guidance about selecting the relevant sections and questions would have been helpful.

Finally, the evaluators found the descriptions of two of the dimensions of the MMT clearly defined, whereas they found the media expression dimension not sufficiently defined. Furthermore, the relationships between some of the dimensions were also not sufficiently explained, particularly between the context and the other two categories. The evaluators also found that the instructions on how to create evaluation protocols were insufficient, thus none of them actually generated such protocols. Also, no information was provided about how to determine which cells of the Taxonomy are relevant to the application under consideration, however only one reviewer said that they would prefer to have selection guidance provided.

4.9.12 Summary of EEM's Effectiveness

A summary of the effectiveness of the three EEMs according to each effectiveness criterion is given in Table 4.18. All ratings given in the table are comparative between the results of the three methods.

Table 4.18: A Summary of the Effectiveness of the three EEMs

EEM / Criterion	MMT	MMCW	IMMC
1. Validity	Reasonable 63% valid problems 37% false alarms	Reasonable 53% valid problems 47% false alarms	Poor 24% valid problems 76% false alarms
2. Thoroughness	▪ Comparative Broad focus ▪ Actual User problem prediction rate: 34.38%	▪ Comparative Narrow focus ▪ Actual User problem prediction rate: 14.06%	▪ Comparative Broad focus ▪ Actual User problem prediction rate: 14.06%
3. Reliability	Lowest	Low	Medium
4. Diagnostic Power	Low	Low	Lowest
5. Tailorability	Indirect to none	Direct	Indirect to none
6. Theoretical Validity	Practice applied & improved	Empirical test validated	Practice applied only

Table 4.18: A summary of the Effectiveness of the three EEMs (continued)

EEM / Criterion	MMT	MMCW	IMMC
7. Cost Effectiveness	Best	Poor	Worst
8. Learnability	Difficult to learn	Difficult to learn	Easy to learn
9. Ease of Use	Reasonably difficult to apply	Difficult to apply	Very easy to apply
10. Expert Specialisation	Some/basic	---	None/implicit
11. Quality of Documentation	Insufficiently defined. Insufficient instructions to use.	Clear guidelines. Insufficient instructions to use.	Sufficiently documented.

4.10 Need for Enhancing the Effectiveness of EEMs for IMM

The results of the study revealed that all three usability evaluation methods showed poor validity and thoroughness. This was characterised by the fact that the experts emphasised aspects which are not so pertinent to the usability of the IMM application, yet missed out important learner problems, including the majority of the most critical ones. Furthermore, in terms of predicting learner comprehension problems, the experts performed quite unsatisfactorily, as the experts missed out two thirds of these problems. This may be due to the fact that the complexity of the learner knowledge construction processes and the context of learning were not adequately addressed by any of the EEMs. On the other hand, a significant number of cosmetic and minor design faults were identified, which did not concern any of the users and did not have an impact on their learning performance. These results should cause concern in every development team wishing to use such methods, especially those who are unable to perform user tests, as the findings indicate that valuable resources could be employed in redesigning software based on such expert predictions without making any improvement to the usability of the IMM applications developed.

The empirical findings presented in this chapter highlight a dire need for improving the effectiveness of existing expert evaluation techniques, particularly their ability to predict valid learner problems. In order to improve their predictive power, deep

understanding of the reasons for lack of success in predicting critical learner problems is needed and measures for improvement should be developed and implemented.

Chapter 5 will present an approach aiming to improve key aspects of the existing EEMs for IMM, which is based on both the results from the empirical study presented in this chapter and theoretical work of how students build mental models of multimedia representations. Chapter 6 will then discuss the results of an empirical validation of an EEM developed based on the propositions made in Chapter 5.

Chapter 5:

Empirically and Theoretically Based Approach to the Improvement of the Effectiveness of Expert Evaluation Methods for IMM

Chapter 5

Empirically and Theoretically Based Approach to the Improvement of the Effectiveness of Expert Evaluation Methods for IMM

5.1 Introduction

The results from the empirical study, described in Chapter 4, revealed the limited effectiveness of all three existing EEMs for IMM studied, particularly in relation to the *validity* and *thoroughness* of the usability predictions made. These limitations indicate that there are a number of aspects which need to be addressed in existing EEMs for IMM in order to achieve highly effective evaluation performance, which produces the results desired by usability practitioners. To improve the effectiveness of expert evaluations, detailed analysis of the characteristics of existing EEMs that limit their performance is required, and actions need to be taken to eradicate them.

The investigation conducted in the second part of this thesis primarily concentrates on improving the *validity* of EEMs for IMM, expressed in terms of increasing the number of valid problems predicted and decreasing the number of false alarms identified during expert evaluations of IMM. The emphasis is on increasing the number of valid comprehension problems predicted by experts. Increasing the number of valid problems predicted also has an impact on the *thoroughness* of expert evaluations. Therefore, the work presented in this and the following chapters also aims to improve the *thoroughness* of EEMs for IMM, particularly where the prediction of comprehension difficulties are concerned.

In order to improve the effectiveness of EEMs for IMM, four hypotheses were formulated to fulfil objective 3 of this thesis:

Objective 3: To formulate hypotheses for enhancing the effectiveness of EEMs for IMM based on the empirical findings produced in Objective 2 and on theories of learning with IMM.

A hypothetico-deductive approach to EEM improvement was adopted. As defined by Campbell and Stanley (1966), the hypothetico-deductive paradigm involves deducing research hypotheses from theory or previous research and testing the hypotheses by means of an experiment, quasi-experiment, or ex post facto design. The major advantage of this paradigm is to confirm or explore causal relationships between variables.

The hypothetico-deductive approach, which contains three steps, is depicted in Figure 5.1.

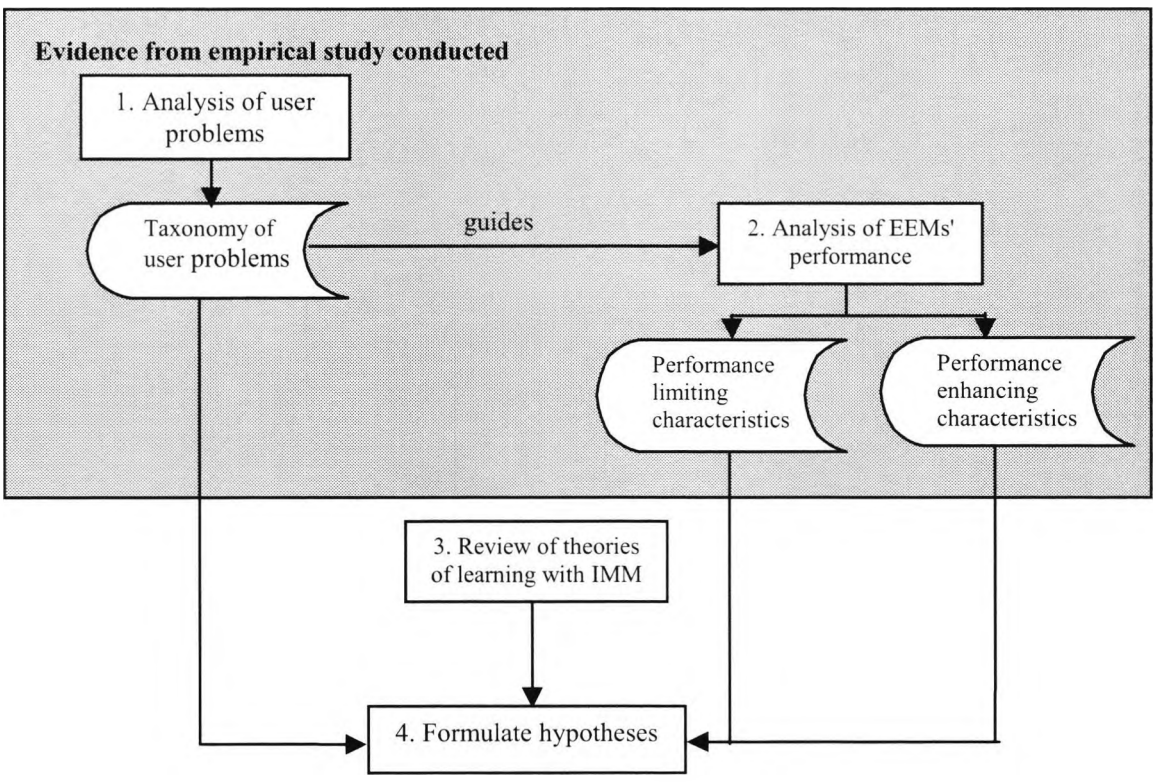


Figure 5.1: Hypothetico-deductive Approach to Improving EEMs' Effectiveness

The first step involves an in-depth analysis of the nature of the usability and comprehension problems the users encountered, describing their structure as well. The analysis conducted is presented in Section 5.2. The second step aims to provide a profound understanding of the characteristics of the existing EEMs, which limited their

validity and thoroughness . Understanding of such characteristics will provide an insight into what aspects of the EEMs need to be addressed in order to improve both criteria. The major performance limiting characteristics are discussed in detail in Section 5.3. On the other hand, there were also aspects of the expert evaluation process which promoted the effective prediction of valid user problems. Understanding what these aspects are can provide an insight into potential enhancements which could be introduced in new EEMs to ensure their effectiveness. Thus, Section 5.4 reveals such effectiveness promoting factors. Finally, a survey of existing theories and empirical studies regarding the cognitive processes of learning with IMM was performed. It can also inform the generation of tentative conjectures as to how the prediction rate of valid user problems can be improved. The review is presented in Section 5.5.

Based on the above analysis, a set of four hypotheses were formulated, stipulating conjectures regarding how to enhance EEMs' effectiveness. The hypotheses are described in Section 5.6. Finally, Section 5.7 presents the methodological approach for testing each hypothesis.

5.2 Analysis of User Problems while Learning with IMM Software

In order to generate hypotheses that will improve the prediction rate of valid user problems, a deeper understanding of the nature and components of user problems experienced while using IMM is required. The primary aim of this analysis is to define requirements for the specification of user problems during expert evaluations of IMM. A secondary purpose of this analysis is to provide a benchmark for comparing predicted problems generated using such EEMs to actual user problems, therefore assisting in the assessment of the validity and thoroughness of EEMs for IMM, which was partially done in the previous chapter.

The user test data, including the video footage of the user interactions and the comprehension test results, were further analysed to define the nature and components of the user difficulties.

In Chapter 4 the user problems were divided into two categories: *usability* and *comprehension*, where the first type occurred directly during the user interaction with

the IMM software, and the second were revealed in the post-exposure knowledge tests. For the purpose of being more specific, the first category of problems will be called *user interaction problems*. Lavery and Cockton (1997) also distinguish between two types of usability problems: *dialogue failures* and *knowledge mismatches.* The authors define *a dialogue failure* as "an observable event that indicates that interaction is not proceeding as envisioned by the designer", whereas a *knowledge mismatch* describes a difficulty "where the user does not have the correct understanding of a particular aspect of the system or task". The *user interaction problem* category introduced in this research adopts characteristics of Lavery and Cockton's *dialogue failures*. However, in the context of this research, the *comprehension problem* category is more specific as it is used to describe difficulties directly related to learners' understanding of the subject-matter content, which differs from Lavery and Cockton's definition of *knowledge mismatches.*

To be able to characterise both types of user problem, each of them is defined in the sections below. To be able to represent each type uniformly common representation formats are also proposed.

5.2.1 Analysis of User Interaction Problems

In this research, a *user interaction problem* is defined as a difficulty occurring during the user dialogue with the user interface, which typically has an immediate effect on their behaviour or performance, which hinders or prevents them from successfully achieving a part of their task.

The usability problem format outlined in Lavery, Cockton and Atkinson (1997), given in Figure 5.2, was used as a basis for specifying the structure of the interaction problems users encountered and for representing them.

As can be seen from Figure 5.2, Lavery, Cockton and Atkinson (1997) distinguish three main components of a usability problem: *a cause,* a possible *breakdown* in the user interaction, and *an outcome.* The authors specify that a *cause* is typically a design fault, which may lead to a breakdown in the user's interaction. Sometimes a design fault may not produce a breakdown, instead it can reduce the usability of the application. A *breakdown* occurs when the user takes an inappropriate step in their interaction. Finally,

an *outcome* can be expressed as a certain behaviour and/or performance. *Behaviour outcomes* address the unfolding of the interaction following a breakdown, whereas *performance outcome* is the result of the user behaviour in relation to the achievement of their goal.

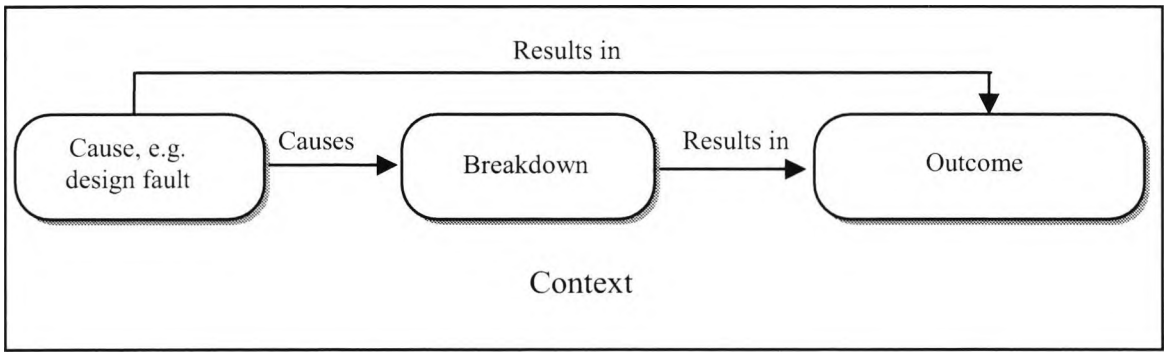


Figure 5.2: Components of a User Interaction Problem
(from Lavery, Cockton and Atkinson, 1997)

To represent user interaction problems the above four components were adopted and defined in a way that reflects specific aspects of the user's interaction with IMM applications.

User Interaction Problem Report Format

Context: Describes any aspects of the context of use which may be responsible for the interaction problem.

Cause: Indicates the design aspect or demand on the user likely to be the cause of the user interaction problem.

Cognitive Breakdown: Is the immediate response of the users at a cognitive level, which prevents them from achieving their goal. In the context of user interaction with IMM the cognitive response may include distracting user's attention from important information, overloading their attention, or the user not being able to identify what action could achieve the goal. A breakdown is the reason for the subsequent undesired user behaviour or performance. It is a mental process triggered by the user interaction, which is manifested in and is responsible for certain external user behaviour and performance, which is

typically inappropriate for achieving their goal. Therefore, there is a close reciprocal relationship between user's cognitive breakdown and their behaviour and performance.

Behavioural Outcome: Is the unfolding of the interaction following a breakdown, i.e. it describes the user's observable actions which result from a cognitive breakdown. For example, after the user forms an incorrect goal, s/he does not perform the correct action to achieve the desired goal.

Performance Outcome: Is the effect on the user's work and the achievement of their tasks and learning goals, resulting from the interaction problem. It describes how the user performance in the task will be affected as a result of experiencing the breakdown and performing incorrect actions. Example performance outcomes include: the user may spend a long time on the task, s/he may fail to achieve their task, or if the problem is minor the user will be able to continue with the task and successfully achieve it. This form of outcome is related to the severity of the breakdown, i.e. can the user recover from it or will their work suffer.

The 51 user interaction problems experienced by the users while using the IMM application were further analysed to describe their causes, cognitive breakdown and resulting outcomes. A full description of all user interaction problems is given in Appendix 5.1. Below is an example of such a problem:

User Interaction Problem # 2:

Context: user started demo and the graph of e^x was displayed, after which they were prompted to "click again for the graph of 2^x".

Cause: the difference between the three graphs in the demo and why 10^x is steeper than the other two graphs is not explained.

Cognitive Breakdown: student has difficulties understanding why the graphs were different.

Behavioural outcome: the student redoes the demo as a result of that and reads the text again.

Performance outcome: Users spent more time on task. Two of the students had problems recognising which graph is which on the comprehension test.

To gain further insight into the nature of each component of user interaction problems, a detailed analysis of the usability problems experienced by the users while interacting with the IMM software was performed. The following sub-sections present categorisations of the three main components of user interaction problems: cognitive breakdowns, behaviour outcomes and performance outcomes. Furthermore, it is also essential to establish what parts of the user interaction problems were successfully predicted by the experts and which were not. This is also revealed in the sections below.

5.2.1.1 Categories of Cognitive Breakdowns

Based on the cause-breakdown-outcome analysis conducted on the usability problems eight categories of breakdown can be distinguished. The first five are particularly relevant to knowledge-based tasks and their achievement using IMM user interfaces. The last three categories were identified by Lavery, Cockton and Atkinson (1997), and although they are also appropriate for procedure-based tasks, the users of the IMM application also encountered them.

1. **Attentional breakdowns**: Can either represent that the user's attention is not drawn to an important part of the presentation or that the user's attention is overloaded. Such occurred in six user interaction problems.

2. **Perceptual breakdowns**: Occur when the user is unable to perceive a part of the display which is hidden from them, or when the display is unpleasant or uncomfortable to view particularly for a long time. Such were found in four user interaction problems.

3. **Information integration breakdowns:** The user cannot establish a link between information presented in two media resources. These occurred in five user interaction problems.

4. **Content comprehension breakdowns:** The user has no or has formed partial or incorrect understanding of the subject matter content. Some comprehension breakdowns may be a result from the previous three types of breakdown, but also may be due to learner's insufficient prior knowledge, or

the material being too complex and not sufficiently explained. Such occurred in eleven user interaction problems.

5. **Affective breakdowns:** Occurs when the interaction problem can generate an emotional response in the user. For example, the user may get frustrated or confused as a result of performing an incorrect action, or may not find learning tasks interesting or challenging. Affective outcomes were expressed in seven cases.

6. **Goal formation breakdowns:** The user does not understand the goal of the presentation or has difficulties establishing the correct goal. Such were found in three user interaction problems.

7. **Goal execution breakdowns:** The user has formed the right goal but is not sure what action will achieve that goal, or believes that a wrong action will achieve the goal. Such occurred in a considerable number of user interaction problems, namely in thirteen.

8. **Action feedback breakdowns:** The user does not recognise the result of or the feedback from the action as being right for the goal, or thinks the feedback is the wrong one. Such occurred in four user interaction problems.

The different types of breakdowns defined above are not mutually exclusive and can occur in combination or sequence of each other.

A limited number of the breakdowns were predicted by the experts using the three EEMs. Three attentional breakdowns were predicted by the two multimedia designers (MMDs) using the MMCW. The cause of one perceptual breakdown was predicted by the subject-matter experts (SMEs) who used the MMT. The information integration and content comprehension breakdowns are related and they proved to be the most difficult to predict, as none of these breakdowns were identified by the experts. These breakdowns are complex in nature, as their identification requires deep analysis of learners' conceptual learning processes. However, identifying such breakdowns is key for predicting comprehension performance problems, as will be revealed later in this

chapter. Finally, some of the experts, particularly the instructional specialists, predicted five affective breakdowns.

5.2.1.2 Categories of Behavioural Outcomes

The following five categories of behavioural outcomes were identified from the user interaction problem set. The first and the last were also identified by Lavery, Cockton and Atkinson (1997), whereas the remaining three were defined based on the problems experienced by the users while learning with the Maths application.

1. User performs a wrong action. Observed in three user interaction problems.

2. User repeatedly performs an action, which could either be right or wrong. This may include reading or viewing a part of the MMUI in an attempt to gain further understanding of the material. Observed in five cases.

3. User skips an important part of the presentation. Observed in three cases.

4. User performs the right action, but uses wrong parameters. Observed in one case.

5. User abandons their task. Observed in one case.

Sometimes more than one of these actions can occur, e.g. user can reread a piece of text and then perform a wrong action.

One of the experts who applied the MMT predicted one valid 'wrong action' and two valid 'repeated actions' behavioural outcomes. In describing the behaviour following a cognitive breakdown, the evaluator however described his own actions rather than the actions the users are likely to take. Finally, the evaluator was not prompted by the MMT to specify these outcomes, instead he used his own judgement and initiative.

5.2.1.3 Categories of Performance Outcomes

The following five categories of performance outcomes were found amongst the user interaction problems:

1. User cannot achieve her/his task without external help. This type is concerned with failures to achieve user interaction tasks, whereas failures to achieve comprehension tasks are dealt with a separate category. Occurred in four cases.

2. User interaction fails, signifying that the user has failed to achieve the goal of their interaction. Occurred in one case.

3. User spends more time performing a task. As a result of that, depending on the severity of the cognitive breakdown, the user may or may not recover and complete the task successfully. Occurred in five cases.

4. Missed interaction, signifying that the user does not attempt to perform an action or a series of actions necessary for achieving their goal. Occurred in four cases.

5. User fails to comprehend a concept. Some problems that users experience with their interaction, particularly those at knowledge level, can have an effect on their learning performance. Therefore, this type of performance outcome is specifically concerned with situations in which users are not adequately supported in achieving their learning goals. Such outcomes occurred in six cases.

   Comprehension outcomes are very important indicators of problems connected with achieving the learning goals set in the application, however they are only manifested when the users are given an opportunity to test and apply their understanding. Furthermore, the outcomes of such difficulties may not be observable immediately after the user interaction, i.e. the user may continue with the task without having formed an adequate understanding of the material. Instead, comprehension outcomes are best manifested during comprehension tests or interactions which require users to demonstrate understanding of the material. Long-term memory

comprehension outcomes are discussed in greater detail in the following section, together with the success rate of their prediction by the experts.

Two performance outcomes of type 3 were predicted by the same evaluator who predicted the behavioural outcomes. Furthermore, the two SMEs who used the MMT predicted three comprehension outcomes. There is no evidence suggesting that the MMT prompted the prediction or the specification of either of these outcomes.

Table 5.1 presents a summary of the component categories and their prediction rate by the three EEMs studied. The findings presented in Table 5.1 show the low prediction rate of user problem components by the expert evaluations. In particular, they highlight that only one of the EEMs - the MMCW, provides any support in predicting cognitive aspects of the implications that design faults may have for target users.

Table 5.1: User Interaction Problem Components and their Prediction Rate by experts

User Interaction Problems Component Type	Number of Problems Experienced	Number of Problems Predicted	Method by Which They Were Predicted
COGNITIVE BREAKDOWNS			
1. Attention breakdowns	6	3	MMCW
2. Perceptual breakdowns	4	1	Own judgement
3. Information integration breakdowns	5	-	-
4. Content comprehension breakdowns	11	-	-
5. Affective breakdowns	7	5	Own judgement
6. Goal formation breakdowns	3	-	-
7. Goal execution breakdowns	13	-	-
8. Action feedback breakdowns	4	-	-
BEHAVIOURAL OUTCOMES			
1. User performs a wrong action	3	1	Own judgement
2. User repeats an action	5	2	Own judgement
3. User skips important part of the MMUI	3	-	-
4. User uses wrong parameters	1	-	-
5. User abandons task	1	-	-
PERFORMANCE OUTCOMES			
1. User cannot achieve their task	4	-	-
2. User interaction fails	1	-	-
3. User spends more time on a task	5	2	Own judgement
4. Missed interaction	4	-	-
5. User fails to comprehend a concept	6	3	Own judgement

5.2.2 Analysis of the Comprehension Problems

As revealed in Chapter 4, the experts were unable to predict nearly two thirds of the user comprehension difficulties. These difficulties were classed as major problems as they represent situations in which the learners failed to achieve the learning goals set in the IMM application. A deep understanding of the nature and structure of these problems is needed to inform the effort to increase the ability of experts to predict such problems during expert evaluations. In the paragraphs below, comprehension problems are firstly defined, after which a format for representing such problems is proposed.

5.2.2.1 Definition of Comprehension Problems

Fisher and Lipson (1986) define a learning error as:

> "... an observable event or performance which, in a way judged to be significant, differs from an expected, ideal ("correct") model of performance."

The authors add that learning errors encompass a broad range of performances exhibited by individuals while learning. An error cannot be perceived in the absence of an expectation or a goal which defines the "correct" response. Error recognition, therefore, requires judgement, often on the part of a teacher, who judges whether the tasks have been performed correctly and whether errors have been made.

Fisher and Lipson (1986) further emphasise that in defining learning errors it is important to distinguish between observable behaviours/performance and underlying causes. They state that there is a close reciprocal relationship between performance and mental representations in that performance helps reveal the existing mental representations, and understanding the current representation is important for eliminating errors and correcting performance. In this sense, Fisher and Lipson (1986) view misconceptions as causes of "incorrect" behaviour or performance, in the same way as cognitive breakdowns trigger incorrect interactions with a user interface and result in partial or complete failure of achieving user tasks. However, in the context of this research, it is also necessary to identify the causes of the misconceptions as well, particularly those due to user interface design features. Therefore, to avoid ambiguity, in this research a cause of a comprehension problem is considered to be the cause of the misconception, e.g. a design fault or a lack of prior knowledge.

In the context of using IMM software, it is defined that a *comprehension problem* represents a difficulty experienced by the users of such software, which occurs as a result of an aspect of the MMUI making a demand on the users' cognitive processes and results in hampering their knowledge construction processes. As a consequence, the users may fail to develop correct understanding of the subject matter, or may develop partial or wrong understanding of the material presented in the software. User comprehension problems differ from other usability problems in that the outcomes of such problems directly affect users' learning performance and the achievement of their learning goals. Furthermore, the outcomes of such difficulties may not be observable immediately after the users' interaction with the application, i.e. the users may continue with their task without having formed adequate understanding of the material, and this may or may not be manifested in their subsequent behaviour. Instead, the outcomes of such problems are best manifested in the users' learning performance following their interaction, for example during comprehension tests or other activities which require the learners to demonstrate understanding of the material.

5.2.2.2 Comprehension Problem Report Format

From the definition presented above and analysis of the comprehension problems experienced by the learners who used the Maths software, four components were identified to represent *comprehension problems* that occur while learning with IMM software. They are as follows: *Context, Cause, Misconception,* and *Comprehension Outcome,* and are presented in Figure 5.3.

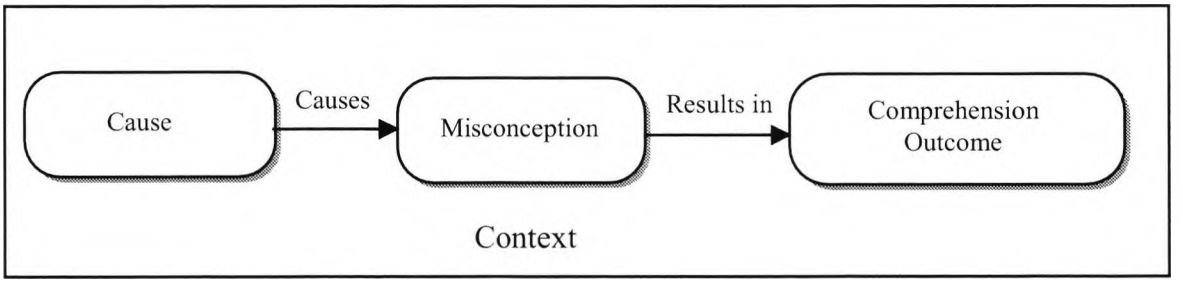


Figure 5.3: Components of a Comprehension Problem

Comprehension Problems Report Format:

Context: Describes any contextual factors that may have been responsible for the comprehension problem, for example learner's prior knowledge of the material.

Cause: The *cause* is whatever is believed to be responsible for the comprehension difficulty. Failures in learning can be attributed to a variety of factors, which can be both internal and external to the learner (Johnson-Laird, 1983; Fisher and Lipson, 1986). Possible causes include a mismatch between learner's prior knowledge or existing mental model of the material and the material represented in the IMM, insufficient or ineffective presentation of the subject-matter, and hindrances in the learning environment. Often there is more than one factor responsible for failure in achieving learning goals. Because of the multiplicity of factors, establishing causality of learning problems is difficult. In the case of evaluating the pedagogical effectiveness of IMM, however, of primary interest are causes which originate from the MMUI, such as design faults, demands on the users' learning processes or interaction difficulties hampering users' concept formation processes.

Misconception: Describes the nature of the comprehension difficulty, which typically consists of an incorrect or an incomplete mental model of the material. In particular, it indicates that the learner has acquired partial, incorrect, or no understanding of the material. Defining misconceptions is complex, as it requires not only substantial knowledge of the subject matter but also understanding of learners' conceptual models. Therefore, subject matter experts are required to specify such misconceptions.

Comprehension Outcome: Represents the observable behaviour or performance exhibited by learners as a result of forming a misconception. It can consist of a sequence of wrong actions, no actions when such are expected, or actions demonstrating incorrect performance. As highlighted above, the success or failure of comprehension outcomes needs to be measured in relation to the learning goals set in the educational application, by assessing how successfully they have been achieved.

Categories of Comprehension Outcomes

A part of Bloom et al.'s (1956) taxonomy of educational objectives can be used to classify comprehension outcomes. The authors define *comprehension* as the situation in which learners are confronted with a communication of some kind, and they are expected to know what is being communicated and to be able to make some use of the material or ideas contained in it. *Comprehension* includes those objectives, behaviours, or responses which represent an understanding of the message contained in the communication. The authors specify three types of comprehension behaviour: *translation*, *interpretation*, and *extrapolation*. *Translation* means that a learner is capable of putting a communication into other language, terms, or other form of communication. It usually involves the giving of meaning to the various parts of a communication, taken in isolation. *Interpretation* involves dealing with a communication as a configuration of concepts whose comprehension may require a reordering of concepts into a new configuration in the mind of the individual. This also involves considering the relative importance of the concepts, their interrelationships, and their relevance to generalisations implied in the original communication. Finally, *extrapolation* includes the making of estimates or predictions based on understanding of the trends or conditions described in the communication. It may also involve the making of inferences with respect to implications, consequences or effects, which are in accordance with the principles described in the communication. Therefore, each comprehension outcome can be described precisely in terms of the type of comprehension difficulty the learners experienced.

Each of the 20 knowledge propositions, which the exponential graphs material was divided into (see Table 4.10), was tested in the post-exposure knowledge test for one or more of these types of comprehension depending on the nature of the concepts. For example, to test students comprehension of propositions 3 and 4 they were examined as to whether they could *translate* exponential equations into graphs and vice versa. Whether the students are able to *extrapolate* exponential graphs from their equations was also tested. (See comprehension test questions in Appendix 4.10).

Detailed definitions of all comprehension problems are presented in Appendix 5.2. The causes of the comprehension problems were inferred based on design features

representing the concepts involved. Sometimes they also refer to user interaction problems, which could have potentially hindered students' understanding of the Maths concepts. As some causes were specified with a degree of uncertainty, such assumed causes are presented in brackets. As a subject matter expert could not be recruited to define the likely misconceptions that could have occurred, the researcher defined only some of them, which are aimed to illustrate the nature of possible misconceptions that may have occurred rather than to define such with accuracy. Finally, the comprehension outcomes describe the type of comprehension behaviour or performance showed by the students during the comprehension tests.

Experts using the three EEMs were able to predict primarily design features that could potentially cause comprehension problems. However, they neither specified what misconceptions learners were likely to form nor how the design of the IMM application could influence learners' behaviour or performance. As a result, only four out of the thirteen comprehension problems (or 31%) were predicted using the EEMs, four of which were predicted using the MMT and one using the MMCW.

An example comprehension problem definition is given below. As the misconceptions are assumed without having sufficient evidence or understanding of the actual ones, they are given in brackets.

Comprehension Problem # 54

Context: During the tests the students are asked to recognise the type of scaling from the equations, however it depends on how many tests the students did for scaling.

Cause: Graph lines are not labelled adequately. Connection between textual explanation of both types of scaling and the graphical representations during tests is not direct or obvious. The differences between the two types of scaling are not explicitly explained.

Misconception: (The students could not grasp how exponential equations are translated into scaled equations. The students also may not have paid enough attention to the differences between the two types of scaling.)

Outcome: All 4 students had difficulties recognising the type of horizontal scaling from its equation. One of them recognised that it was scaling, but swapped the two types - horizontal and vertical.

5.3 Effectiveness Limiting Characteristics of the EEMs for IMM

The data presented in the previous section revealed that the usability evaluators were unable to predict most types of breakdowns and behavioural outcomes, all types of performance outcomes, particularly those to do with comprehension of the subject matter. On the other hand, the experts identified a number of false alarms, which did not have any effect on the users' interaction or performance with the IMM application. As highlighted earlier, it is essential to identify and analyse the characteristics pertinent to the EEMs which could have limited their validity, and which are responsible for the identification of false alarms. Such an analysis can then be used to propose improvements to the methods' effectiveness.

The following sections present four characteristics of the EEMs that can explain the reasons for the prediction of a significant number of false alarms, for under-specifying usability problems, and for failing to discover certain valid user problems. The characteristics were identified from a further analysis of the EEMs' performance data presented in Chapter 4.

5.3.1 None of the EEMs adequately supports evaluators in inferring the impact that design faults may have on the intended users, including their behaviour and learning performance.

As revealed in Sections 5.2 and 4.5.5 (Diagnostic Power of the EEMs) the majority of the problems predicted by the experts do not specify the effects that design faults may have on target users. This limitation is considered to be the main reason why a significant number of the predicted problems were under-specified, a considerable proportion of which were found to be false alarms, as discussed in Section 4.7. The main reason for this is that all the EEMs focus primarily on aspects of the user interface design and consider its quality in isolation from the users who will be using the application and the wider context in which it will be used. This tendency was revealed in the lack of adequate tailorability to context, discussed in the previous chapter. This indicates that the evaluators identified possible design faults without attempting to confirm whether they would actually cause difficulties to users in the course of achieving their learning tasks. Consequently, a number of these design faults were not confirmed by the user tests, and therefore were considered to be false alarms.

To be able to increase the number of valid usability problems, and to reduce the number of false alarms predicted by experts, a detailed analysis of the focus of each EEM was conducted. Further analysis of whether the EEM explicitly encourages evaluators to infer the effects on the intended users or not, the findings of which are presented below, was also performed.

- **IMMC**

  The largest proportion of problems that do not identify the likely effect of design faults on users were identified while applying the IMMC. In particular, 76% of the total number of problems predicted using this method do not specify such effects. A possible explanation is that the majority of the checklist questions require evaluators to check whether the design follows certain principles of good design and includes certain features, in isolation of who will use the software and how it will be used. Example of such questions are: 'is the feedback offered in a consistent way?' and 'are facilities such as accelerator keys made available to the users?' Thus, the quality of these features is considered independently, without considering whether they support the users and the achievement of their tasks. Therefore, most of the questions ask whether certain design features are present or not, assuming that if they are present then the user interaction will be effective and the content understandable, without taking into account individual differences and characteristics of the learners or the particular context of learning.

- **MMT**

  It was found that 68% of the problems predicted using the MMT do not reveal what implications the design faults may have for the intended users. As no explicit evaluation criteria are provided in the taxonomy, the focus of the predictions depended solely on the evaluators' expertise, although they were guided by the structure and categories of the three dimensions. It was found that the more experienced evaluators, especially the two subject matter experts, considered the learners and the likely effect of the content and its representation on them, whereas the less experienced evaluators did not.

- **MMCW**

  58%, or just over half, of the problems predicted using the MMCW do not point to the likely effect the design problems may have on the intended users. The emphasis of the attentional and contact points guidelines is on the design of individual media and how different types of media can be linked together in order to focus attention on or establish sufficient links between pieces of information. However, the implications of such design aspects for the user's cognitive load and cognitive processes of attention, perception and integration of visual and verbal inputs are not explicitly explained in the guidelines. The emphasis of the media selection guidelines is on what media resources to select in order to represent certain types of information and why certain media should be selected depending on the suitability of their symbolic systems to represent certain types of data. Again, the implications of the media choice for user's comprehension is not brought up in the guidelines. Such implications are only explained at the beginning when the model of comprehension is explained, and in the actual method these are implicit. Therefore, the evaluators using the method are not generally asked explicitly to consider cognitive implications, instead they are made implicit, i.e. by following the guidelines it is assumed that it will be ensured that the user will not be distracted and that important information will be brought up.

5.3.2 There are evaluation criteria which require evaluators to provide their subjective opinion about users' attitude, needs and personal learning goals, which often resulted in inaccurate predictions and false alarms.

Nineteen of the ninety questions in the IMMC ask the evaluators to judge how the design would affect the users in terms of their feeling of engagement, whether the tasks, the content and the style of interaction are appropriate to the users' needs, and whether the users can build their own plan of achievement and can make personally relevant decisions about aspects of the UI. An example of such a question is: "can the user identify with the goals and objectives and build their own personal plan for achievement?" The evaluators who applied the method tended to disagree on their answers to such questions. Furthermore, the evaluators' opinions also often contradicted the user responses. As a result, the majority of the issues identified using such questions were classed as false alarms. Most of the false predictions were made by the MMDs,

however the SMEs also identified two false alarms. A likely reason for the evaluators' misjudgements is that such questions tend to require a considerable amount of subjective judgement on the part of the evaluators. As they do not constitute typical users, the evaluators had difficulties predicting users' attitude and behaviour with accuracy. To be able to judge users' attitude towards certain aspects of the software, as well as whether their learning needs are met, evaluators need to be well acquainted with the target user group. The MMDs only had a short user profile on which to base their judgements. Therefore, they were providing their own opinions rather than predicting what the likely user reactions might be. Stereotyping of users and substituting experts' own experiences with the application were also observed by Hertzum and Jacobsen (1999). The authors observed that each of the four questions of the Cognitive Walkthrough method (Wharton et al., 1994) drives the evaluators to think of the user's behaviour in a certain situation. When the user description becomes too fuzzy or lacks sufficient detail to judge user's behaviour, the evaluator unintentionally substitutes the description with a particular user much like him/herself. Thus, evaluators tend to produce success stories if they imagine themselves having no troubles using the design feature in question, and they report problems when they imagine themselves having troubles in the particular situation. Although SMEs were more familiar with typical users as they have taught students, they also made wrong assumptions believing that learners' attention and concentration could not be maintained consistently, which was not the case with the students who took part in the study. However, the experimental nature of the evaluation could have influenced that. Based on the above, it could be concluded that accurate predictions of how design aspects may impact users' attitude and satisfaction of their needs require experts to be well acquainted with the intended user group.

A second drawback of such evaluation criteria is that they ask evaluators to imagine user response in general, rather than in the context of specific realistic learning situations and tasks. This adds to the fuzziness of the predictions, and due to the lack of detail in which predictions need to be made the evaluators need to approximate not only who the users are, but also what tasks they are likely to be performing and in what learning conditions. This provides more leeway for subjectivity and misjudgement. Therefore, questions of such nature also need to be in context rather than generic.

5.3.3 None of the EEMs encourages the consideration of relevant user characteristics, such as prior knowledge or cognitive capabilities.

Another important limitation of all three EEMs is that none of them suggests that the learners' prior domain knowledge, the complexity of the subject-matter content or the cognitive task requirements of the content should be considered. Furthermore, none of the EEMs correlates these aspects to how media resources could be used and designed to represent complex concepts in order to enable students to comprehend them more easily. This limitation could explain why two thirds of the comprehension and learning support problems that the users experienced were not predicted by the evaluators, including the subject matter experts. In particular, the experts rarely considered relevant user characteristics, such as *students' prior knowledge* of the material, while evaluating the usability of the Maths application. For example, most students requested more support on Scaling and Translation, especially Scaling, since they had no previous knowledge of these concepts. However, although the experts were advised to assume that the students had little or no prior knowledge of the subject matter, none of them could envisage where students with little prior knowledge may need further explanation of the material. Furthermore, none of the three evaluation methods explicitly asked the evaluators to consider students' prior knowledge. Kalyuga, Chandler and Sweller (1998) argue that whether instructional representations are intelligible often depends on the prior level of experience of the learner, and whether they have sufficient knowledge to understand the representations. Therefore, exploiting information about users' prior domain knowledge during usability evaluation of IMM could potentially improve inferences regarding how understandable the material is and improve predictions of potential problems that learners with low prior knowledge may experience.

Another very important factor which the experts did not consider was the *cognitive task requirements* of the multimedia representation of the Maths material. The cognitive task requirements represent demands placed on the users' cognitive processes by the MMUI, which users need to perform in order to process the information presented to them. This includes not only the subject matter content, but also user interface components, such as navigation and dialogue aids. One factor found to influence the comprehension of the material was the complexity of the Maths concepts and the way this complexity was represented in the IMM user interface. The greater the complexity of the material the greater are the cognitive task requirements on the students, and the more support they

will need to be given by the presentations in order to comprehend the complex concepts. Therefore, instructional multimedia presentations need to be designed paying particular attention to human working memory limitations (Kalyuga, Chandler and Sweller, 2000), and reducing cognitive load for the target learners (Rogers and Scaife, 1999). Kalyuga, Chandler and Sweller (2000) also clearly state that the most instructionally effective media presentation is the one that poses the lowest unnecessary cognitive load on the target learners.

In the case of the Maths material used in the evaluations the concepts of Reflection are the simplest because a single change of (+) or (-) signs is involved. The principles of Translation are slightly more complex as the movement of the graphs depends on whether a constant is a positive or a negative number. Finally, the principles of Scaling are the most complex of the three, since the movement of the graphs again depends on whether a constant is positive or negative, but there is also a difference in the vertical intercept. In spite of these differences, the material is represented in the same way for each type of transformation – using text and supporting graphs in the demo/tests. Thus, more complex concepts are not explained sufficiently and intricacies are not highlighted in the presentation. For example, in Scaling the movement of the graphs is represented graphically, however, as shown in Figure 5.4, no emphasis is placed on the position of the vertical intercept which is an important difference to the other two methods of transformation. Therefore, students with limited prior knowledge of the concept of transformation may have experienced difficulties linking the textual explanation and the graphical representation of the vertical intercept. Cheng (1999) emphasises that an effective external representation must help students to integrate information about a domain and should reveal the nature of the connection between laws of the domain; this has not been achieved by the media representation in Figure 5.4, as the circles and arrows indicate.

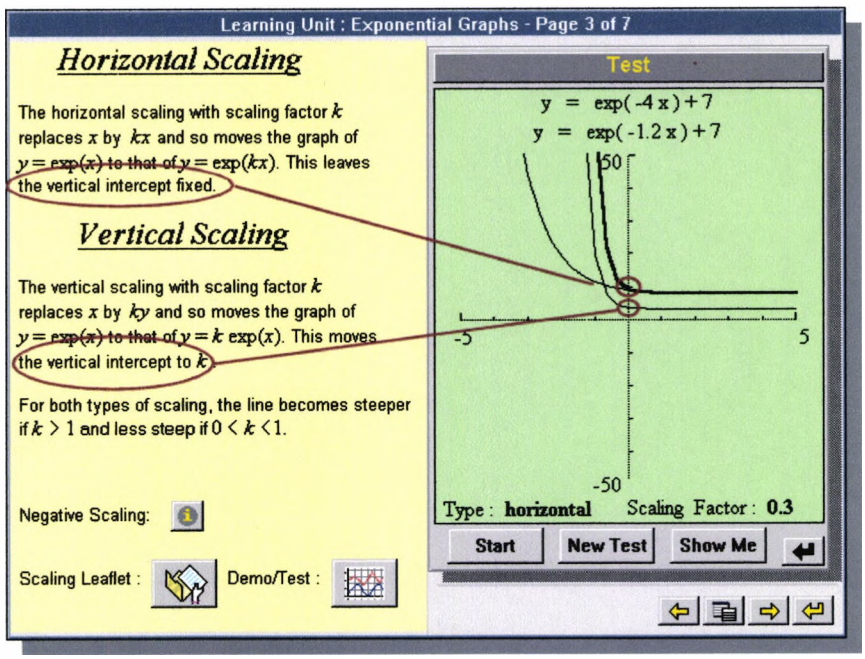


Figure 5.4: Limitations in representing Scaling (NAG ©)

The comprehension test results showed that all students grasped the concepts of Reflection, the majority of them got Translation right as well, however most of them experienced difficulties with understanding Scaling. These results suggest that the external media representation places demands on learners' cognitive capabilities and thus can either facilitate or hamper their understanding of the material presented. Effective usability evaluation of IMM needs to consider these demands and ensure that users cognitive load is minimised.

5.3.4 All the EEMs lack a uniform format for reporting usability problems.

Because none of the EEMs suggests how usability problems should be reported, inconsistencies were observed in the level of detail and the level of granularity of the different problem descriptions. A significant number of the predicted problems were under-specified, either not revealing the likely impact of design faults on the users or not specifying precisely the cause of a usability problem. From a usability engineering perspective, the implications of providing insufficient details in reporting usability problems are twofold. Firstly, under-specifying the impact of usability problems can often lead to unconfirmed design faults, which do not affect real use of the application, which were classified as false alarms. This jeopardises the cost-effectiveness of the development process, as valuable time and effort could be expended without improving

the effectiveness of the final product. Secondly, not describing the precise cause of usability problems can hamper the redesign process, due to uncertainty about the potential effectiveness of redesign solutions. Both implications can potentially undermine the credibility of and confidence in the usability evaluation process. From a research perspective, both the inconsistencies in the level of detail and the discrepancies in the level of granularity of the predicted and observed problems caused difficulties in matching the two problem sets.

Jeffries (1994), Mack and Montaniz (1994) and Lavery and Cockton (1997) all advocate that uniform problem reports should be introduced during expert usability evaluations. These researchers emphasise that usability problem reports not only ease problem analysis, but also clearly impact the quality of predictions. Lavery and Cockton (1997) also acknowledge that a structured representation firstly makes independent matches of predicted and observed problems more realistic, and therefore improves the credibility of validations of expert evaluations. Secondly, it encourages design analysts to produce more detailed and comprehensive predictions. Therefore, a uniform usability problem format should be introduced to ensure completeness and consistency of usability problem reports.

5.4 Effectiveness Promoting Factors

In order to be able to enhance the effectiveness of existing EEMs for IMM and build new more effective methods, we also need to understand what aspects of the expert evaluation process promote the effective prediction of valid user problems. Such understanding can then provide an insight into potential enhancements which could be introduced in existing EEMs and built into new EEMs to ensure their effectiveness.

A small proportion of the usability problems predicted by the experts did specify the likely impact that design faults may have on users. In particular, in the case of the MMCW, 32% of the predicted problems specify the likely impact, as did 27% of those predicted using the MMT, and just 3% of the ones identified using the IMMC. The main issue of interest is to identify the reasons why the impact was predicted in these cases in order to draw conclusions regarding how such predictions can be made more systematic during expert usability evaluations. In the analysis presented below, it is also discussed

whether the effects specified were influenced by an aspect of the EEM used or primarily by the evaluator's judgement.

In Section 5.2 it was revealed that some of the evaluators successfully predicted a few cognitive breakdowns, some behavioural outcomes and a limited number of potential comprehension difficulties. Below is the analysis of each of these successful predictions.

5.4.1 Prediction of Cognitive Breakdowns

The majority of the fully specified problems described possible breakdowns in the user interaction, i.e. immediate reactions that users may exhibit as a result of the design of the MMUI, which can prevent them from achieving the desired result. Such breakdowns typically concern users' attention, attitude or emotional state (e.g. confusion), difficulty understanding pop-up messages and icon meanings, or recognising differences between objects in the diagrams. Example breakdowns which were specified by the evaluators who assessed the usability of the IMM application include:

> "Text changes too fast for the users to read", and
> "The application jumps around a lot, notably when leaflets are employed. User focus is disjoint, and a concept of where you are and how much you have completed of the work eludes you. ".

Most of the breakdowns were predicted by the MMDs. Five were identified by the experts using the MMT, all of which were experienced by the users. However, due to the nature of the MMT the identification of the breakdowns could be attributed solely to the judgement of the experts rather then the method. The evaluators who applied the MMCW specified four breakdowns, only two of which could be directly attributed to attentional guidelines included in the method. The other two breakdowns were not experienced by the users, and a further one contradicted the users' reaction. The latter prediction was made by the less experienced multimedia designer who tended to misinterpret some of the guidelines and therefore generated all of the false alarms predicted using the MMCW. Finally, two breakdowns were specified by the evaluators using the IMMC, both prompted by the checklist questions, however neither of them

was experienced by the users although they were ratified by the independent judges as valid.

As revealed above, the prediction of cognitive breakdowns was random, as it appeared to depend on the experts' judgement rather than the EEMs, and was not always accurate. The success of such predictions will depend on how well the complexity of the users' cognitive processes and their cognitive capabilities are understood by the evaluators. Therefore, adequate support for this should be provided by EEMs, which should be grounded in empirical evidence of how people learn with IMM.

5.4.2 Prediction of Behavioural Outcomes

The experts sometimes also predicted what actions could be taken as a result of a design problem. The majority of these were specified by the experts using the MMT. Therefore, they were not due to the usage of the method, instead due to the expert's own initiative and judgement. An example of a behavioural outcome is: "It says 'finally click again to see the graph' and I clicked on the graph pane, not on the button, which is confusing". The majority of the predicted behavioural outcomes are, however, descriptions of the evaluators' own behaviour when using the application, which assume that the target users will behave in a similar way. In spite of that, the users who took part in the evaluation behaved in a similar manner on most occasions, and only in one case they did not perform any of the actions specified by the expert. The success rate of the prediction of behavioural outcomes could be explained by the fact that behavioural descriptions contain procedural information by nature, which are *actionable,* as defined by Cockton and Woolrych (2001). As a result, behavioural outcomes are reasonably easy to describe, which is further eased by the fact that often there are only a limited number of actions that could be performed when a problem in the interaction occurs. Therefore, evaluators need to be systematically encouraged and supported by EEMs to infer the likely user behaviour following an interaction problem in order to avoid describing their own behaviour.

5.4.3 Prediction of Comprehension Problems

Three potential comprehension problems were predicted by the SMEs and two by MMDs, whose prior knowledge in Maths was limited. The first group specified where learners with little prior knowledge of the material could experience confusion in trying

to grasp the Maths concepts. However, they did not specify the nature of the misconceptions likely to occur. The second group predicted that one concept was not explained well and identified an area where students may need further learning support. Whereas the first group relied on their pedagogical expertise to envisage likely comprehension difficulties that learners with little prior knowledge may have (without experiencing such difficulties themselves), the second group were in fact novice learners, who described comprehension difficulties they experienced themselves while interacting with the Maths software. Both cases suggest that it is vital to consider the learners' prior subject-matter knowledge for the prediction of valid comprehension problem, which further supports the suggestions made in Section 5.3.3.

Below is an example of a comprehension problem predicted, which is also illustrated in Figure 5.5:

"The notation has been just changed from exp(x) to e^x without any comment. If I was a novice that would throw me completely. They haven't said that you can write exp() as e^x. If this is for novices, it ought to say that there are two different notations and what they are."

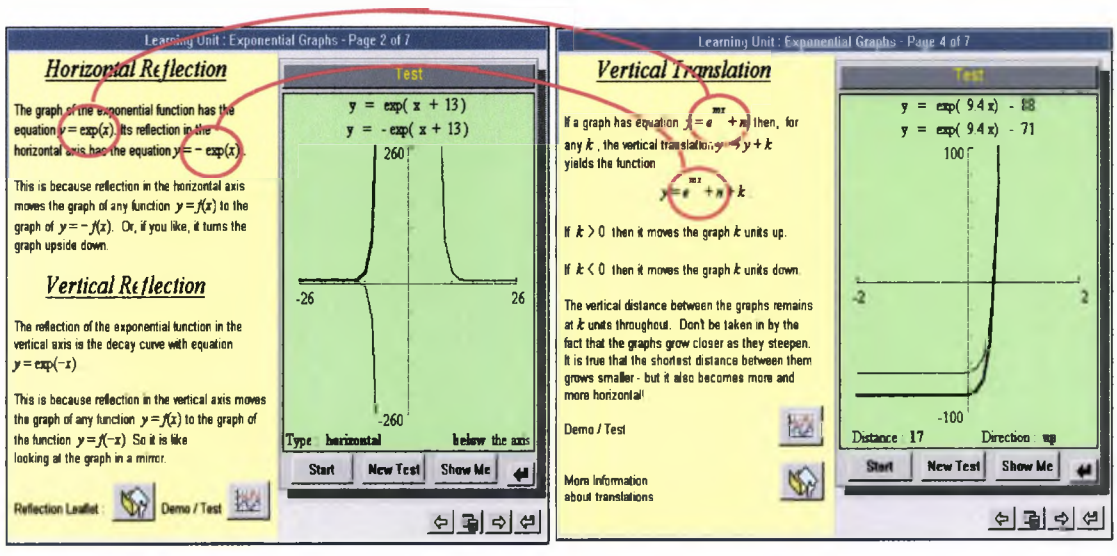


Figure 5.5: Sample Screens Illustrating Inconsistency of Maths Notation Used

5.5 Theoretical Frameworks of Learning with Multimedia

The primary aim of the investigation conducted in the second part of this thesis is to improve the ability of EEMs for IMM to predict valid user problems, including not only user interaction problems, but also difficulties in comprehension and learning with IMM. Therefore, an understanding of how people interact and learn with IMM is necessary to inform the process of assessing how effectively the design of IMM supports learner's interaction and comprehension processes.

The process by which a learner comprehends the information presented in a MMUI is a cognitive one. Thus, in order to produce adequate evaluation methods for analysing the effectiveness of MM presentations, a model of learning with MMUIs is required. This section presents a survey of existing instructional and cognitive psychology studies concerning the cognitive processes of learning with MM. These studies are used to inform the generation of hypothetical propositions regarding how the effectiveness of EEMs for IMM can be improved. The hypotheses made are presented later in this chapter.

Furthermore, in a learner-centred approach to the developing instructional multimedia, the primary focus should be on the learner and their needs (Soloway, 1998), their existing knowledge and experience, pre-determined learning style, and individual approach to learning (Wild and Quinn, 1998), as well as on their particular information processing capabilities and limitations (Scaife and Rogers, 1996). Therefore, usability evaluators of IMM need to assess whether the learners' needs have been taken into consideration, and whether the diversity of their skills and abilities has been adequately addressed. Therefore, this section also describes what characteristics of the target learner population are important to consider when assessing the effectiveness of such software.

5.5.1 Cognitive Frameworks of Learning

Craig, Mehrens and Clarizio (1975) distinguish between the process of learning and its product. The products include facts, concepts and principles, skills, attitudes and values, and behaviour of various kinds. The products also include both what one is capable of and what one is predisposed to do. There are three main elements of learning: *the learner*, *the stimulus* and *the response*. The response is the nervous, including mental

and muscular, activity of the learner that results from stimulation. Because the precise nature of mental and muscular activity is often difficult to observe and describe, frequently the effects of this activity are referred to which are called a performance. Learning is commonly inferred from the observation of a performance following exposure to a stimulus situation.

Craig, Mehrens and Clarizio (1975) specify that there are five phases of learning: attention, perception, acquisition, retention and transfer, which are depicted in Figure 5.6. The interaction of the learner with stimuli in the environment at each stage is different.

- **Attention:** is a necessary preparatory step, it directs students' attention to some stimuli than others. It prepares students for the perception phase of learning.
- **Perception:** the input to the senses is registered and meaning is added. What is perceived depends partly on prior learning and partly on what stimuli one attends to. It often involves a complex interrelating of information from the environment and information retained from prior learning. The student's perception of a situation has affective components also.
- **Acquisition:** in this phase the student acquires a new capability or disposition.
- **Retention:** something learned cannot be demonstrated or used unless it is retained for at least a brief time. There are two types of retention: short- and long-term. Efforts to improve retention should give attention to what is learned initially (short-term) and how this learning is organised (structure of information), to the distribution and type of practice (interactivity) or review (assessment), and to what is learned before and later.
- **Transfer:** educational efforts that carry over from one situation to another. Transfer cannot occur unless there is some permanency to the effects of initial learning, hence transfer and retention are closely related phenomena. In theoretical questions concerning the permanency of learning, retention refers to later reproduction of learning products in situations exactly like that in which the original learning occurred, whereas transfer refers to the effects of learning in different situations. In reality, however, no two situations are exactly alike, so any demonstration of retention must involve transfer to some degree.

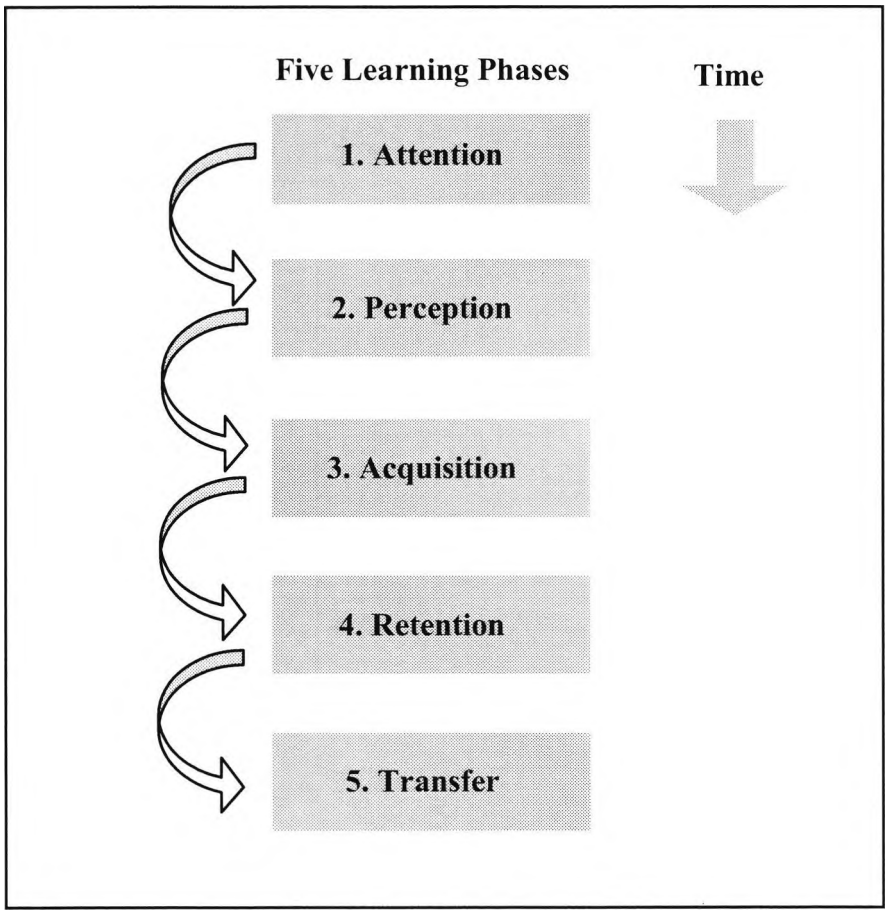


Figure 5.6: Five Cognitive Phases of Learning (Craig, Mehrens and Clarizio, 1975)

Kozma (1991) and other advocates of the Constructivist learning paradigm such as Jonassen (Jonassen, 1998; Jonassen and Rohrer-Murphy, 1999) and Honebein, Duffy and Fishman (Honebein, Duffy and Fishman, 1993) describe learning as an active, constructive process whereby the learner strategically manages the available resources to create new knowledge by extracting information from the environment and integrating it with information already stored in their memory. This process is constrained by cognitive factors such as the duration and amount of information in short-term memory, the task-relevant information that is available in long-term memory, the structure of this information, the procedures that are activated to operate on it, etc. Consequently, the process of learning is sensitive to characteristics of the external environment, such as availability of specific information at a given time, the duration of that availability, the way the information is structured and the ease with which it can be searched.

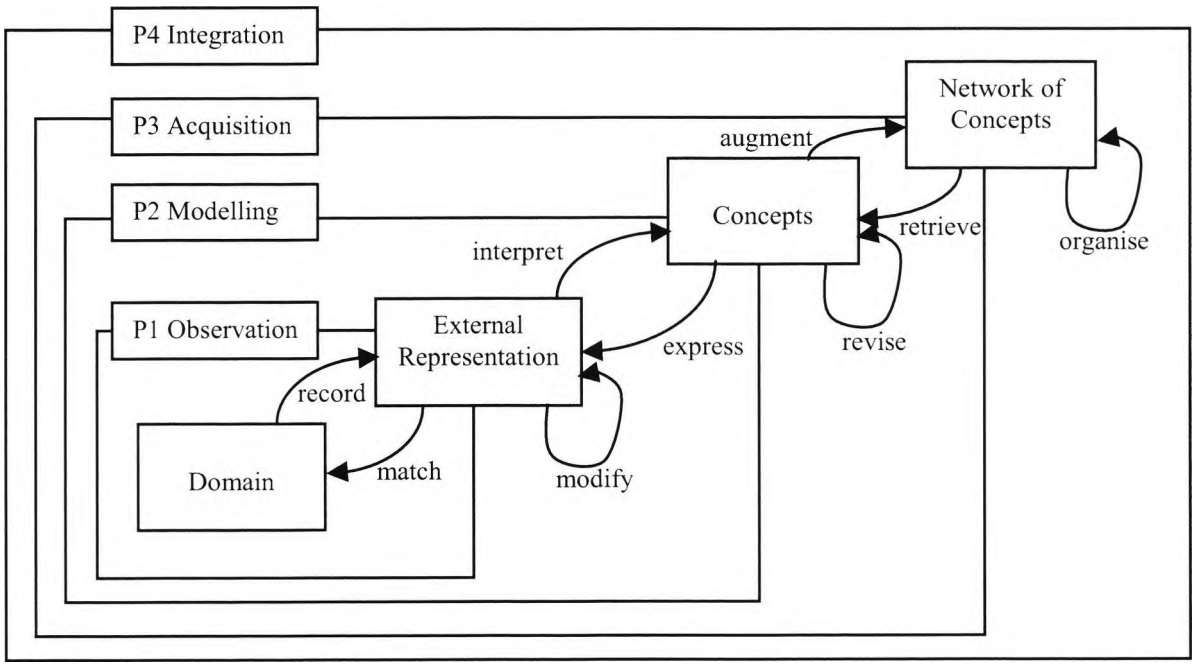


Figure 5.7: Components, Processes and Sub-processes of Conceptual Learning (Cheng, 1999)

Cheng (1999) also highlights the critical role of external representations on conceptual learning. The author emphasises that the external representations used for learning can substantially determine what is learnt and how easily this occurs. Therefore, to improve conceptual learning it needs to be ensured that effective representations are used. Cheng (1999) proposes a framework of conceptual learning, which is presented in Figure 5.7. Figure 5.7 illustrates the major tasks facing learners during conceptual learning. It also shows the complexity which designers and evaluators of external representations for learning face in ensuring their effectiveness. This is achieved by depicting the major relationships between the external sources of knowledge exhibited in books, computer software and physical models (Domain and External Representation elements of the model) and the internal network of concepts the learner is gradually developing (Concepts and Network of Concepts elements). Cheng (1999) emphasises that, as learning is a constructive process, consideration must be given to the role of the external representations used for the domain, and the role of the individual's Concept schema. Due to the limitations of human information processing, sound consideration of both components is necessary in developing and evaluating the effectiveness of external representations.

Cheng (1999) specified four main processes involved in conceptual learning, which are analogous to some of the phases described by Craig, Mehrens and Clarizio (1975).

- **Observation:** generates and checks expressions that are descriptions of phenomena.
- **Modelling:** follows observation and involves generating new expressions to tie together descriptions of particular phenomena or concepts represented in the external sources.
- **Acquisition:** involves mentally constructing a new concept.
- **Integration:** is a process of adding a concept to the network of concepts or of modifying the structure of an existing network of concepts in some way.

Therefore, Cheng (1999) suggests that to achieve pedagogically effective learning experiences it needs to be ensured that external representations of concepts adequately support all of the above processes of learning.

5.5.2 Cognitive Theories of Learning with Multimedia

In recent years, multimedia representations have been increasingly used to provide motivating and engaging learning experiences. The effective design of IMM learning environments is based on the assumption that users' comprehension and recall of information increases when information is presented with two media sources concurrently. This is derived from the Dual Coding theory (Clark and Paivio, 1991; Verdi et al., 1996), which postulates that humans possess two distinct information-processing systems: one that represents information verbally and one that represents information visually.

A number of studies have been performed examining the effect of different media combinations on learners' recall and comprehension of information (e.g. Spencer, 1991; Mayer, 1994; Scaife and Rogers, 1996; Moreno, and Mayer, 1999). The findings of such studies provide evidence regarding how people attend to and perceive information from visual and audio media and the implications of different combinations of static and dynamic, visual and verbal media on users' cognitive processes. For example, Large et al. (1996, 1995, 1994), and Mayer and Anderson (1992) studied the effect of combining different media on students' comprehension, recall and problem-solving capabilities. Large et al. (1995) examined how animation can enhance text when explaining complex

descriptive information. They found that such combinations of media significantly improved recall, as subjects were encouraged to build representation connections. The results of such studies are significant for the effective design of IMM, as they promote understanding of the impact that various combinations of media can have on users' comprehension and recall of information. However, their application to the design and evaluation of IMM is limited as they have examined only a limited amount of media combinations (e.g. text and animation) under a restricted number of factors (e.g. procedural information). Therefore, more systematic frameworks of how people learn with MM are required to inform more effective design and evaluation of IMM.

One such framework was developed by Faraday and Sutcliffe (1997b). The authors studied the cognitive processes users engage in while using MM presentations in order to determine their implications for the design of such presentations (Faraday and Sutcliffe, 1996, 1997a). Their cognitive framework of MM comprehension, presented in the top section of Figure 5.8, follows some of the cognitive processes identified in the previous section which are required for users to understand a multimedia presentation composed of visual and auditory stimuli. The structure of the framework is based on models of cognitive architecture and information processing. These models suggest that several processes are required to move from images and language in the world to meaning in the learner. The framework has three main components: (i) attention, perception and short-term memory, (ii) working memory and (iii) comprehension. The issues raised in the framework can be linked to key issues in the evaluation of IMM presentations. Based on this framework the authors have created the MM Cognitive Walkthrough method, studied in the previous chapter.

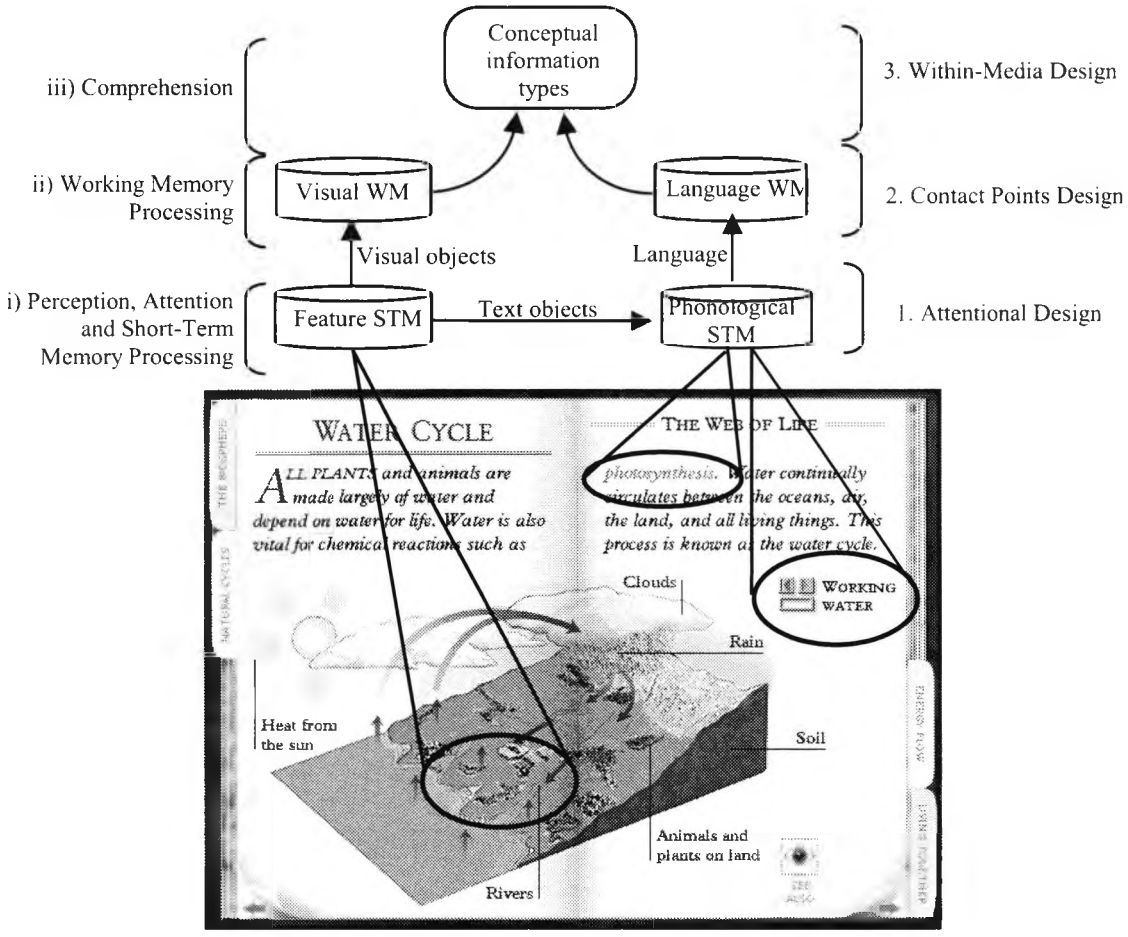


Figure 5.8: Overview of MM Cognitive Framework (Faraday and Sutcliffe's, 1997) (top) and Related Evaluation Issues in a MM Representation of the Water Cycle (bottom) (from DK© Encyclopaedia of Nature)

The role of each component in the framework and its implications for IMM evaluation are briefly discussed in the following paragraphs.

i) Perception, Attention & Short-Term Memory

The framework first covers attention to and perception of different media and combinations of media. Attentional processing has two separate systems for dealing with auditory and visual input. The auditory part of the presentation is processed by phoneme recognisers, which extract language from other sounds. The phonemes are held in the Short-Term Memory (STM) phonological loop, which has a limited duration buffering capacity. The visual part of the presentation is scanned serially . The visual information is processed by low level visual feature recognisers which extract colour, edges and texture. This is stored in a visual STM, called 'feature STM' in Figure 5.7, which also has a limited duration capacity. The issues for IMM evaluation concern

'attentional design' - how elements of the presentation should be designed to make important information more salient, or 'emphasise' important details of the subject matter content; and what restrictions STM places on how auditory and visual media elements can be presented together within a presentation.

ii) Working Memory

The second part of the framework deals with how working memory is used to combine visual and verbal media together. Visual and verbal inputs are translated into units of meaning, called 'propositions' in Faraday and Sutcliffe (1997). Working memory acts as a storage buffer for these propositions. This allows coherence within and between these propositions; it is used to resolve co-references such as anaphor. 'Contact points' describe ways in which visual and verbal media are integrated by co-reference. Issues for MM evaluation include how to combine media to form a single whole, and how to effectively design 'contact points'.

iii) Comprehension

Finally, the semantic content of the propositions will be constructed within a mental model. A mental model is a representation derived from the contents of the proposition, the user's long-term memory knowledge and other existing mental models. Mental models are held for a longer duration in an episodic memory store. The evaluation issues concern individual media design, in particular how media resources should be designed to aid comprehension of the information.

Although the framework presented above has proven very useful for designing effective multimedia presentations (Faraday, 1997), it has one major limitation. It does not model how users interact with IMM presentations in order to achieve their goals. Researchers, such as Jih and Reeves (1992), and Narayanan and Hegarty (1998) have investigated issues of users' interaction with MM learning environments. According to Jih and Reeves (1992) in order to make meaningful responses to interactive learning software, learners must cope with and integrate three types of cognitive load: the content of the information, the structure of the program, and the response strategies available. In order to respond to any one of these demands, the student must perceive options, conceptualise a choice and take some physical action. The human-computer interface is

the vehicle that allows perceptual, conceptual and physical contact between the user and the software.

The authors continue that users seek new information in ways that depend on and are limited by their current mental models. They believe that mental models affect such factors as the effort people devote to tasks, their persistence, expectations and prediction of results, and the level of satisfaction after task execution. Furthermore, when learners possess an adequate mental model of the structure of hypermedia or other complex IMM, they are less likely to become disoriented and they are more likely to learn. Jih and Reeves (1992) conclude that an ideal user's working mental model is the one that is consistent with the conceptual model of the interface developed by designers.

Narayanan and Hegarty (1998) also consider how comprehension and navigation strategies are related in a way allowing appropriate descriptions and depictions to be delivered to the user at the relevant stage of comprehension. The authors suggest that the way to achieve that is by balancing between linear and non-linear information access.

5.5.3 The Learner of IMM Applications

The typical users of instructional multimedia software are not developers of software systems, neither are they knowledgeable workers. Instead, they are people who would like to be informed, learn or explore an area of interest. The users of multimedia systems are likely to have a variety of word-processing skills, but they are not likely to be very skilled in desktop publishing, graphics or video and audio production and editing. Furthermore, they will not have extensive knowledge of the material represented, unless they are teachers or trainers who will be using the software to teach students. Therefore, they will have different levels of abilities and prior domain knowledge. The differences exhibited by people can affect the interaction between a human being and a computer system (Morris et al., 1994). For instance, *user's individual aptitudes and abilities* may influence the degree to which the use of visual technologies are beneficial (Morris et al., 1994). *Age* has proven to have an influence on user attitude towards and satisfaction of the use of various media in educational software (Passig and Levin, 2001). *The technical, educational and domain-specific background* of the user population must be considered. Many concepts taught in school

or in educational computer systems assume certain knowledge and abilities that are necessary to grasp the new concepts. Furthermore, Kalyuga, Chandler and Sweller (1998) argue that whether IMM representations are intelligible often depends on the *prior level of experience* of the learner, and whether they have sufficient knowledge to understand the representations. Giouvanakis, Samaras and Tarabanis (2001) also emphasise the importance of the level of the learners' *prior knowledge* for the design of pedagogically effective IMM interfaces. Therefore, exploiting information about users' *prior domain knowledge* during usability evaluation of IMM can improve inferences regarding how understandable the material will be for the learners and predictions of potential problems that learners with low prior knowledge may experience.

The impact of individual differences in students' abilities on learning is one of the major themes in the field of instructional software design. Learning is a function of the learner, the content to be learned, the features of the instruction and the context of learning. Each learner brings his or her unique characteristics to the learning environment. The cognitive demands of human learning are complex and involve numerous variables which interact with learner characteristics (Jih and Reeves, 1992). According to the authors, learning is influenced by three dimensions of individual differences: a) *personalistic factors*, like prior knowledge and experiences, b) *affective factors*, such as motivation and attitudes, and c) *physiological factors*, e.g. eye-hand co-ordination and visual acuity.

Other individual differences that are especially relevant to the study of learner interactions with interactive multimedia learning systems are *learning styles, learner's existing knowledge and experience, their acquired individual approach to learning, their motivation, cognitive and perceptual characteristics* and *learner specific needs* (Wild and Quinn, 1998).

Therefore, the design and evaluation of high quality MMUIs should take differences in learners' characteristics into account throughout the entire development cycle. As highlighted in Section 5.3.3, one main limitation of the three EEMs for IMM studied is that none of them encourages or adequately supports evaluators in considering relevant user characteristics while assessing the effectiveness of IMM. The lack of consideration of learner's cognitive abilities and their prior domain knowledge was found to be the

main reason why such a significant number of comprehension problems were missed by the usability evaluators. Therefore, efforts directed towards improving the effectiveness of such EEMs should propose ways of building explicit evaluation support, which should ensure that the diversity of learners' characteristics are adequately addressed by the IMM application.

5.6 Hypotheses for Improving the Effectiveness of EEMs for IMM

The previous sections presented comprehensive analysis of the major factors that constrain the ability of EEMs for IMM to predict valid user interaction and comprehension problems, as well as those factors which support the prediction of such problems. Based on these findings, desirable characteristics for EEMs for IMM were formulated and four hypothetical predictions were made as to how implementing them in such methods could enhance their effectiveness. The taxonomy of user problems was also used to provide specific requirements of what components of usability problems need to be predicted, and how such problems should be specified. Finally, the four hypotheses were also based on the theoretical frameworks presented in the previous section. This aims to ensure the validity of the hypotheses as the deductions made are drawn from previous empirical research and theory in the field.

As defined by Strauss (1993), a hypothesis is "a provisional answer to a question about conceptual relationships", or "a tentative proposition set forth as a possible explanation for an occurrence" as defined by Leedy (1993). Therefore, the hypotheses presented below describe tentative suppositions of how evaluation methods' validity and thoroughness can be improved. Strauss (1993) also explains that hypotheses are both provisional and conditional. Therefore, each definition presented below states both precisely what part of methods' performance can potentially be improved (i.e. the provisional result) and the factors (i.e. the condition/s) hypothesised to improve their performance. The hypotheses in effect propose characteristics that need to be built into EEMs for IMM to overcome limitations 1, 3 and 4, as described in Section 5.3. Section 5.7 will discuss how the hypotheses were tested, and will explain the circumstances under which each hypothesis can be totally accepted, partially accepted or rejected.

Each hypothesis is concerned with improving the prediction rate of a different component/s of user interaction or comprehension problems, as depicted in Figure 5.9. The arrows in Figure 5.9 show which hypothesis addresses the prediction of which component.

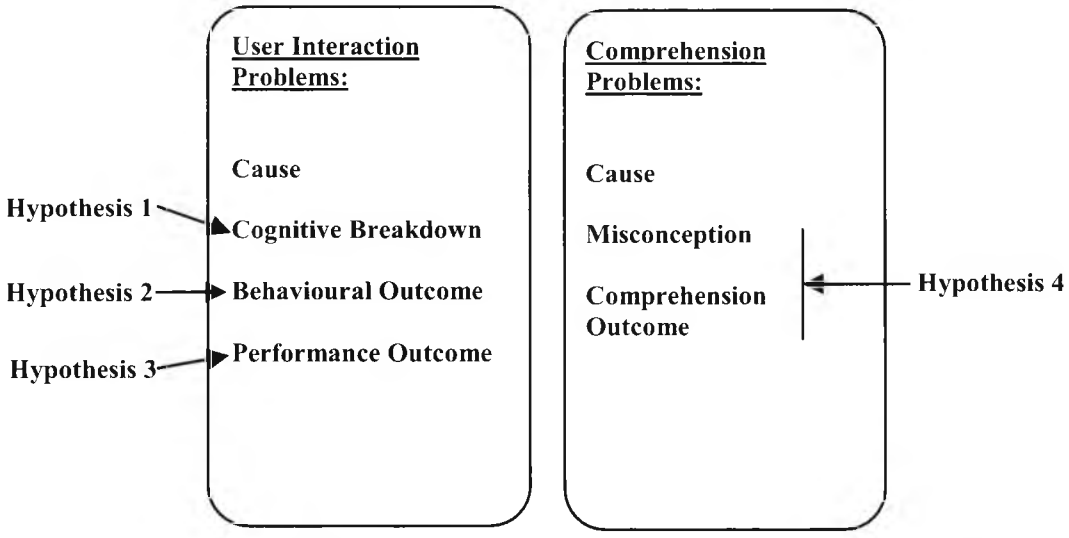


Figure 5.9: Usability Problem Components and the Hypothesis Supporting the Prediction of Each One

All four hypotheses aim to increase the prediction rate of components concerned with the effect of an interaction or a comprehension problem on the user, their behaviour and their performance. They aim to support evaluators in making accurate inferences regarding what the likely effect of a problem may be on the target user. In this way, the hypotheses propose ways of improving the *validity* of EEMs for IMM by increasing the number of valid problems and decreasing the number of false alarms. As an extension of increasing the number of valid problems predicted using EEMs, their *thoroughness* can also be enhanced to a degree. Hypothesis 4 aims specifically to increase the number of valid comprehension problems predicted during expert evaluations of IMM, thus aiming to enhance the *thoroughness* of EEMs in this specific area.

Each hypothesis is now described in detail.

5.6.1 Hypothesis 1

Definition: The prediction rate of valid *attentional, perceptual, information integration and comprehension breakdowns* can be improved by asking experts to explicitly consider how different forms of media representations affect users' cognitive processes of attention, perception, integration and comprehension of information presented in various media.

The aim of this hypothesis is to support experts in identifying breakdowns or difficulties occurring in the early cognitive phases of learning with MMUIs, namely attention, perception, integration and comprehension of information. Breakdowns in these phases can lead to failures in long-term learning, thus it is important to identify them. It is hypothesised that by systematically providing guidance in an EEM regarding how the design of the multimedia user interface and the users' interaction with it can influence their cognitive processes, evaluators would be able to infer how design aspects can potentially hinder such processes. As suggested in Section 5.4.1, such guidance needs to be based on cognitive science theories of how people process cognitively information presented in multiple media to ensure that it is sound. The hypothesis assumes that experts with little formal cognitive psychology training would be able to make inferences about how particular design aspects could influence certain users' cognitive processes, by providing them with propositions regarding how people process information presented in various media.

To be able to predict such breakdowns, evaluators need to consider the three main elements of learning with IMM: the learner (including their information processing abilities and limitations), the external multimedia representations which act as stimuli on the learner, and the learner's response that results from the stimulation. Faraday and Sutcliffe's (1997) cognitive framework, presented in Section 5.5, can be used as a basis for developing evaluation criteria for assessing how effective external IMM presentations support learner's cognitive processes and utilise their information processing abilities.

The conjectures made in this hypothesis also draw upon the findings presented in Sections 5.4.1, 5.3.1 and 5.3.4. In particular, it was revealed in Section 5.4.1. that two

valid attentional breakdowns were identified using the MMCW. A limitation of the evaluation guidelines used to predict those breakdowns, is that the implications of design aspects for the user's cognitive load and cognitive processes of attention, perception and integration of visual and verbal inputs are not explicitly explained. Therefore, it is hypothesised that making such implications explicit and asking evaluators to consider such implications can potentially increase the prediction rate of cognitive breakdowns. The latter can be achieved by introducing a problem report format, explicitly requiring evaluators to specify breakdowns of user interaction with MMUIs, as proposed in Section 5.3.4. Supporting evaluators in considering cognitive implications of interaction with MMUIs and explicitly asking them to specify cognitive breakdowns resulting from such interactions are inextricably connected, and can be viewed as two sequential steps of the process of specifying user cognitive breakdowns of interaction. Therefore, this hypothesis suggests that both conditions need to be supported adequately by EEMs.

5.6.2 Hypothesis 2

Definition: The prediction rate of valid *behavioural outcomes* can be improved by ensuring evaluators are sufficiently acquainted with the target user group, and by supporting them in inferring the likely user behaviour following a breakdown in users' interaction with the MMUI.

This hypothesis has two conditions, namely i) Evaluators need to know the intended users well, and ii) Evaluators need to be explicitly asked to specify the likely behaviour outcomes. Although, the conditions are quite different in nature, they are both considered necessary for achieving the hypothesised improvement. Similarly to Hypothesis 1, both conditions need to be present and adequately supported.

This hypothesis presumes that if evaluators are familiar with the target user group they will be able to infer more accurately the likely user behaviour following a breakdown. Knowing the target users well is likely to reduce the effect of experts substituting assumptions regarding likely user behaviour with their own behaviour, as observed in the method comparison study (see Section 5.4.2). This is hypothesised to make expert

predictions more accurate rather than biased with evaluator's own behaviour, as suggested by Hertzum and Jacobsen (1999).

As also revealed in Section 5.4.2, the evaluators rarely specified behavioural outcomes, and when they were specified it was due to the evaluators' own initiative. Therefore, to ensure the prediction of behavioural outcomes is more systematic, evaluators need to be explicitly encouraged to specify what actions users are likely to take following a breakdown.

According to Jih and Reeves (1992), in order to make meaningful responses to IMM, learners must cope with and integrate three different types of cognitive load: the content of the presentation, the structure of the user interface, and the response strategies available to them. In order to respond to any one of these demands, the learner must perceive options, conceptualise a choice, and take some physical action. When a user experiences a breakdown in the interaction they are cognitively hampered in some way, e.g. they experience difficulties understanding the material, which means that there is a gap between the level of the content and the learner's prior knowledge of the material, or there is a mismatch between the user's expectations and the actual system response. The user's behaviour following such breakdowns is dependent on the severity of the breakdown, which influences the user response strategy. The user response strategy is typically determined by the user's conceptual model of the system, the response options available to them, user's goals and the task they are required to do. Thus, all these factors need to be considered by usability evaluators when determining the likely user behaviour following a breakdown.

5.6.3 Hypothesis 3

Definition: The prediction rate of valid *performance outcomes* can be improved by ensuring evaluators are sufficiently acquainted with the target user group, and by supporting them in inferring the implications for the users' task resulting from the behaviour they are likely to adopt.

The findings of the user problem set analysis revealed that behavioural and performance outcomes are very tightly dependent on each other, i.e. what the user will do as a result

of a cognitive breakdown will determine whether they will be able to recover and still achieve their goals or not. Therefore, there is a strong dependence between the two, and the severity of the first directly influences the outcome of the second. Further analysis of the cause-breakdown-outcome sets of user interaction problems revealed that there are patterns in the type of behaviour outcome and the severity of the performance outcome. For example, if a user repeats an action (e.g. re-reads text or repeats tests), their learning performance is likely to improve. However, if a user skips a part of the presentation their performance is affected severely. For instance, their learning suffered, or they could not continue with their task without external help. Therefore, evaluators need to be encouraged to consider the relationship between the two together when inferring performance outcomes.

5.6.4 Hypothesis 4

Definition: The prediction rate of valid *comprehension difficulties* can be improved by supporting evaluators in predicting the implications of IMM design aspects on learners' comprehension of the content, particularly considering the needs of learners with little prior knowledge of the subject matter.

The term 'comprehension difficulty' is introduced to describe learning difficulties predicted by some of the experts. It differs from the term ' comprehension problem', which describes the problems experienced by the learners on the knowledge tests, as defined in Section 5.2.2.2. These problems were defined in terms of the misconception and the learning performance outcome observed. The misconceptions depend largely on the learners' prior conceptual framework and other factors both internal and external to the learners, whereas the learning performance outcomes are tightly dependent on the knowledge assessment approach which was adopted. Therefore, experts cannot be expected to predict exact learner misconceptions or learning performance outcomes with any degree of specificity. Instead, evaluators may be able to predict areas of the IMM design which are likely to cause learners concept comprehension difficulties.

Therefore, this hypothesis aims to test whether evaluators are capable of systematically predicting where learners with low prior domain knowledge may have difficulties in acquiring new knowledge of the subject matter. The aim of this hypothesis is to build features within EEMs for IMM that will explicitly support experts to make more

systematic and accurate predictions of areas where learners with low prior domain knowledge may experience difficulties comprehending the material.

As discussed in Section 5.5.3, learners' prior domain knowledge has an effect on how comprehensible they will find the MM presentation (Kalyuga, Chandler and Sweller, 1998). Therefore, in predicting what areas of IMM presentations may cause comprehension difficulties to target learners, their prior knowledge needs to be considered to ensure that their level of prior knowledge is adequately addressed. Furthermore, the results of the evaluation studies presented in Sections 5.3.3. and 5.4.3 further emphasise the need to consider students' prior knowledge of the material when evaluating the instructional effectiveness of IMM user interfaces. Therefore, the second condition of this hypothesis suggests that learners' prior domain knowledge should be considered in order to improve the prediction rate of comprehension difficulties. Finally, although a number of other learner characteristics need to be considered, as discussed in Section 5.5.3, this hypothesis will only test whether considering learners' prior domain knowledge can contribute to predicting a greater number of comprehension difficulties.

5.7 Methodological Approach for Testing the Four Hypotheses

To be able to test whether improvement of the validity and thoroughness of EEMs for IMM can be achieved based on the conjectures made in the four hypotheses, it was decided to develop an EEM for IMM that implements all the conditions specified in the hypotheses. The evaluation results produced by the new EEM need to be validated against those produced by the user tests, which will reveal whether improvement in the number of valid user problems predicted is achieved and to what extent. After that, the method's predictive power needs to be tested under similar conditions as the previous study.

This section proposes an approach to testing each hypothesis, and explains the conditions under which each hypothesis can be totally accepted, partially accepted or rejected, which are summarised in Table 5.2. After that, the following chapter will describe the development of a new EEM and the empirical study validating its effectiveness.

Table 5.2: Conditions for Accepting and Rejecting Each Hypothesis

	Totally Accept	Partially Accept	Reject
Hypothesis 1	If cognitive breakdowns are predicted for > 90% of predicted problems, AND If >90% of predicted breakdowns are valid.	If cognitive breakdowns are predicted for <90% of predicted problems, AND/OR If <90% of the predicted breakdowns are valid.	If no cognitive breakdowns are predicted, OR If all of those predicted are invalid.
Hypothesis 2	If behavioural outcomes are predicted to >80% of problems, AND If >90% of predicted behavioural outcomes are valid.	If behavioural outcomes are predicted to <80% of problems, AND/OR If <90% of predicted behavioural outcomes are valid.	If no behavioural outcomes are predicted, OR If all of those predicted are invalid.
Hypothesis 3	If performance outcomes are predicted to >80% of problems, AND If >90% of predicted performance outcomes are valid.	If performance outcomes are predicted to <80% of problems, AND/OR If <90% of predicted performance outcomes are valid.	If no performance outcomes are predicted, OR If all of those predicted are invalid.
Hypothesis 4	If >90% of comprehension difficulties predicted are valid.	If <90% of comprehension difficulties predicted are valid.	If no comprehension difficulties are predicted, OR If all those predicted are invalid.

5.7.1 Testing Hypothesis 1

In order to test this hypothesis, evaluation criteria will be created as part of the new EEM to explain not only the cognitive aspects of multimedia design, but also to guide evaluators in inferring the likely implications of the design on users' cognitive processes. Furthermore, a uniform problem report format will be introduced, which requires evaluators to explicitly specify the likely cognitive breakdowns incurred as a result of the usability problems.

After the new EEM is applied to the evaluation of the IMM software, the problems predicted by the experts will be analysed to establish whether they specify any cognitive breakdowns users may experience. The predicted problems will also be compared to those experienced by the users to establish the validity of the cognitive breakdowns predicted.

The hypothesis will be **totally accepted** if over 90% of the problems predicted by experts specify cognitive breakdowns and at least 90% of those identified are valid ones. 10% discrepancy between the predicted and observed breakdowns is allowed as

the experienced problem set may be incomplete. This could be due either to the small sample of users who took part, or to the fact that the users may have experienced the cognitive breakdowns but they were not captured on the video tapes.

The hypothesis will be **partially accepted** if there are breakdowns predicted, but to a small number of problems and/or if there are breakdowns predicted which are not valid ones. This can signify that either the evaluation criteria defined in the new EEM are not entirely appropriate or that they are appropriate but cannot be used easily in the particular evaluation context.

The hypothesis will be **rejected** if no breakdowns are predicted, or all the ones that are predicted are invalid ones. Such results may indicate either that the evaluation criteria are totally inappropriate, or that it is unreasonable to expect evaluators to predict cognitive breakdowns with the guidance provided in the new EEM.

5.7.2 Testing Hypothesis 2

To test this hypothesis, evaluators need to be presented with a detailed user specification which can provide them with a better understanding of the target users. The problem report format which will be introduced as part of the new EEM needs to include a section in which evaluators will be asked to specify potential behavioural outcomes. finally, guidance for predicting behavioural outcomes will also be included in the new EEM.

The problems predicted by the experts using the new EEM will need to be analysed to establish whether they specify any behavioural outcomes that users may experience. The predicted problems will also be compared to those experienced by the users to establish the validity of the behavioural outcomes predicted.

The hypothesis will be **totally accepted** if over 80% of the problems predicted specify the likely behavioural outcomes and at least 90% of those specified are valid ones. As not all breakdowns may result in immediately observable behaviour it is assumed that in up to 20% of the problems predicted such may not be specified.

The hypothesis will be **partially accepted** if there are behavioural outcomes predicted but for a small number of problems, and/or if there are behavioural outcomes predicted but not all are valid. The first result may indicate that the problem report format or the guidance is insufficient to ensure all behavioural outcomes are specified. The second result may signify that the user description is incomplete or insufficient for understanding how users are likely to behave when they encounter a problem in their interaction.

The hypothesis will be **rejected** if no behavioural outcomes are predicted, or all the ones specified are invalid. Such a result may indicate that all conditions failed to produce the hypothesised improvement and a different approach to testing the conjecture is necessary.

5.7.3 Testing Hypothesis 3

To test the hypothesis evaluators again will need to be presented with a detailed user specification and the problem report format will need to include a section in which evaluators will be asked to specify potential performance outcomes.

The problems predicted by the experts using the new EEM will be analysed to establish whether they specify any performance outcomes users may experience. The predicted problems will also be compared to those experienced by the users to establish the validity of the performance outcomes predicted.

The hypothesis will be **totally accepted** if over 80% of the problems predicted specify likely performance outcomes, and at least 90% of those identified are valid ones. It is assumed that such outcomes may not be specified to 20% of the problems, as not all problems may have resulted in immediately measurable performance outcomes. Furthermore, due to the small sample of users who took part in the user tests, some valid performance outcomes may not have been identified.

The hypothesis will be **partially accepted** if there are performance outcomes predicted but not for all problems, and/or if there are performance outcomes predicted which are not valid ones. The first result can indicate that the problem report format is insufficient to ensure all performance outcomes are specified, and better guidance is required. The

second result may signify that the user description is incomplete or insufficient for understanding how the achievement of the user tasks will be affected when they encounter a problem in their interaction.

The hypothesis will be **rejected** if no performance outcomes are predicted. Such a result may indicate that both conditions failed to produce the hypothesised improvement, and a different approach to testing the conjecture is necessary.

5.7.4 Testing Hypothesis 4

To test this hypothesis the evaluation criteria of the new method should encourage evaluators to consider how aspects of the design of the MMUI will support or hinder learners, particularly those with little prior knowledge, to comprehend the material.

The problems predicted by the experts using the new EEM will be analysed to establish whether they specify any comprehension difficulties learners may experience. The predicted problems will also be compared to the comprehension problems experienced by the users to establish the validity of the comprehension difficulties predicted, if any.

The hypothesis will be **totally accepted** if at least 90% of the comprehension difficulties identified by the experts were experienced by the learners. 10% of inaccuracy is allowed, as the users who participated in the study may not have verbalised all difficulties they experienced, or due to the small sample of users not all potential comprehension difficulties may have been discovered.

The hypothesis will be **partially accepted** some of the comprehension difficulties identified are invalid, i.e. contradict the users' experience during their interaction with the IMM application (e.g. students did reasonably well on the tests included in the application) or the learners showed good understanding of the material during the post-exposure comprehension tests. The hypothesis will also be partially accepted if comprehension outcomes are defined only to some problems posing such difficulties to students. Such a result will indicate either that some of the suggestions made in the evaluation criteria are inappropriate, or that the evaluators made inaccurate assumptions regarding the learners' ability to comprehend the material. The latter would depend on

the subject matter knowledge and the instructional expertise of the evaluators, and not on aspects of the EEM.

The hypothesis will be **rejected** if all comprehension difficulties predicted are invalid or none have been predicted. The first result may indicate that the guidance given in the new EEM is inappropriate or that the evaluators made wrong assumptions about the learners' ability to comprehend the material. The second result may indicate that either the conjecture made in the hypothesis is invalid or that the approach adopted in testing the hypothesis is inappropriate.

5.8 Conclusions

This chapter promotes understanding into the nature of user problems experienced while learning with IMM, what characteristics are undesirable in an EEM for IMM as they limit its performance, and what features are desirable to implement in an EEM, as they promote effective prediction of valid user problems. Based on that rationale, four hypotheses were defined proposing conjectures as to how to improve the predictive power of EEMs for IMM by eliminating limiting features and building performance-enhancing ones into such methods. The conjectures made in the hypotheses are founded not only on empirical but also on theoretical work, thus ensuring the validity of the hypotheses.

The following chapter presents how a EEM was developed to implement the performance enabling characteristics defined in the four hypotheses, together with an empirical study which was conducted to test the hypotheses, following the methodological approach proposed in Section 5.7.

Chapter 6

Creation and Validation of a New Expert Evaluation Method for Instructional Multimedia

Chapter 6

Creation and Validation of a New Expert Evaluation Method for Instructional Multimedia

6.1 Introduction

This chapter presents how a method for expert usability evaluation for IMM was created to fulfil objective 4 of this research:

> *Objective 4:* To develop an EEM for IMM, which aims to improve the performance of the existing EEMs by implementing the characteristics specified in the hypotheses defined in Objective 3.

The main aim of the newly created method is to improve on the effectiveness of the existing EEMs by implementing the conjectures made in the hypotheses defined in Chapter 5. The creation of a conceptually new EEM was not considered feasible within the time limits of this PhD research. Therefore, it was decided to base the method on the Multimedia Cognitive Walkthrough developed by Faraday and Sutcliffe (1997), and consequently the new EEM was called Educational Multimedia Cognitive Walkthrough (EMMCW). The rationale for selecting the MMCW and how the EMMCW method was developed are described in Section 6.2.

An empirical study validating the validity, the diagnostic power and the thoroughness of the EMMCW was conducted, which achieves the final objective of this research:

> *Objective 5:* To validate the effectiveness of the EEM developed as part of Objective 4.

As the evaluation method implements the four hypotheses defined in Chapter 5, by empirically validating the EEM the four hypotheses are tested. The performance of the

developed EEM is measured in two ways, as depicted in Figure 6.1. Firstly, the method's validity and actual thoroughness are established by validating the predicted usability and learning problems against the problems experienced by the users during the user tests described in Chapter 4. These findings are presented in Section 6.4. Secondly, the predictions made using the new EEM are compared to those produced using the original MMCW and to the other two EEMs studied to establish whether improvement over their validity, actual thoroughness and diagnostic power has been achieved, and to measure its extent. The results of this comparison and the conclusions made are presented in Section 6.5. Finally, based on the results of the expert evaluations conducted, conclusions are drawn in regards to whether to accept or reject the four hypotheses defined in Chapter 5. These conclusions are outlined in Section 6.6.

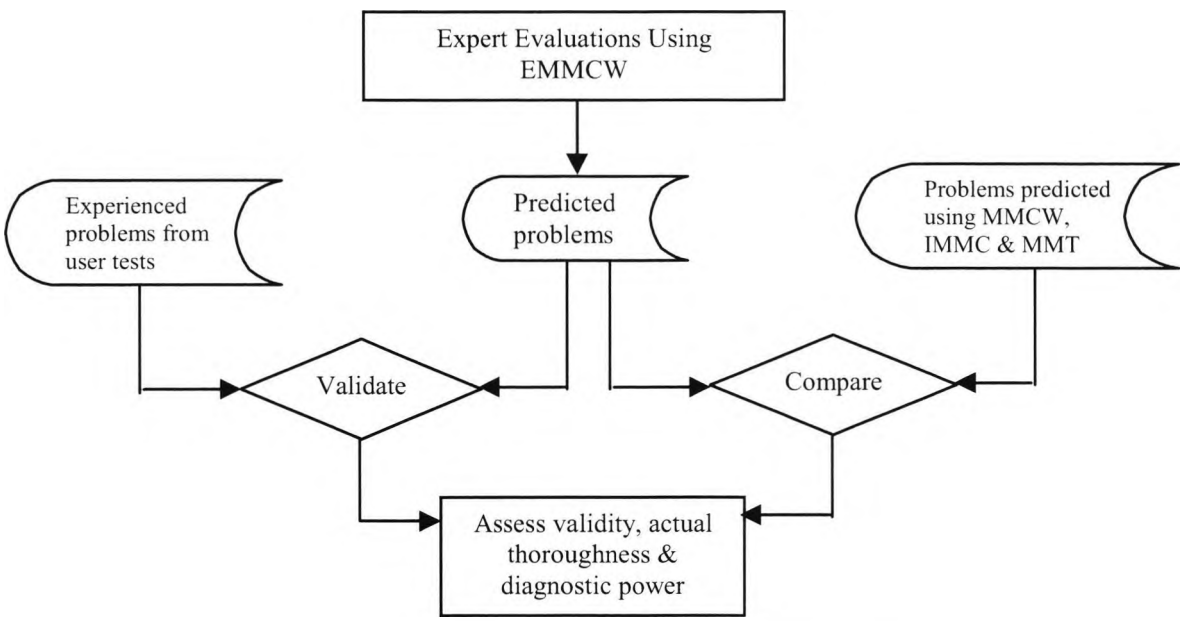


Figure 6.1: Method of Assessing the Effectiveness of the EMMCW

The methodological approach for validating predicted problems presented in Figure 6.1 differs from the one implemented in the previous study in that no instructional specialists were involved to validate the predicted problems (to complete Step 2 of the validation process described in Chapter 3). This was because it was not possible to find instructional specialists to validate the predicted problems at the time the study was completed. Therefore, to be able to compare the validity ratios of the EMMCW and the ratios of the existing EEMs studied, only the validity calculations from Step 1 were compared. The conclusions regarding the *validity* and *invalidity* of the predicted problems identified in this study represent the ability of the new EEM to predict valid

user problems in comparison to the three existing EEMs. However, the new EEM's ability to predict valid instructional problems, which the users did not encounter, could not be reflected on.

6.2 Creation of an Expert Evaluation Method for IMM

This section firstly explains the reasons why the original Multimedia Cognitive Walkthrough method was chosen to be the basis for developing an EEM for IMM. Then, it describes how the EMMCW was created to implement the conjectures made in the hypotheses defined in Chapter 5.

6.2.1 Rationale for Selecting the MMCW

The Multimedia Cognitive Walkthrough (Faraday and Sutcliffe, 1997) was considered to be the most appropriate of the three EEMs because it focuses on cognitive processing of information presented in multimedia presentations, while the other two EEMs do not consider such aspects. The main aim of Hypotheses 1 and 4 is to improve the prediction rate of cognitive implications of MMUI on users information processing in general, and their comprehension of the subject-matter in particular. Therefore, the MMCW was found to be a particularly suitable basis for testing these hypotheses, despite the method's unsatisfactory performance achieved during the empirical investigation reported in Chapter 4. The method's potential for predicting cognitive breakdowns and comprehension difficulties was considered to be the highest amongst the three EEMs, which was supported by the fact that it is the only method which enabled the prediction of valid attentional breakdowns and one valid comprehension problem.

6.2.2 Principles of the EMMCW Method

The most significant modification implemented in the EMMCW is a change of the approach to evaluation from predominantly prescriptive to *analytic*. The process of reviewing how well an MMUI supports a specific group of users in achieving their learning goals and inferring areas of potential difficulties is an analytic one. It requires evaluators to consider a number of contextually dependent factors, such as who the users are, what their prior domain knowledge is, as well as other aspects of the learning environment. The evaluators then need to predict what cognitive processes users are likely to engage in during their interaction with various MM resources, based on which

they need to make a judgement as to whether the users are likely to be hampered in their knowledge construction processes. As Fisher and Lipson (1986) emphasise, learning error detection and recognition is a process of judgement. Therefore, such an analysis asks for an approach to IMM evaluation that adequately supports experts in considering the complexity of the learning context, and in reasoning about how well the MMUI design is likely to support users in achieving their tasks. The experts then can infer the likely problems the target users may encounter in the learning process.

The original MMCW proposes that an attentional graph is created depicting all media resources included in a presentation, and then annotating the timing of presenting each medium, its duration and any contact points established between various media objects. The MMCW also contains a set of guidelines for Attentional Design, Media Integration Design and Media Selection. These guidelines are prescriptive in nature as they encapsulate good MMUI design principles and require evaluators to establish whether aspects of the MMUI design in consideration conform or not with these principles. A major limitation of these guidelines is that they do not encourage the consideration of the users' tasks, the users' background or other aspects of the context of use. The original MMCW method does not support context-dependent usability evaluation of MM presentations. Therefore, it was decided to adopt a different approach to evaluation which guides the analysis of the effectiveness of IMM by supporting experts in reasoning about how effectively the MMUI design assists a particular group of learners in achieving their learning goals. This can then encourage evaluators to assert their qualified judgement regarding the effectiveness of the design of IMM and to draw conclusions as to what potential problems users may encounter. In this way, the EMMCW promotes context-sensitive evaluation of IMM.

To achieve an analytic usability evaluation, it was proposed to adopt *analytic-style questions* that encourage the consideration of the learners, particularly their prior domain knowledge, and that support judgement regarding the effectiveness of IMM. The questions resemble in principle the four questions asked in the Wharton et al.'s (1994) Cognitive Walkthrough method regarding users' interaction with a UI in achieving individual tasks. Furthermore, as part of the Heuristic Walkthrough method, Sears (1997) also proposes four "thought-focusing" questions which aim to guide evaluators in concentrating on certain parts of the users' interaction while performing

their tasks. How the analytic questions, included in the EMMCW, were formulated is described in the next section.

The EMMCW method is intended to be used by two group of experts: firstly by multimedia and user interface designers, and secondly by instructional specialists and instructional designers. Although both groups of evaluators are expected to analyse the design of IMM applications and their cognitive implications for users, the first group is required to focus more on the design of the media resources and their integration, and the implications on users' cognitive processes. The second group is required to emphasise primarily on the instructional efficacy of the media resources and the implications of the content design for knowledge acquisition. As the professional background of these two distinct groups of evaluators is different, two versions of the method were created, each one specialised for one of the two groups. The expert specialisation is aimed firstly to ensure that the expertise of each type of evaluator is utilised adequately, and that they are asked to focus on issues within the realm of their expertise. This should reduce the potential of making false predictions due to lack of expertise, which is a threat to the validity of expert predictions in methods without expert specialisation, as highlighted in Chapter 3. Secondly, the expert specialisation is aimed to ensure that the method is understandable by each type of expert. Details of the expert specialisation are also given in the relevant sections below. The two versions of the EMMCW are given in Appendix 6.1.

The main element of the original MMCW which was adopted in the EMMCW was the cognitive framework of MM presentations, as it is considered fundamental for providing a basis for reasoning about the effectiveness of such user interfaces. Three major aspects, different to the original method, were implemented in the EMMCW method to satisfy the conditions of the four hypotheses defined in the previous chapter. Firstly, as mentioned above, analytic-style evaluation questions were formulated, based on the evaluation guidelines included in the original method. Secondly, two usability report formats were integrated in the method's procedure of evaluation and specification of usability problems. Finally, the evaluation procedure was amended to specify how the new evaluation criteria and the formats should be used for conducting expert evaluations of IMM. Each aspect of the EMMCW is described in detail below.

6.2.3 Formulation of Analytic Evaluation Questions

The framework of cognitive processes of MM presentations, proposed in the original MMCW, specifies three key processes: attention to, integration of and comprehension of information. Three main sections of evaluation guidelines are then proposed in the original MMCW: Evaluation of Attentional Design, Evaluation of Media Integration and Evaluation of Individual Media Design, which reflect how users' cognitive processes are affected by the design of different MM resources. The three main sections were preserved. However, at the beginning of each section more detailed explanation was composed, which highlights the main principles of how users interact with IMM and how their cognitive processes are affected during such interaction, as well as pointing out major areas of concerns. A set of questions were then formulated in each section guiding the analysis of different media resources, such as images, text & animation, and encouraging evaluators to reason about the implications of their design on users' cognitive processes of attention, perception, information integration and comprehension of concepts represented. In this way the main condition stipulated in Hypothesis 1 was implemented in the new method.

In total 46 evaluation questions were formulated to guide evaluators to consider the implications of specific components of the MMUI design on these cognitive processes while performing learning tasks. The questions are based on the design guidelines included in the MMCW. The relevant design guidelines are also included after each set of questions to provide evaluators with more explanation about basic principles of media design and integration.

The analytic questionnaire aims to support reasoning about the implications of the MM design on learners, which will allow them to infer the causes and effects of potential usability problems. This is achieved by dividing each evaluation question into two parts. The first part points evaluators to consider aspects of the MM design, which can potentially cause problems to the user. The second part of each question then encourages evaluators to identify what effect the MM design aspect may have on the user's cognitive processes. The latter aims to support the evaluators in identifying cognitive breakdowns that users may experience in their interaction with IMM, which implements the condition of Hypothesis 1, and also aims to improve methods' limitation 1, as discussed in Section 5.3. Crane and Isaacson (1990) also emphasise that design

effectiveness decisions benefit from knowledge about the existing competencies of the target audience and profit from responses of the target group to alternative representation formats and interactive strategies. For some questions, the second part is presented in a separate question indented underneath the original question. The following is an example of a set of questions for evaluating highlighting techniques for text. They are followed by a design principle based on empirical evidence of attention to text (Faraday, 1997), supporting not only clarity of the questions but also decision-making regarding the effectiveness of the MMUI design.

1.1 Are important words, phrases or concepts highlighted to ensure the user will pay particular attention to them?

1.2 Are there important words and concepts which are not highlighted at all?

1.3 If so, do you think that the user may fail to read them?

1.4 Are there any parts of the text which are highlighted but are not important for the user's task?

1.5 Is that likely to divert user's attention from more important parts in the text?

Design principle: In order to ensure that the user will extract all important information their attention needs to be drawn explicitly to key words and phrases. Effects which can be used to make particular words or phrases stand out include bold, large fonts or underlining.

The first question aims to assess the effectiveness of attentional design techniques used in the context of the user task, whereas the second question asks the evaluators to infer what implication these techniques may have on users' attention.

The evaluation questions formulated in the last section of the EMMCW: Evaluation of Individual Media Design, aim directly to support evaluators in reasoning whether the design of individual media supports or hinders learners in developing adequate understanding of the subject matter concepts represented. In the process evaluators are also explicitly asked to take into account learners' prior knowledge of the subject matter. Both characteristics promote the prediction of comprehension difficulties learners are likely to experience, therefore implementing the conditions set in

Hypothesis 4. Furthermore, this aims to improve the first part of limitation 3, as defined in Section 5.3. An example question designed to check whether textual descriptions promote understanding of the concepts explained is:

> 5.3 Are the concept descriptions at the right level for the students to understand, given their prior knowledge of the material?

Two versions of the evaluation questionnaire were created - one using terminology suitable for MM and UI designers, and the other using terminology familiar to instructional specialists. For example, the first version used terms such as 'users', 'visual' and 'verbal' media, 'contact points', 'user interface', whereas the second one used respectively - 'students', 'images' and 'text', 'links', 'presentations', as instructional experts may not be familiar with HCI terminology.

When formulating the evaluation questionnaire, care was taken not to bias the questions towards the Maths application. To ensure that, the questions were derived based on the guidelines included in the original MMCW, and no specific questions regarding particular aspects of the Maths application were formulated.

6.2.4 Introduction of Usability Report Formats

The second main aspect of the EMMCW developed was to integrate usability problem report formats into the approach. The uniform problem report formats firstly aim to ensure the completeness of the problem definitions, by explicitly encouraging evaluators to specify all necessary components of usability problems. Secondly, the problem report formats aim to ensure that the likely effect of MMUI designs on the users' cognitive processes, behaviour and performance is explicitly specified. This would encourage evaluators to consider whether design faults will have an effect on the users or not. This could reduce the likelihood of experts making false predictions as they may be able to rule out minor design issues which are unlikely to impact the users. Finally, for the purpose of this research, introducing a uniform problem report format, consistent with those in which the user problems were specified, aims to facilitate the comparison between the predicted and experienced problems necessary for establishing the effectiveness of the first. In all these respects, the formats aim to directly improve limitation 4 of the EEMs.

As highlighted above, the two target groups of evaluators are required to focus on different implications of the design of IMM on users while performing learning tasks. As multimedia and user interface designers are considered more capable of predicting user dialogue problems, they were asked to define usability problems specifying the following components:

- **Problem Description:** a general description of the usability problem.
- **Cause:** the cause of the problem, e.g. the design fault or feature likely to cause the problem, if it has not been specified in the problem description.
- **Effect:** the immediate effect on the users' cognitive processes of attention, perception, information integration or comprehension. It is equivalent to cognitive breakdown in the user interaction problem format. The term *effect* was used instead of *a cognitive breakdown* as the method is designed to be used by MM designers and other evaluators who may not have sound cognitive psychology knowledge.
- **Actions:** the actions users are likely to take as a result of the cognitive breakdown. It is equivalent to behavioural outcome in the user interaction problem format. Similarly to the previous component, the use of the term *behavioural outcome* was not used to avoid confusion.
- **Performance Outcome:** the effects on the user's work and the achievement of their tasks resulting from the problem.

As instructional designers and instructional experts are expected to focus primarily on the comprehensibility of the MM presentation, they were asked to specify the following four components:

- **Problem Description:** a general description of the usability problem.
- **Cause:** the cause of the problem, e.g. the design fault or feature likely to cause the problem, if it has not been specified in the problem description.
- **Comprehension Difficulty:** a description of the potential problem learners may experience understanding the subject matter, including what part of it they are likely to have difficulties comprehending. The term *comprehension difficulty* is used instead of *misconception* or *comprehension outcome*. This is because the aim of Hypothesis 4 is to test whether evaluators can systematically predict

areas of the IMM presentation which can hamper learners' understanding, rather than their prediction rate of specific misconceptions or how students behaviour or performance may be affected as a result of not forming complete understanding of the content. This was highlighted in Section 5.6.

This format does not include behavioural outcomes, as instructional specialists do not typically have user interface design expertise, and therefore they may not be in a position to specify the likely actions a user may take following a comprehension difficulty.

6.2.5 Modification of the Evaluation Procedure

The evaluation procedure was modified to support the analytic nature of the new evaluation method. The suggested procedure is depicted in Figure 6.2. It is proposed that MMUIs are analysed segment by segment. The granularity of a segment needs to be determined individually by the evaluators depending on the structure of the presentation being evaluated.

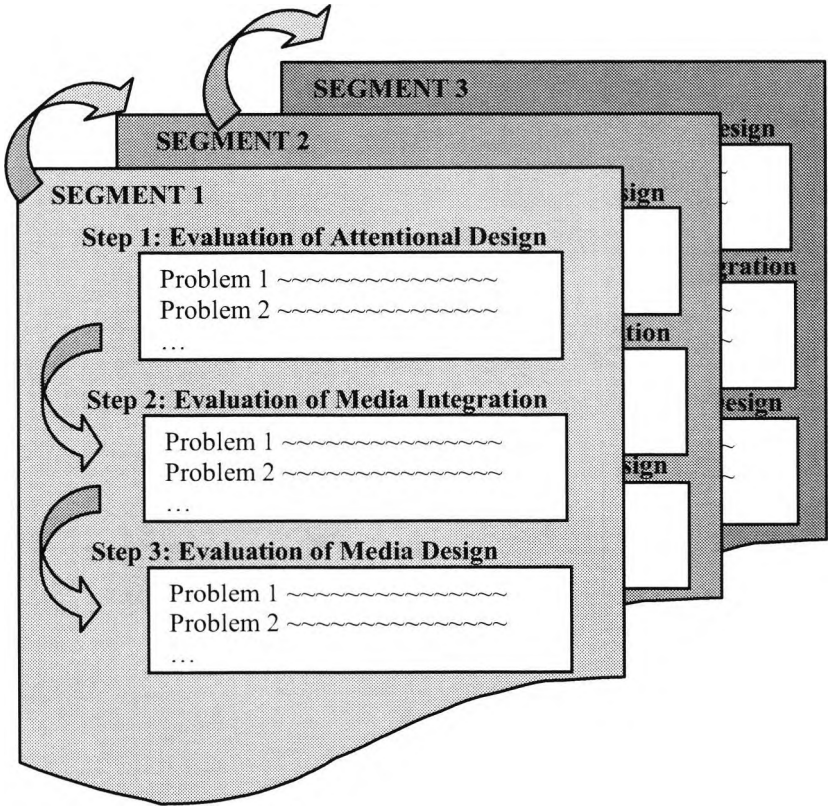


Figure 6.2: Analytic Evaluation Procedure Proposed in the EMMCW

The three steps of evaluation proposed in the original method were preserved. They are:

1. Evaluation of the Attentional Design,
2. Evaluation of the Media Integration,
3. Evaluation of Individual Media Design.

The creation of attentional graphs, however, was excluded from the new procedure, as it proved unsuitable for interactive presentations, as revealed by the empirical study reported in Chapter 4. Instead, it is suggested that evaluators analyse each segment of the MMUI following the three steps before moving to the next segment. In each step, evaluators need to use those questions which they consider relevant to the application and infer potential user problems. It is important to emphasise that evaluators were asked to specify usability problems resulting from their analysis prompted by the evaluation questions, rather than simply provide answers to each question they consider relevant. Finally, when specifying each usability problem, evaluators are asked to follow the uniform problem report format.

Figure 6.3 summarises the changes made in the original MMCW and the new aspects introduced in the two versions of the EMMCW developed.

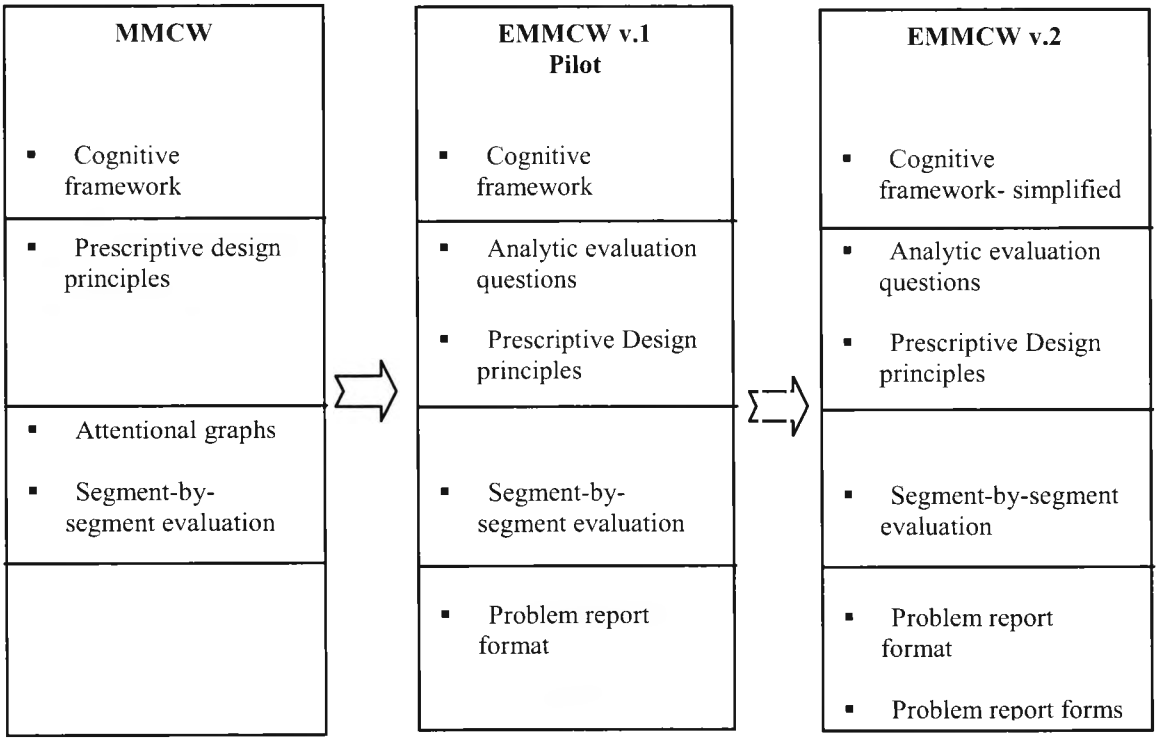


Figure 6.3: Comparison of Original MMCW and the EMMCW

6.3 Approach of EEM Validation Study

This section presents the methodological background of the empirical validation study by describing the pilot study conducted, the experts who took part in the main study, its procedure and the post-evaluation interviews conducted.

6.3.1 Pilot Expert Evaluation

A pilot evaluation was performed to test how clear and understandable are the evaluation procedure and the evaluation questionnaire included in the EMMCW. The aim was to gather procedural and ease-of-use information rather than to assess the validity of the evaluation results produced.

For the purpose, one female PhD student from the Centre for Human-Computer Interaction Design at City University was asked to evaluate the *Exponential Graphs Section* of the Maths software using the EMMCW. The evaluator had some HCI training and one year of undergraduate project experience in usability evaluation of IMM software. She familiarised herself with the method in her own time. After that, the evaluator was instructed to assess the usability of the Maths application using the EMMCW method. She was firstly given a context description (see Appendix 6.2) describing the target learner group and the context in which the application is typically used. The same information was provided to the evaluators who used the original MMCW. The evaluation then progressed in two stages:

- **Stage 1:** The evaluator was asked to specify only the causes and the likely *cognitive breakdowns* resulting from the usability problems. This is aimed to test Hypothesis 1.

- **Stage 2:** A detailed user profile was created, based on discussions with the creators of the Maths software and a lecturer who used the software as part of his teaching method. The description specifies the user's age range, prior knowledge in the Maths material taught and relevant concepts, computer literacy, English language abilities, as well as the context in which the software is typically used for learning Maths. The user profile can be found in Appendix 6.3. It was created to fulfil the second condition in Hypotheses 2 and 3.

During the second stage of the pilot the evaluator was given the detailed user description and was asked to specify the likely *behavioural* and *performance outcomes* for the problems already identified. This is aimed to test Hypotheses 2 and 3. The reason for the separation was to exclude the influence of the detailed user specification on the definition of the likely *cognitive breakdowns*, as providing such is not a condition in Hypothesis 1.

At the end of the evaluation session, a short interview was administered to elicit the evaluator's opinion on the comprehensibility of the cognitive framework, the clarity and ease of applying the evaluation procedure and the evaluation questionnaire.

The procedure of evaluation using the EMMCW was studied in the pilot, as well as the problem report format and a high level view of the predicted problems rather than their content. The predicted problem set, therefore, was not compared to the user test data.

During the pilot, it was found that although in the *Guidelines for Evaluation* section it is explicitly specified that evaluators need to identify problems which can potentially hinder users' cognitive processes, the evaluator answered the questions in turn providing positive and negative comments about the interface design. This result indicates that either the procedure was not sufficiently clearly defined or could be due to the evaluator's inexperience in performing usability evaluation. Furthermore, the evaluator inconsistently specified the causes of the problems identified and the likely cognitive breakdowns although explicitly asked to specify them for each problem in the *Guidelines for Evaluation*. None of the problem descriptions also specified the location of the problems, which can further hinder the problem rectification process.

Based on these results, it was decided to amend the procedure of evaluation to highlight further to evaluators that the aim of the evaluation is to analyse the user interface and infer likely usability problems rather than to provide answers to the relevant evaluation questions. Blank forms for specifying usability problems following the format already defined were also introduced to ensure that all components of usability problems are specified, including their cause, likely cognitive breakdown, behavioural and performance outcomes, together with the location of the problem. An example form is given in Appendix 6.4.

During the post-evaluation interview the evaluator expressed that the cognitive framework provided was "reasonably clear" and that the concepts of contact points between visual and verbal media were not easily understandable and she had to re-read their explanation a couple of times. Therefore, Faraday and Sutcliffe's (1997) cognitive framework was modified to exclude Short-Term, Working and Long-Term memory information processing concepts, with which multimedia designers and instructional specialists are not typically familiar. The new description of the framework explains the way people learn with multimedia, emphasising on the three fundamental processes: attention to different media resources, integration of information presented in visual and verbal media, and finally comprehension of the information or meaning formation. The simplified framework is aimed to be more understandable to evaluators with little or no knowledge of cognitive science, as typical multimedia and instructional designers tend not to have such training (Dimitrova and Sutcliffe, 1999). Furthermore, a greater emphasis was placed on how media resources support the processes of learning and comprehension, as it was considered that the original framework does not sufficiently highlight the last cognitive stage of comprehension of information presented in MM presentations.

Providing the detailed user description after the problems have been identified did not result in specifying many behavioural or performance outcomes. In fact, the evaluator did not specify any behavioural outcomes and identified performance outcomes to only 10% of the problems predicted. One possible explanation could be that she may have had difficulty remembering the specific nature of the problem and the actions that followed after that during the second stage of the evaluation. Furthermore, the evaluator's analytic reasoning processes about each problem were split in two parts - firstly inferring likely breakdowns to all problems and later revisiting each problem to specify likely behavioural and performance outcomes. This may have hampered the evaluator in fully identifying all implications of the same problem, as this was not done contiguously. Therefore, it was decided to provide evaluators with the detailed user description at the beginning of the evaluation process.

Finally, during the interview the evaluator also explained that although most of the questions could be applied to the Maths application, she found the section regarding

Attention to Moving Image irrelevant. Therefore, this section was removed from the questionnaire.

6.3.2 The Experts

As explained earlier, the EMMCW method is aimed to be used by two types of evaluators - user interface experts (including multimedia designers and HCI experts) and instructional specialists (including SMEs). Therefore, four evaluators were recruited to apply the EMMCW, different from the ones who participated in the first study. Two of them (both female) were MSc students in Electronic Publishing who had previous training in a range of usability inspection methods, and both have used Heuristic Evaluation, Checklists and Style Guidelines for evaluating software. Both had been involved in designing multimedia applications for one year, one of whom also had experience evaluating IMM on 10 projects of various size. The other two evaluators were PhD students both of who had knowledge of the Maths material and were tutoring on a Maths course within the School of Engineering at City University at the time of the study. One of them had also been teaching Maths formally or informally for 10 years, whereas the other had limited experience in teaching. The latter however, had five years of software engineering experience.

None of the evaluators had any prior knowledge of the original MMCW or either of the other two methods used in the previous study. Prior to the experiment, each evaluator familiarised themselves with the EMMCW in their own time.

Finally, none of the evaluators was familiar with the Maths application before the experiment. Therefore, the last two conditions were kept uniform with the previous evaluation study.

6.3.3 Procedure of Expert Evaluations

As none of the evaluators had previous experience with the Maths software, they were given 15 minutes to familiarise themselves with it. The evaluators were also given a context description (see Appendix 6.2), the detailed user description, and a task instructions sheet (see Appendix 6.5), to familiarise themselves with the goals of the application, who is likely to use it and in what environment. After that, each evaluator was asked to apply the EMMCW to evaluate the *Exponential Graphs section* of the

Maths application in order to uncover potential usability problems. The evaluators were asked to follow the evaluation procedure described in the method. The MMDs were also given problem report forms (see Appendix 6.4.1) in which to specify the usability problems they identify, describing the problem first, then specifying its cause, location, the breakdown resulting from the problem, and the likely behavioural and performance outcomes. The SMEs were given another report form (see Appendix 6.4.2) which allows them to specify any comprehension difficulties they identify, giving a description of the difficulty, its cause and the nature of the misconception learners are likely to form. Both groups were also asked to write down the evaluation criterion that led them to identify the error. The latter was necessary to differentiate which errors were identified using the EEM and which using evaluators' own judgement.

Each evaluator worked individually in the same office, under the same conditions to those of the previous study. Thus, environmental factors were kept the same for all usability evaluations.

6.3.4 Post Evaluation Interview

To be able to assess the ease-of-use of the EMMCW and to gather information about how it was applied by each evaluator, semi-structured interviews were conducted immediately after each evaluation session. The aim was also to collect data about the evaluators' experience and attitude towards key aspects of the EEM. As the purpose of the validation study was not to assess the effectiveness of the EEM according to every effectiveness criterion specified in Chapter 3 but to test its predictive power, a shorter questionnaire was created, which is given in Appendix 6.6. It contains multi-scale questions aimed at eliciting specific information regarding how understandable the cognitive framework was, how easy-to-use was the structured problem report format, and how comprehensible the evaluation questionnaire was. The interview questionnaire also contains open-ended questions aimed at probing evaluators to describe their experience using the EEM and to provide any suggestions for improving the EEM.

At the end, the evaluators were given a questionnaire (the same as those in Appendix 4.2), which gathered information on their prior knowledge of and experience with evaluation methods, as well as any relevant qualifications and experience they have.

The same questionnaires were given to the evaluators who participated in the previous study.

6.3.5 Method of Analysis of Expert Evaluation Data

6.3.5.1 Refining of the Predicted Problem Set

The original problem set predicted by the evaluators was refined to exclude any repetitions made by individual evaluators and also any bogus problems (Cockton and Woolrych, 2001) if such have been identified by the evaluators. In the same way as in the previous study, a distinction was made between problems that were identified as a result of applying each EEM and those that were identified due to the evaluators' own expertise. The distinction between the two was made based on the information provided in the problem report forms, where the evaluators explicitly indicated which evaluation question prompted them to identify each problem. The problem descriptions (PDs) for which no such information was given were assumed to be based solely on the evaluators' expertise. The number of EEM-derived problems was then counted.

6.3.5.2 Validation of the Predicted Problems

To establish whether the problems predicted using the EMMCW are valid or not, each component of the predicted PDs (as defined in Appendixes 5.1 and 5.2) was compared to each part of the user interaction and comprehension problems from the user tests already conducted as redefined using the problem report format (see Appendixes 5.1 and 5.2). A predicted problem was classified as valid if any of its components was successfully matched to any component of the user problem data. The six matching rules used to compare the predicted and experienced problems in the previous study were used to match the problems predicted using the EMMCW. As the user and the expert problems were specified using the same problem report format, the scale of degree of certainty of problem matches was not used in this study. Independent instructional specialists were not available to perform Step 2 of the problem validation (as defined in Chapter 3) in this study, therefore second validation of instructional problems could not be performed. Thus, the validity of problems related to the Maths content which were not experienced by the users could not be established with certainty.

As a result of the validation, the predicted problems were divided into:

- Problems experienced by the users. These problems are valid ones.
- Content-related issues not experienced by the users. As these problems could not be ratified by instructional experts, their validity or falsity cannot be established with certainty.
- Design faults not experienced by the users. As in the previous study, this set of problems will be classified as false alarms, as no user test data exists to support them.
- False alarms, which contradicted or were explicitly rejected by the user test data.

The results from the validation are presented in Section 6.4.

6.3.5.3 Detailed Problem Matching

This step was introduced in this study in order to establish the prediction rate of each of the components specified in the four hypotheses, formulated in Chapter 5. In this way, any improvement in predicting cognitive breakdowns, behavioural outcomes, performance outcomes, and comprehension difficulties could be measured. To achieve that, every component of the problems predicted using the EMMCW (as defined in Appendixes 5.1 and 5.2) was compared again to every component in the user problem descriptions using the matching rules specified in Chapter 4. Two rating scales were also used. The first was used for matching the problems predicted by the MMDs to the user interaction problems. This scale ranged from 0 (no components matching) to 4 (all four components matching). The second scale was applied for matching the problems predicted by the SMEs and the comprehension problems experienced by the users. This scale ranged from 0 (no components matching) to 3 (the cause, comprehension breakdown and comprehension difficulty matching). A similar rating scale was suggested and used by Lavery, Cockton and Atkinson (1997) for matching expert and user problems.

As the user interaction and the comprehension problems were specified using different report formats, the prediction rate results for each one will be presented separately in the next section. After that, for the user interaction problems identified by the MMDs, the prediction rate of cognitive breakdowns, behavioural outcome and performance outcomes will be discussed individually. For each component, the number of correctly

predicted instances, the number of instances predicted but not experienced by the users, and the number of instances for which the component was not specified will be established. After that, the likely reasons for the successful or unsuccessful prediction of each component will be discussed.

Similarly, for the comprehension problems identified by the SMEs, the prediction rate of misconceptions and comprehension problems which were experienced by the users and those which were not experienced will be discussed, together with other components predicted by the SMEs.

6.3.5.4 Comparison of the Results to those from the MMCW

The results produced using the EMMCW method will be compared to those produced using the Multimedia Cognitive Walkthrough method to identify whether improvement has been observed in terms of increasing the number of valid problems predicted and decreasing the number of false alarms identified. Only the problems predicted by the MMDs using the MMCW and the EMMCW will be compared. This is because the original MMCW was applied only by a multimedia designer and an HCI expert, and not by SMEs. Therefore, to preserve the internal validity of the conclusions reached the problem set produced by the SMEs will be excluded from the comparison.

Both sets of results will be compared based on their *validity*, *actual thoroughness*, and *diagnostic power*, which are the effectiveness criteria aimed to be improved by the hypotheses formulated in Chapter 5.

6.3.5.5 Comparison of the Results to those from all EEMs Studied

To establish whether overall enhancement of the effectiveness of the EMMCW method has been achieved, the results achieved using the new EEM will be compared to those achieved by all three existing EEMs studied in this thesis. All four EEMs will be compared based on the three criteria investigated in the hypotheses, namely *validity*, *diagnostic power* and *actual thoroughness* of predictions made. The compared data will include the results produced by both MMDs and SMEs who applied the EMMCW.

If an improvement in these measures is observed, then the EMMCW can be considered more effective than the other three methods. The four hypotheses can then be

considered as accepted based on the evidence produced in the comparison. Alternatively, the hypotheses will be rejected.

6.4 Results from the Expert Evaluations using the EMMCW

This section first presents the results from the expert evaluations in terms of the total number of problems predicted and how thoroughly they were specified. Then, the predicted problems are validated against the user problem set, followed by a detailed analysis of the predicted cognitive breakdowns, behavioural outcomes, performance outcomes, and comprehension difficulties.

6.4.1 Number of Problems Predicted

In total 66 non-unique problems were predicted by the four evaluators. Table 6.1 presents the results produced by individual evaluators, as well as by each type of evaluator - the multimedia designers (MMDs) and the Maths tutors acting as subject matter experts (SMEs). The problem sets were refined to exclude any duplications made by the evaluators. For example, the two MMDs identified four common problems, therefore in total they predicted 22 unique usability problems. The Maths tutors identified two duplicating problems, therefore the total number of unique problems they predicted is 38. Finally, both types of evaluators identified two common problems. No bogus problems were identified in this study. The final total of unique problems predicted using the EMMCW is 58, as indicated at the bottom of Table 6.1. A complete list of the problems predicted using the EMMCW can be found in Appendixes 6.7.1 (a list of problems predicted by the MMDs) and 6.7.2 (a list of problems predicted by the SMEs).

Table 6.1: Problems Predicted by Evaluators using the EMMCW

Expert type	Total # predicted by expert	# predicted using EEM	% predicted using EEM	# from own judgement	% from own judgement
MMD1	15	14	93%	1	7%
MMD2	11	10	91%	1	9%
# of common problems	**4**	**4**			
Total unique by MMDs:	**22**	**21**	**95%**	**1**	**5%**
SME1	23	23	100%	0	0%
SME2	17	16	94%	1	6%
# of common problems	**2**	**2**			
Total unique by SMEs:	**38**	**37**	**97%**	**1**	**3%**

Total Unique Predicted Using EMMCW:	**58**

An interesting observation is that the evaluators only specified one problem each using solely their own judgement. This could indicate either that they could relate nearly all of the problems they identified to the evaluation criteria provided, or that they limited their predictions primarily to problems prompted by the evaluation questions. The latter is supported by a comment made by the second MMD during the interview that "you could get carried away by specifying many problem even very little ones, but I didn't feel that was necessary". This indicates that this evaluator prioritised the problems and specified only those she believed were important.

6.4.2 Validation of Predicted Problems

The results from the problem validation are presented in Figure 6.4. The last column in Appendixes 6.7.1 and 6.7.2 indicates for each valid problem predicted using the EMMCW which user interaction or comprehension problem it has been matched to.

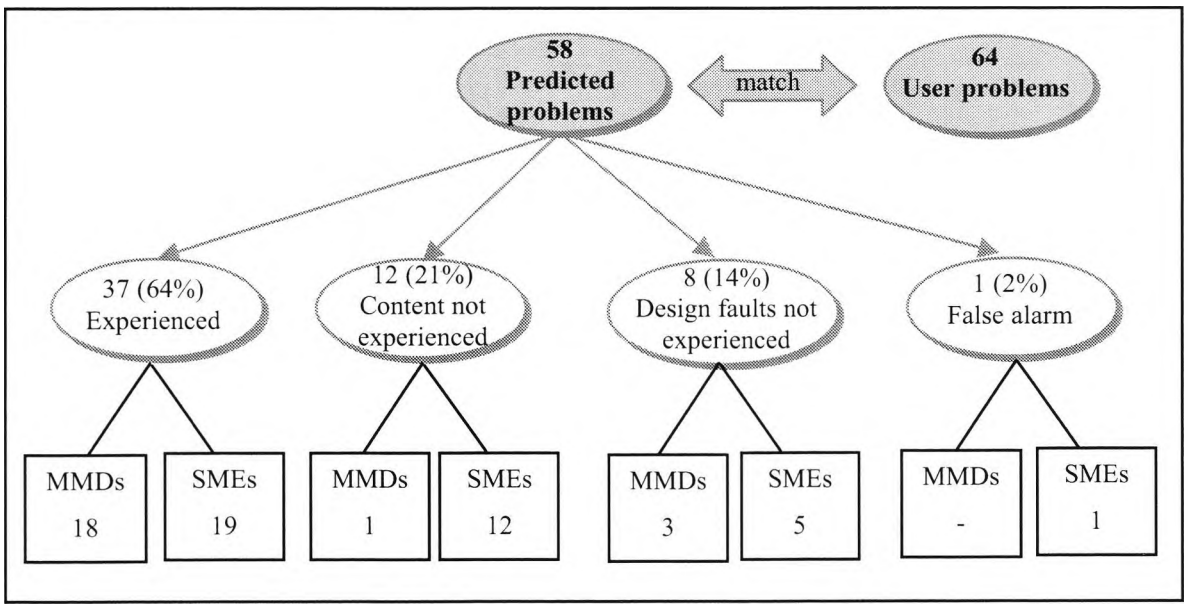


Figure 6.4: Validation of Problems Predicted Using the EMMCW

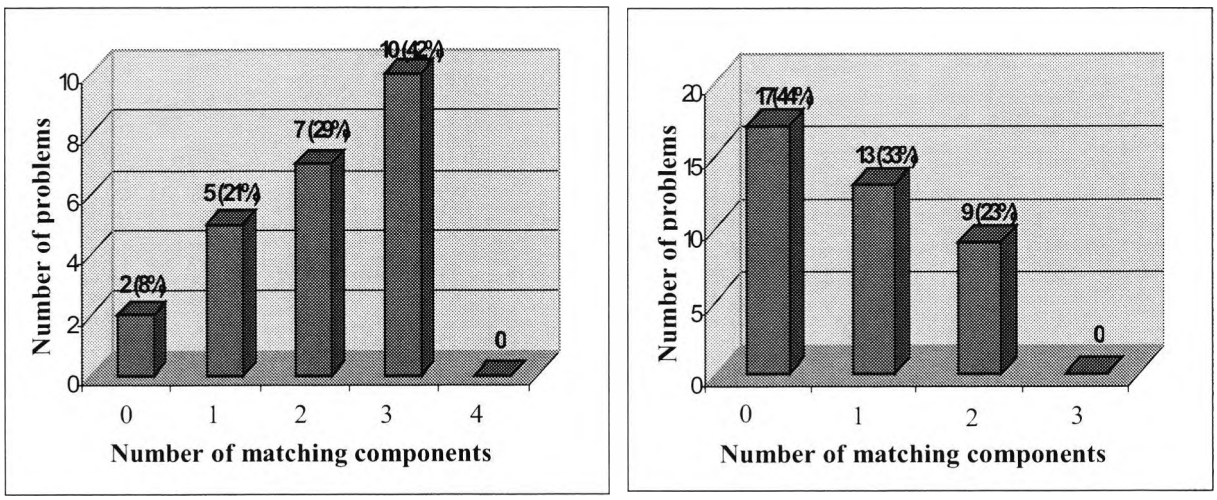
Figure 6.4 shows that nearly two thirds of the predicted problems were experienced by the users either during their interaction with the IMM or during the knowledge test following that. It needs to be acknowledged that as only four users took part in the user tests, the user problem set may not be complete. Therefore, it is likely that more predicted problems could have been validated should more users have been involved. 21% of the predicted problems are concerned with issues of the Maths content and the way it is represented, however they could not be validated with certainty as no independent judges were available to do so. None of these problems contradicts the user test data, and thus there is a potential that some of these issues are valid ones. Therefore, the proportion of valid problems predicted using the EMMCW could be higher than 64%. The evaluators also identified some design faults, which were not experienced by the users. Finally, one problem predicted by one of the SMEs contradicted the user test data and was classified as a false alarm. This false alarm is an interaction problem which was specified by a Maths tutor, who did not have enough expertise in UI design to specify such problems with accuracy.

6.4.2.1 Detailed Problem Matching

The above information only presents a high level view of the total number of valid problems predicted using the EMMCW. It does not reveal how many of the predicted

cognitive breakdowns, behavioural outcomes, performance outcomes, and comprehension difficulties were experienced by the users.

The results of the detailed problem matching are presented in Figure 6.5.



a) For User Interaction Problems predicted by MMDs

b) For Comprehension Difficulties predicted by SMEs

Figure 6.5: Problem Matching Results

All calculations are based on the total number of non-unique problems predicted by the MMDs and the SMEs using the EEM, given in column 3 in Table 6.1, as each evaluator specified different components although sometimes referring to the same problem.

6.4.2.2 Validation of User Interaction Problems

In almost all cases, the problems predicted by the MMDs correctly specify a cognitive breakdown, a behavioural outcome or a performance outcome, or a combination of the three. However, as can be seen from Figure 6.5 (a), there were no thorough matches between the problems predicted by the MMDs and those experienced by the users. Nearly half of the predicted problems (or 42%) correctly specify three of the four components used to describe the user difficulties. A further 29% of the user interaction problems predicted identified correctly two components of the user problems. Finally, 21% of them specified only one type of impact on users, which usually is the performance outcome of a user comprehension problem. It needs to be noted that the user comprehension difficulties were only described by an assumed cause, hypothetical

misconception and a comprehension outcome observed on the knowledge tests, and no behavioural outcomes could be described. Therefore, only three components of the user comprehension difficulties were matched to interaction problems predicted by the MMDs, the first two of which could have been inaccurate. Thus, the matched user interaction problems have successfully identified potential comprehension difficulties, although could not be classified as thoroughly matching them.

6.4.2.3 Validation of Comprehension Difficulties

Three components were matched from the problems predicted by the SMEs, namely problem causes, cognitive breakdowns and misconceptions. Three instead of two components were matched as the SMEs sometimes specified the likely cognitive breakdown resulting from a problem although they were not asked to do so in the problem format. Figure 6.5(b) reveals that these evaluators did not predict completely any of the problems experienced by the users. For all problems that were successfully matched the evaluators accurately predicted their causes. In 8 cases they also predicted cognitive breakdowns which may occur as a result of the design problem, 6 of which are valid content comprehension breakdowns, but they also specified 1 valid attentional and 1 valid goal execution breakdowns. The comprehension breakdowns indicate areas where learners may experience problems understanding the material, however as they do not specify precisely what concept could be misunderstood they were not classified as misconceptions. Finally, one of the SMEs identified 1 misconception which learners can form as a result of the design. Therefore, the SMEs were quite successful at predicting valid causes to comprehension difficulties, but not at specifying the misconceptions which can result from them. The potential reasons for this are discussed in Section 6.4.6.

In the following sections a more thorough analysis of the prediction rate of the four components tested in the hypotheses is presented, together with a discussion of whether they were experienced, not experienced by the users or contradicted their experience. The sections below do not discuss matches of problem causes due to two reasons. Firstly, the prediction of causes is not the object of study in any of the hypotheses as the prediction of the impact of MMUI design on users is the main focus of all four hypotheses. Secondly, discrepancies were discovered between some of the causes of the predicted problems to those of the experienced ones.

6.4.3 Prediction of Cognitive Breakdowns

Figure 6.6 presents the number of correctly predicted, predicted but not experienced by users and not specified cognitive breakdowns.

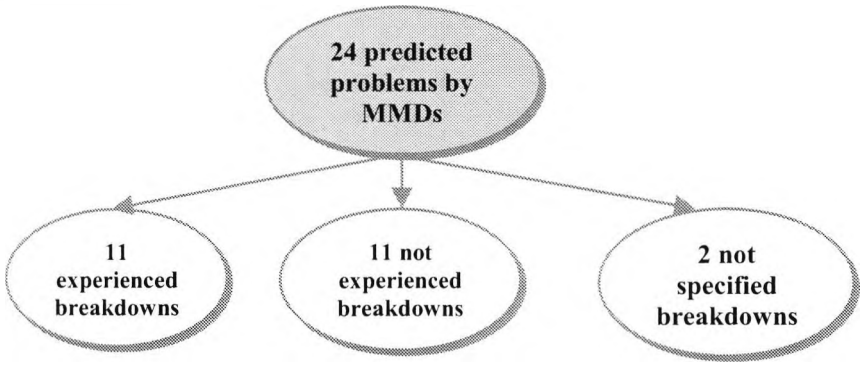


Figure 6.6: Distribution of Cognitive Breakdowns

It was found that exactly half of the cognitive breakdowns predicted by the MMDs while using the EMMCW are accurate. The successfully predicted ones are of four categories: comprehension (8 such were predicted), affective (1 problem causing frustration was identified), attention (1 was predicted) and action feedback (1 was predicted). Furthermore, as mentioned above, the SMEs predicted 8 valid cognitive breakdowns although they were not expected to do so, 6 of which are valid content comprehension ones, 1 deals with attentional design issues, and 1 is a valid goal execution breakdown. These results reveal that a notable number of content comprehension breakdowns were predicted, which is a positive improvement over the performance of the three EEMs studied in the previous study, with the aid of which none such were predicted.

One surprising result is that only two valid attentional breakdowns and no perceptual or information integration ones were identified. In a study examining the effect of introducing detailed task descriptions to guide evaluations using the Cognitive Walkthrough method, Sears and Hess (1999) observed a similar result. The authors found that evaluators identified significantly fewer problems related to difficulties locating information important for achieving the user tasks. In this study the limited number of predicted breakdowns referring to early stages of cognitive information processing (attention to and perception of information) could be due to the fact that more importance is given to the later stage of comprehension in the EMMCW method in comparison to the original MMCW. Therefore, evaluators specified more

comprehension related problems (breakdowns) occurring in the later stages of learning with IMM, instead of only considering the initial stages of attention and perception. This conclusion is supported by the fact that 14 valid comprehension breakdowns were predicted using the EMMCW, whereas none such were identified using the original MMCW. As developing correct understanding of the subject matter is a primary goal of IMM software, identifying hindrance in comprehension processes is of utmost importance and the results revealed in this section suggest that with the aid of the EMMCW a considerable amount of these problems can be predicted.

Another possible explanation is that it is not explained sufficiently in the method description and instructions that the second question of each pair of questions aids the identification of breakdowns.

In order to identify whether the evaluation criteria directly influenced the prediction of the breakdowns, the PDs were analysed together with question numbers which prompted the identification of the problems as specified by the evaluators. It was found that only in half of the cases the identification of breakdowns was attributed to the second part of question pairs, which was designed to prompt evaluators to consider the implications of design aspects on users' cognitive processes. There are two possible explanations of this phenomenon. Firstly, the evaluators may not have specified in their problem reports that the subsidiary question actually promoted the identification of breakdowns they specified, as they may have thought this is not necessary. In that case more breakdowns could have actually been attributed to the evaluation questions, although this was not explicitly indicated by the evaluators. Secondly, it may not have been clear to the evaluators that the second question of each pair directly prompts them to identify the likely effect design features may have on users' cognitive processes. This is briefly explained in the *Evaluation Method Overview* section of the EMMCW, however it is not emphasised explicitly in the specified procedure for evaluation.

As can be seen from Figure 6.6, 11 other cognitive breakdowns were predicted by the MMDs but could not be matched to those experienced by the users. Similarly, four more attentional breakdowns could not be validated by the user test data. There are two possible explanations of this result. Firstly, it could be due to the fact that it was not possible to specify some of the user cognitive breakdowns with accuracy when

reviewing the videotapes containing users interactions and verbal protocols. As eye-tracking data could not be collected, some breakdowns were inferred based on the data captured on the videotapes with a degree of uncertainty. Therefore, some of the cognitive breakdowns predicted by the evaluators may be valid ones, but there was not sufficient user data collected to validate them. A second reason could be that none of the four evaluators actually had any formal training in cognitive psychology, and perhaps the evaluation questions do not sufficiently support such evaluators in inferring cognitive implications of MM design on users. And as a consequence some of the cognitive breakdowns predicted may be inaccurate.

6.4.4 Prediction of Behavioural Outcomes

Figure 6.7 presents the number of correctly predicted, predicted but not experienced by users and not specified behavioural outcomes.

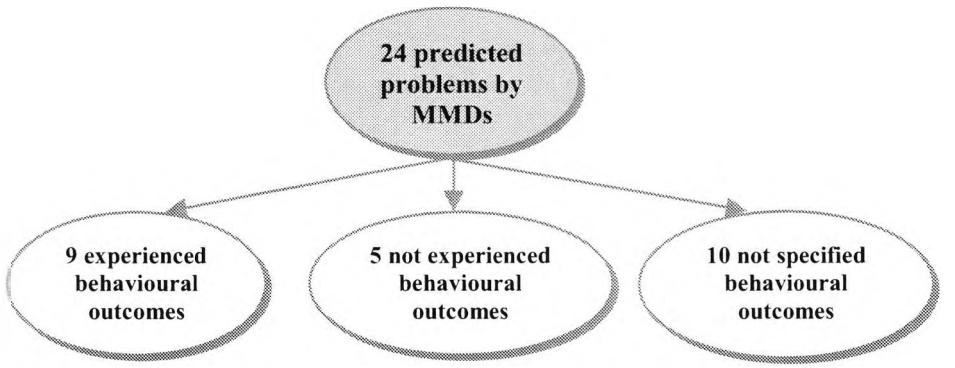


Figure 6.7: Distribution of Behavioural Outcomes

The MMDs using the EMMCW method correctly specified 9 behavioural outcomes. All of them describe repeated actions which users may perform as a result of experiencing a media interaction or a comprehension breakdown, which is the second of five types described in Chapter 5. None of the other types of behavioural outcomes were predicted in the study. The repeated action outcomes appear most intuitive to detect and specify compared to the other types, as their prediction requires considering the dialogue with the UI and the available choices to users. The prediction of the other four types seems more complex. For example, to predict types 1 and 5 (users performing wrong actions and users stop performing their task respectively), the evaluator needs to have a considerable understanding of the user tasks. Additionally, to predict type 3 breakdown (user skips important information) detailed knowledge of the content and which concepts are important is required, together with accurate prediction of attentional

breakdowns. Furthermore, the severity of the attentional breakdowns will determine the nature of the behavioural outcome, however the EEM does not clearly explain how both are related.

In a further two cases, the evaluator combined their description of likely user behaviour with likely performance outcomes resulting from such behaviour. This shows the close connection between the two components.

From Figure 6.7 it can also be seen that 5 of the behavioural outcomes predicted could not be matched to the actual actions taken by the users. One of these cases is a content related problem, which could not be matched with the user test data, thus this outcome could potentially be valid. In further two cases the evaluator specified the performance rather than the behavioural outcomes. In the final case the evaluator specified a breakdown rather than the user actions likely to result from it. The specification of performance instead of behavioural outcomes could be due to the fact that both behavioural and performance outcomes are tightly connected and the difference between the two may not be well explained in the EEM. Such similarity was observed by Lavery and Cockton (1997) as a result of which the authors combined the two types of outcome in one. Furthermore, the heading in the problem format, prompting evaluators to specify behaviour actions, was "actions" which requires observable actions to be specified. However, one of the MMDs may have misunderstood that as she specified a mental activity in the form of a breakdown in this field. Finally, all these behavioural outcomes were defined by the less experienced MMD, therefore her lack of usability evaluation experience could have contributed to the wrongly specified outcomes.

As explained in the method's evaluation procedure, not all usability problems result in users taking observable actions, which could explain why the evaluators did not define such for ten problems they identified. These problems are concerned with content comprehension, media integration or attentional issues, and therefore provoke cognitive actions rather than observable physical ones.

6.4.5 Prediction of Performance Outcomes

Figure 6.8 presents the number of correctly predicted, predicted but not experienced by users and not specified performance outcomes.

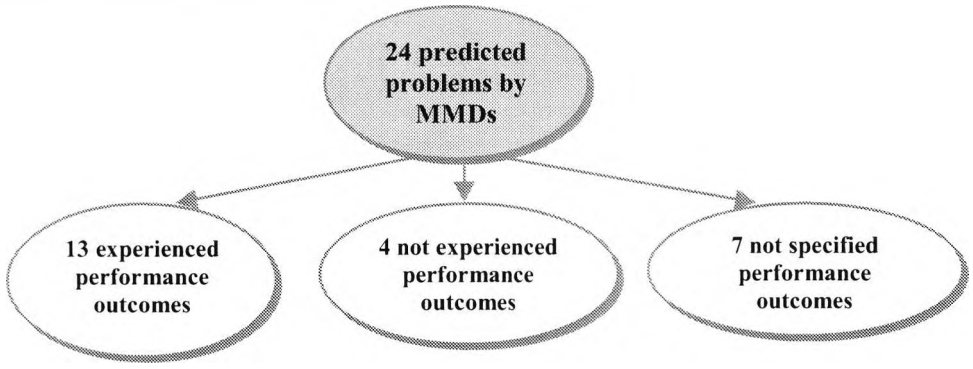


Figure 6.8: Distribution of Performance Outcomes

As can be seen from Figure 6.8 the two MMDs correctly specified 13 performance outcomes using the EMMCW. They fall into two categories: increased time to perform user's task and content comprehension, which refer to performance outcome types 3 and 5, as described in Chapter 5. In particular, the evaluators predicted that in 8 cases it could take longer for the users to perform and possibly achieve their goal. They also predicted that in 6 cases as a result of a usability problem users may experience difficulties understanding the Maths concepts, where in one problem both performance outcomes were predicted to be possible. The prediction of the first type is connected to the type of behavioural outcome predicted by the evaluators, i.e. if the users repeat an action or a series of actions in an attempt to achieve their goal, they are spending more time on the tasks which may or may not result in the achievement of their goal. The connection between observable actions and resulting performance outcomes was highlighted both in the EEM's *Specifying Usability Problems* section and in the instructions given to the evaluators, which contributed to the successful prediction of the 8 performance outcomes. The prediction of the 6 outcomes on users' comprehension of the subject matter could be attributed to the emphasis of the comprehensibility of multimedia presentations in the EEM, particularly of the last section: *Evaluation of Individual Media Design*. Although the comprehension outcomes specified by the MMDs were of a general nature, suggesting that a design feature may slow down comprehension, they may provide an indication to IMM developers that the problems specified can hamper understanding and the achievement of the learning goals. Such

information can then help to prioritise these problems as severe and that they need to be addressed during redesign stages.

Figure 6.8 also shows that 4 performance outcomes could not be matched to the user test data. In two cases these are content problems which could not be validated by the user problems, however this does not eliminate the possibility that they are valid problems. In a further one case the evaluator specified a breakdown in the interaction they experienced, and in the final case the same evaluator described the software's performance rather than the user's. As in the case of the behavioural outcomes, the wrongly specified performance outcomes were defined by the less experienced MMD, therefore their specification could be due to the lack of evaluation experience on the part of the evaluator. This evaluator also did not specify performance outcomes to 6 of the problems she predicted. This indicates that less experienced usability evaluators need further support or training in identifying such outcomes. The fact that for seven problems, performance outcomes were not specified may also indicate that further support is required for the prediction of the other two types of performance outcomes: "user cannot achieve their goal without external help" and "user task fails", both of which suggest that users have encountered severe problems in their interaction. For the prediction of such outcomes, evaluators need not only detailed knowledge of the target user group but also sufficient understanding of the tasks they need to perform. As the nature of the tasks performed with IMM are related to learning and achieving educational goals, which are complex and ill-structured in comparison to procedural tasks, this needs to be reflected in EEMs for IMM.

6.4.6 Prediction of Comprehension Difficulties

Figure 6.9 presents the number and type of the components of comprehension and user interaction problems correctly predicted by the SMEs, as well as the number and nature of the problems they predicted which were not experienced by the users.

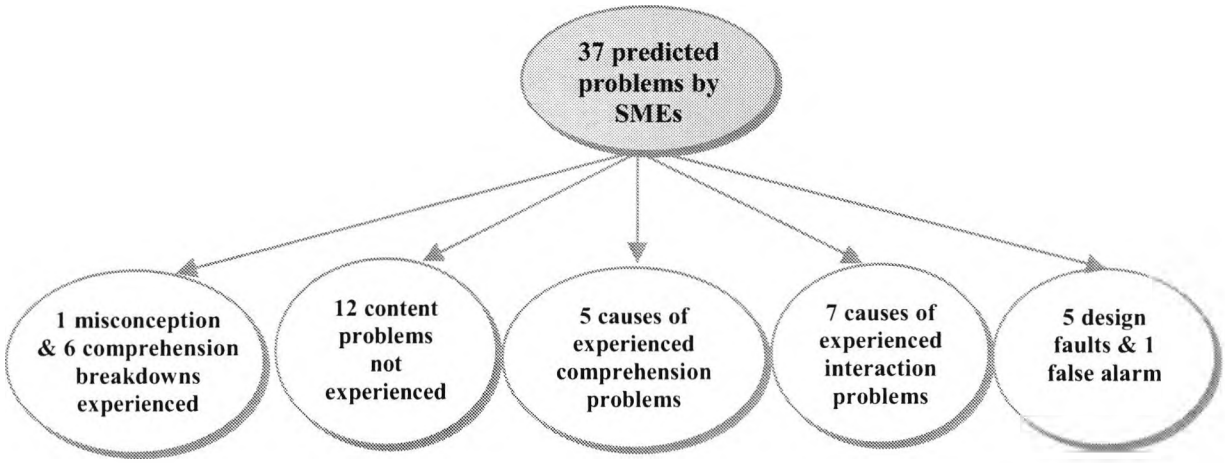


Figure 6.9: Distribution of Comprehension Difficulties

Although the Maths tutors were explicitly requested to specify misconceptions likely to result from the design of the Maths software, only one valid misconception was predicted by one of them. For a further six problems the tutors predicted that learners may experience difficulties comprehending the Maths content as a result of the MMUI design, without defining the nature of the misconception or what concepts precisely learners may have difficulties comprehending. Therefore, such effects were classified as comprehension breakdowns rather than misconceptions. There are a number of reasons for this result. Firstly, the evaluation method and its criteria explicitly prompt evaluators to consider whether the individual media design can hamper learners with low prior knowledge in building correct understanding of the material. Therefore, the Maths tutors accurately identified 6 comprehension breakdowns describing such potential problems, which were indeed experienced by some of the students. However, the tutors could not define precisely what concept could be misunderstood or how students' learning performance may be affected in such cases. As discussed in Chapter 5, the prediction of comprehension problems is complex as it requires knowledge of learners' existing conceptual network, as well as an ability to determine whether the MMUI design is likely to hamper learners with a low level of prior knowledge to build adequate mental models of the subject matter. Sufficient information regarding the target learner group's conceptual networks was not provided in the user description given to the evaluators. Furthermore, the framework provided in the EEM describing how people learn with MM and the evaluation questions included do not explicitly support evaluators in inferring how multimedia representations can support or hamper learners in developing mental models of the subject matter represented. Finally,

although the two Maths tutors who participated in the evaluations had knowledge of the Maths material, they did not possess extensive pedagogical knowledge to enable them to make such predictions. One of the tutors, who had 10 years of experience of teaching Maths formally and informally, explained during the post-evaluation interview that as misconceptions are consequences of design problems, she did not feel the need to specify them. However, it is explicitly specified in the evaluator instructions and the problem report forms that they need to describe misconceptions where possible. The above results could be due either to the lack of evaluation experience of the SME or to the fact that the SME problem report format is not adequately formulated or is insufficient to ensure that SMEs specify likely misconceptions learners may formulate as a result of a design fault.

The two evaluators also predicted 12 problems which relate to the effectiveness of the Maths content and were also depicted in Figure 6.4. From the user test data it appears that the users did not experience these content-related problems. Independent instructional specialists could not be involved to validate this set of problems as in the previous study, however there is a potential that some of them describe valid instructional issues, which will need to be taken into consideration while redesigning the application in order to improve its pedagogical effectiveness. Therefore, involving SMEs to use the EMMCW can be quite beneficial in uncovering areas of the design likely to hinder learners in understanding the subject matter or in achieving their learning goals, although with limited power in specifying the precise effect on comprehension. Due to that, IMM developers may not be able to correctly distinguish between problems which are critical to learners' understanding and those which can only slow their comprehension.

The tutors were also successful at specifying causes of experienced comprehension problems, as they defined valid causes of 5 problems all regarding the instructional effectiveness of the MM presentation. The SMEs however did not specify what likely misconceptions could result from these issues. The likely reasons for this were discussed earlier in this sub-section.

Both Maths tutors, but particularly the one with some software engineering experience, also identified 7 problems which deal with user interaction issues with the MMUI,

which were experienced by the users. They are mostly concerned with affordance of interface objects, the comprehensibility of interface elements, the screen layout, the structure of the presentation, and one also deals with the lack of adequate learning support provided to learners. Such issues can indirectly hamper users in achieving their learning goals, therefore identifying them during formative stages of IMM development can contribute to improving the efficacy of the design. However, as the evaluators specified mostly the design features which are likely to cause interaction problems, and only in three cases they also specified the likely cognitive breakdowns, IMM developers would not be able to prioritise how urgently such problems require attention.

In terms of the thoroughness of predicting valid user problems, the Maths tutors who applied the EMMCW identified quite a significant number of such problems, both comprehension and user interaction ones. In particular, they predicted 10 of the 13 comprehension difficulties experienced by the students (or 77%) and 11 of the 51 user interaction problems (or 22%).

6.5 Discussion of the Effectiveness of the EMMCW

This section compares the results produced using the EMMCW firstly to the original MMCW, and after that to the other two EEMs studied in the previous empirical study, and discusses the enhancement achieved in terms of the *validity, diagnostic power* and *actual thoroughness* of the predictions made. At the end, some general comments about the performance of the developed EEM are also presented.

6.5.1 Comparison of the Results to Original MMCW

The results produced using the EMMCW method were compared to those produced using the original Multimedia Cognitive Walkthrough method to identify whether improvement has been observed in terms of increasing the number of valid problems predicted and decreasing the number of false alarms identified.

It was observed that the problem descriptions produced using the EMMCW were more specific referring to concrete design features, whereas those made using the original MMCW were specified at a higher level of granularity. For example, compare the following two problem descriptions:

EMMCW: "The summary information on Page 1 (leaflet) is not highlighted and obvious. Navigation buttons take prominence on page."

MMCW: "No objects are highlighted anywhere other than by sequence as being more important than others. Very few highlighting techniques are used."

The introduction of the problem report format could have influenced this outcome to a great extent. Lower level of detail problem specification is more preferred as it eases problem location and identification, and can aid more successful redesign based on precise specification of design faults. However, the difference in the level of granularity of the two problem descriptions hampered to an extent the comparison between the two predicted problem sets. The results of the comparison are presented in Table 6.2.

6.5.1.1 Discussion of Validity

As can be seen from Table 6.2, a similar number of problems were identified using the original and the EMMCW, as the MMDs who used the EMMCW identified three problems more than those who used the MMCW. In terms of the validity of the predictions, the difference is quite significant, as with the aid of the EMMCW twice as many valid user problems were predicted. Therefore, 100% improvement is observed in the number of valid user problems predicted. Considerable improvement in reducing the number of false alarms was also achieved, as 9 false alarms were identified using the MMCW (as shown in Table 4.15 in Chapter 4), whereas 3 problems predicted using the EMMCW were classified as false alarms (see Table 6.4 for design fault distribution). As the content-related problems predicted by the MMDs using the EMMCW could not be validated using independent judges, no final conclusions can be drawn regarding the number of valid problems or false alarms predicted.

Table 6.2: Comparison of the Validity of Problems Predicted Using the Original MMCW vs. those Predicted using the EMMCW

	MMCW	EMMCW
Total # of problems predicted	19	22
Total # of predicted problems experienced by users	9 (47%)	18 (82%)
Total # of content related problems not experienced by users	1 (6%) (ratified by judges)	1 (4%) (not ratified by judges)
Total # of false alarms	9 (47%)	3 (14%)

6.5.1.2 Discussion of Diagnostic Power

As mentioned above, nearly half of the predicted problems (or 42%) correctly specify three of the four components used to describe the effects on the users caused by the design problems. This is a significant improvement in comparison to the original MMCW with only 3% of the problems predicted being fully specified and 76% of the problems not specifying any likely effect on users. In particular, the MMDs who applied the EMMCW specified considerably more comprehension breakdowns, 8 in total, whereas none were predicted using the original MMCW. As the goal of IMM is to train and educate learners, giving adequate support to MMDs to predict potential problems in comprehensibility of the multimedia representations can be very beneficial. A distinct advantage of such an EEM is that, if SMEs are not available to perform formative evaluation potential comprehension difficulties could still be identified and rectified.

In terms of predicting valid behavioural and performance outcomes a great improvement over the performance of the MMCW was also achieved. Nine valid behavioural outcomes were identified using the EMMCW, whereas none were specified using the MMCW. Therefore, considerable increase was observed, which is primarily due to the introduction of the problem report format. Furthermore, 13 valid performance outcomes were specified using the EMMCW, nearly half to do with comprehension, which is again a significant improvement over the results of the MMCW, as only one of these was specified using this method.

6.5.1.3 Discussion of Actual Thoroughness

Finally, in terms of the number and type of user problems predicted some improvement was observed with the use of the EMMCW by the MMDs. In particular, 15 of the user

problems found during the user tests were predicted, or 23%, including 5 comprehension difficulties. In comparison, 9 or 14% of the user problems were predicted using the MMCW, only one of which is a comprehension difficulty. This result is not surprising as both EEMs focus on media design and media integration issues, therefore aspects of the user interaction with MMUIs are not strongly emphasised. The interaction problems identified with the EMMCW are concerned with affordance, comprehensibility of presentation objects, misleading functionality, individual media design and screen layout.

6.5.2 Comparison of the Results to all Existing EEMs Studied

It is also important to establish whether overall enhancement of the effectiveness of the EMMCW method has been achieved, not only in comparison to the original MMCW. Therefore, the results presented in this chapter were collectively compared to those achieved by all three existing EEMs studied in this thesis. All four EEMs are again compared based on the three criteria investigated in the hypotheses, namely *validity*, *diagnostic power* and *actual thoroughness* of predictions made.

6.5.2.1 Discussion of Validity

The comparison of the validity of the predictions made using the new and the existing EEMs is based on the validations performed against the user test data only, as no independent judges were available to further validate the 12 content-related problems predicted using the EMMCW (see Figure 6.4). The 8 design faults identified using the new EEM were classified as false alarms, as explained in Section 6.4.2. Therefore, the total number of false alarms identified using the EMMCW is 9, including one problem which contradicted the users' experience. The data about the validity of the three existing methods is taken from Table 4.15, in Chapter 4. The results are presented in Table 6.3.

The validation results presented in Table 6.3 show that positive improvement has been achieved in terms of both increasing the number of valid user problems predicted and decreasing the number of invalid user problems identified with the use of the developed method. Whereas with the use of the existing three EEMs under half of the predictions made were valid, with the use of the EMMCW nearly two thirds of the predictions revealed user problems that were actually experienced by typical users. A more

significant improvement was observed in reducing the number of false alarms predicted, as with the use of the EMMCW between 43% and 78% fewer false alarms were identified. It needs to be acknowledged again that as the content-related problems predicted using the EMMCW could not be validated using independent judges, the final total of valid problems and false alarms could not be established. Despite that, the evidence presented in Table 6.3 shows that the credibility of the evaluation results achieved using the EMMCW would be greater than the other three EEMs, and IMM and usability practitioners can apply the method with greater confidence in the potential benefit of using the method.

Table 6.3: User Data Validation of Predicted Problems for Each EEM

	EMMCW	MMT	MMCW	IMMC
Total # of problems predicted	58	56	19	70
Total # of predicted problems experienced by users	37 (64%)	27 (48%)	9 (47%)	16 (23%)
Total # of false alarms	9 (16%)	21 (37%)	9 (47%)	53 (76%)

6.5.2.2 Discussion of Diagnostic Power

The diagnostic power of the four EEMs was compared based on the number of fully-specified problems (FS), the number of problems specifying the effect design faults may have on users without specifying the cause (US-NC), the number of problems specifying only problem causes without the likely effect on users (US-NE), and those only revealing redesign suggestions (RS). The results of which are presented in Table 6.4.

Table 6.4: Diagnosed vs. Under-Specified Predicted Problems for all Four EEMs

Diagnostic Category / EEM	FS	% FS	US-NE	% US-NE	US-NC	% US-NC	RS	RS %
EMMCW	46	79%	10	17%	2	4%	0	0%
MMT	15	27%	38	68%	0	0%	3	5%
MMCW	6	32%	11	58%	2	10%	0	0%
IMMC	2	3%	53	76%	13	18%	2	3%

Based on the findings revealed in Table 6.4 it can be concluded that with the application of the EMMCW a substantial improvement has been achieved in terms of specifying the causes and the likely implications on users of usability problems. This is primarily due to requesting evaluators to specify problems using the two problem report formats suggested and to providing them with problem report forms enabling them to describe

usability problems in these formats. As uniform problem reports formats were not introduced in any of the other three EEMs, the diagnostic power of all of them is significantly lower. Had such format been applied, the diagnostic rate of the existing EEMs could have been higher.

6.5.2.3 Discussion of Actual Thoroughness

Figure 6.10 illustrates the prediction rate of user problems shown by the four evaluation methods. The summary on the right-hand side indicates the total number of user problems predicted and the percentage of the total number of user problems found in the IMM application.

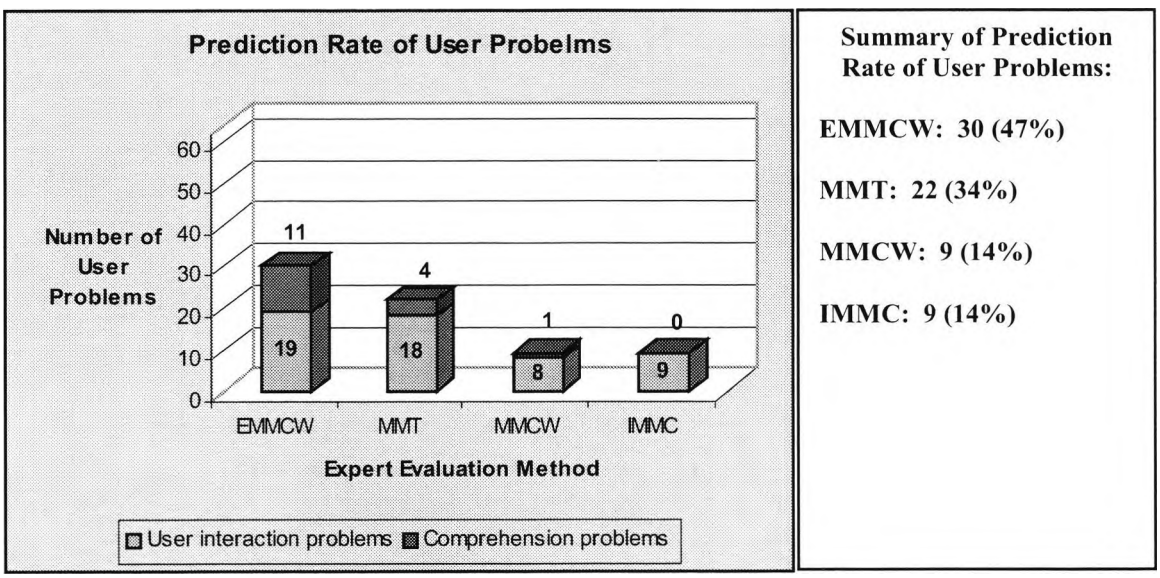


Figure 6.10: Coverage of User Problems per EEM

The summary presented in Figure 6.10 shows that with the aid of the EMMCW nearly half of the problems experienced by the users were uncovered, which is an improvement in comparison to the three existing EEMs. The chart presented on the left-hand side in Figure 6.10 reveals that a significant increase in the prediction of comprehension problems was achieved using the developed EEM, as it discovered nearly all 13 comprehension difficulties the learners experienced. These difficulties were all classified as major problems, as they indicate that the students could not achieve some of the learning goals of the IMM software. Identifying and rectifying potential comprehension difficulties early in the development of educational software is vital for ensuring the pedagogical effectiveness of such software. Therefore, using the EMMCW during formative stages of development of IMM could ensure that a good quality

product is delivered to students. The prediction rate of user interaction problems of the EMMCW is similar to the one achieved using the MMT, however it greatly improves the performance of the original MMCW and the IMMC. There is still a need to further improve the method's thoroughness in order to achieve better results, compatible with those achieved by established usability evaluation methods such as Heuristic Evaluation. Nielsen (1994) quotes a level of thoroughness of Heuristic Evaluation as high as 75% of the actual user problems when using eight usability evaluators, and as high as 90% when fifteen evaluators are involved.

Finally, it needs to be highlighted that both types of evaluators who applied the EMMCW predicted both types of user problems. Therefore, experts of only one type could be asked to use the method if project resources do not allow the involvement of both. However, it is recommended that at least 4-5 experts from different professional backgrounds be involved in usability evaluation using the EMMCW to achieve a sufficient level of thoroughness of the evaluation.

6.5.3 General Comments about the Effectiveness of the EMMCW

From the expert evaluations it was revealed that the second questions of each pair may not have always prompted the identification of as many cognitive breakdowns as expected. However, the considerable number of valid causes identified shows that the first part of the evaluation questions was more successful at identifying usability problems and specifying the precise causes of such.

Furthermore, it needs to be noted that the MMDs who applied the EMMCW were very successful at instantiating the evaluation questions in the context of the Maths application, as they did not simply provide answers to the questions. Instead, they used them in reasoning about the design and provided low level, specific problem descriptions.

Finally, the use of the two problem report formats introduced in the EMMCW was found useful but to a limited extent. It seemed to have encouraged the MMDs to specify cognitive breakdowns and outcomes to nearly all predicted problems, which was pointed as one reason for the higher number of valid user problem predicted using the EMMCW. However, the MMDs may have become confused with the difference

between behavioural and performance outcomes, as they sometimes specified one in the place of the other. As mentioned earlier, Lavery and Cockton (1997) have resolved this close similarity between the two by merging them into one category "outcome" for specifying the difficulty the user experiences as a result of a dialogue failure or a knowledge mismatch. The SME problem report format did not achieve the expected result, as the two SMEs rarely specified the likely misconceptions the user may form as a result of a design fault. As explained earlier, there are a number of reasons for this, including the lack of experts' evaluation experience. The improvement of the prediction rate of misconceptions is complex and could be beyond the formulation of a new report format.

6.6 Implications of Results for Enhancing the Effectiveness of EEMs for IMM

The previous sections provide evidence of and highlight the extent of the improvement in the effectiveness of the evaluation results achieved by applying the EMMCW. This section revisits the four hypotheses formulated in Chapter 5 and draws conclusions as to whether each hypothesis can be totally accepted, partially accepted or rejected based on the findings of the expert evaluations reported in this chapter.

6.6.1 Conclusions Regarding Hypothesis 1

Hypothesis 1: The prediction rate of valid *attentional, perceptual, information integration and comprehension breakdowns* can be improved by asking experts to explicitly consider how different forms of media representations affect users' cognitive processes of attention, perception, information integration and comprehension of information presented in various media.

The evidence presented in Section 6.4.3 suggests that the MMDs predicted cognitive breakdowns to 92% of the problems they identified. This high proportion shows a significant improvement in the number of cognitive breakdowns predicted over the existing EEMs. Only half of these breakdowns were validated by the user test data. Due to inability to involve independent judges and the lack of sufficient eye-tracking data, the remaining half of the predicted breakdowns could not be validated or invalidated with certainty. Therefore, as there is no sufficient data to totally accept Hypothesis 1, it

is **partially accepted**. The prediction of such a high proportion of cognitive breakdowns can be primarily attributed to the evaluation questions provided in the new EEM. The usability problem report format introduced as part of the evaluation process also encouraged their specification. However, the evaluation data produced by the SMEs suggests that in some cases the evaluation questionnaire is sufficient for specifying cognitive breakdowns, as they specified a few cognitive breakdowns without being explicitly prompted to do so by the report format they were given.

6.6.2 Conclusions Regarding Hypothesis 2

Hypothesis 2: The prediction rate of valid *behavioural outcomes* can be improved by ensuring evaluators are well acquainted with the target user group, and by supporting them in inferring the likely user behaviour following a breakdown in users' interaction with the MMUI.

The data presented in Section 6.4.4 highlight that the MMDs specified behavioural outcomes to 58% of the problems they predicted, and in 64% of these cases the outcomes were experienced by the users. Therefore, Hypothesis 2 can be **partially accepted**. These results show that providing evaluators with a user specification and encouraging them to specify behavioural outcomes on an evaluation form supports the predictions of valid behavioural outcomes to a degree. However, further support is required to aid evaluators in predicting all types of behavioural outcomes. However, as the data from the expert evaluations confirmed, not all usability problems encountered during interaction with IMM result in immediately observable physical actions on the part of users, therefore defining behavioural outcomes to all problems predicted would be unreasonable to expect.

6.6.3 Conclusions Regarding Hypothesis 3

Hypothesis 3: The prediction rate of valid *task performance outcomes* can be improved by ensuring evaluators are well acquainted with the target user group, and supporting them in inferring the implications on the users' task resulting from the behaviour they are likely to adopt.

In Section 6.4.5 evidence was presented that the MMDs specified performance outcomes to 71% of the problems they identified, 70% of which were experienced by

the users. Therefore, Hypothesis 2 can be **partially accepted.** The prediction rate of performance outcomes is better than the one of behavioural outcomes. The results indicate that providing evaluators with a user specification and encouraging them to specify performance outcomes on an evaluation form proved more effective for promoting the specification of valid performance outcomes. The fact that the connection between observable actions and resulting performance outcomes is highlighted in the EEM's *Specifying Usability Problems* section and in the instructions given to the evaluators, proved important for the successful prediction of the 8 performance outcomes. Therefore, providing more specific information of what such connections might be and giving evaluators more guidance in identifying the likely performance resulting from different type of behaviour could further increase the validity and thoroughness of their predictions.

6.6.4 Conclusions Regarding Hypothesis 4

Hypothesis 4: The prediction rate of *comprehension difficulties* can increase by supporting evaluators in predicting the implications of IMM design aspects on learners' comprehension of the content, particularly considering the needs of learners with little prior knowledge of the subject-matter.

Finally, the results which were revealed in Section 6.4.6 indicate that the SMEs specified the likely outcomes on learners' comprehension only to 18% of the problems they identified. The MMDs, however, were more successful at specifying possible implications on learners' comprehension of media design features, as they specified comprehension breakdowns and/or comprehension outcomes to 11 of the 24 problems they predicted, or 46%. However, these descriptions were only high level ones, warning that design features may slow or hamper learners' comprehension. Based on these evidence Hypothesis 4 can also be **partially accepted.**

The main reasons why the SMEs were not as successful at predicting learner misconceptions include the complexity of the learning process, the insufficient support for identifying such provided by the EMMCW, and the lack of pedagogical knowledge and experience particularly on the part of one of the Maths tutors. To improve the prediction rate of valid comprehension problems guidance grounded in theoretical frameworks of learning with IMM is required, such as those outlined in Chapter 5. This

will provide evaluators with necessary understanding of how learners build mental models of multimedia representations, and how the major factors of IMM design - the content, the instructional approach adopted, the learner and their cognitive processes, the context of use and the multimedia application – all relate to each other. Furthermore, specific instructions as to how to use the rationale provided in such framework for reasoning about the effectiveness of IMM design is also needed.

6.7 Conclusions

This chapter described how the EMMCW was created based on the MMCW method developed by Faraday and Sutcliffe (1997). It also presented an empirical study validating the predictive power of the EEM developed in this research, called Educational Multimedia Cognitive Walkthrough, aiming at improving the effectiveness of the existing EEMs for IMM software. The evaluation results presented above revealed that with the use of the EMMCW positive improvement has been achieved in terms of both increasing the number of valid user problems predicted and decreasing the number of invalid user problems identified. In particular, it was found that the prediction rate of valid comprehension problems was significantly improved in comparison to the three EEMs studied in Chapter 4. The instructional effectiveness of IMM applications is of vital importance for the success of such software. Therefore, the developed evaluation method can promote the development of more pedagogically sound IMM solutions, by supporting the early identification of problems threatening the instructional effectiveness. Another enhancement achieved by the use of the EMMCW was in the proportion of fully specified usability problems. Therefore, software developers would be better informed of the causes of usability problems and their likely implications on users, based on which developers could make better decisions as to how to improve the quality of the design of the software being developed.

Despite the positive enhancement in effectiveness achieved, none of the four hypotheses formulated in Chapter 5 could be totally accepted. This indicates that further improvements are necessary to achieve a higher level of validity and thoroughness of expert predictions of usability problems using this method. Therefore, suggestions of how the EMMCW can be further improved are proposed in the next chapter.

Chapter 7

Discussion and Conclusions

Chapter 7

Discussion and Conclusions

7.1 Summary

This thesis was divided into two consecutive parts. In the first part the effectiveness of existing EEMs for IMM was investigated. For the purpose, a comprehensive framework for the assessment of the effectiveness of such EEMs was first proposed. The framework was based on a review of previous EEM effectiveness studies and literature on effective design of interactive multimedia learning systems. The framework, containing eleven effectiveness measurement criteria, was presented in Chapter 3. Using this, the effectiveness of three existing usability evaluation methods for IMM - MMT, IMMC and MMCW, was studied empirically and conclusions were drawn regarding their effectiveness according to each of the eleven criteria defined in the effectiveness framework. This investigation was presented in Chapter 4. The results of the empirical study highlighted the main strengths of the three EEMs, and identified the main limitations which constrain their effectiveness in key areas such as validity, thoroughness, reliability and diagnostic power.

The second part of this thesis research aimed to develop a method for the usability evaluation of IMM applications, which is more effective in predicting valid usability problems than the existing methods studied. The conceptual approach for enhancing the effectiveness of EEMs for IMM, which was used to develop the new EEM, was given in Chapter 5. Four hypotheses, each of which proposed a way of enhancing specific elements of method's effectiveness, were then formulated. The hypotheses were drawn from evidence of the empirical study conducted during the first part, and were also informed by existing theoretical work on how people learn with IMM. A new EEM for IMM was then developed, called Educational Multimedia Cognitive Walkthrough. It derives from one of the three existing methods studied, namely the Multimedia Cognitive Walkthrough (Faraday and Sutcliffe, 1997), and embodies the effectiveness enhancing characteristics proposed in the four hypotheses. How the method was

developed was explained in Chapter 6. Finally, the effectiveness of the developed EEM was tested empirically, under similar conditions as the existing EEMs, and conclusions were drawn as to its effectiveness, and the extent of the improvement achieved. The results were encouraging and highlighted positive improvement in the performance of the new EEM over the existing ones, in terms of increasing the number of valid user problems predicted, decreasing the number of false alarms identified, and providing more thorough specifications of the usability problems predicted. The validation study was also presented in Chapter 6.

Therefore, this thesis has met its main aim to investigate into and improve the effectiveness of existing expert evaluation methods of Instructional Multimedia. The five objectives set at the onset of this research have also been successfully achieved. Namely, an EEM effectiveness framework was developed and used for assessing and comparing the performance of three existing EEMs. Four hypotheses for rectifying the limiting features of existing EEMs for IMM were formulated and a new evaluation method was developed based on the recommendations made in the hypotheses. The thesis concluded by demonstrating that the new EEM could help identify a significant proportion of existing user problems in IMM user interfaces, which if rectified could positively improve the usability and instructional effectiveness of such software before it is released to be used by learners.

The remainder of this chapter discusses the contributions and the implications of the thesis research. Limitations of the research are also detailed, followed by suggestions for possible future research on the theory and practice of effective expert evaluation of IMM.

7.2 Contributions and Implications of Research

The thesis research has helped to define the effectiveness of expert evaluation methods for IMM, and promote our understanding of the nature of the processes, the knowledge required by evaluators, and how the knowledge can be used to enhance the effectiveness of expert evaluations. The research has important implications for the formative evaluation of IMM and research for developing more effective expert evaluation methods.

The following sub-sections present the contributions of the thesis research.

7.2.1 Producing a Systematic Framework for Measuring the Effectiveness of Formative EEMs for IMM

Chapters 2 and 3 of this thesis argue that in order to assess and improve the effectiveness of EEMs for IMM, a coherent framework promoting our understanding of what constitutes an effective EEM is necessary, together with systematic procedures for measuring the effectiveness of such methods.

No one has previously tried to bring existing literature on effective evaluation of IMM into a coherent framework for measuring and comparing the merit of EEMs for IMM. Previous critical reviews of such evaluation techniques were quoted in Chapter 2, however none has looked at a comprehensive set of factors and rarely empirical measures of their merit have been used. Studies comparing the effectiveness of traditional usability inspection methods were also reviewed in Chapter 2. However, it was argued that the criteria used cannot be applied directly for assessing the effectiveness of EEMs for IMM, as they do not address specific aspects of the usability evaluation process of IMM.

The framework proposed in Chapter 3 provides a set of eleven criteria (see Figure 7.1) and procedures for assessing the effectiveness of formative usability evaluation methods, particularly those for IMM. The criteria are primarily concerned with establishing the quality of the results produced during expert evaluations, namely their validity, thoroughness, reliability and completeness of the usability problem reports (diagnostic power). These criteria were considered to be particularly important for EEMs' performance and were classified as *primary factors,* highlighted in Figure 7.1. Another set of criteria aims to assess the usefulness of EEMs, particularly their cost-effectiveness, ease of use and ease of learning. These criteria were classified as secondary as they do not deal directly with the effectiveness of the evaluation results produced using EEMs. All criteria presented in Figure 7.1 were used to study the performance and usefulness of the existing EEMs, which was covered in Chapter 4. However, in the second part of the thesis research particular emphasis was placed on

improving the effectiveness of three primary factors and one secondary factor, which was addressed partially. These criteria are highlighted in Figure 7.1.

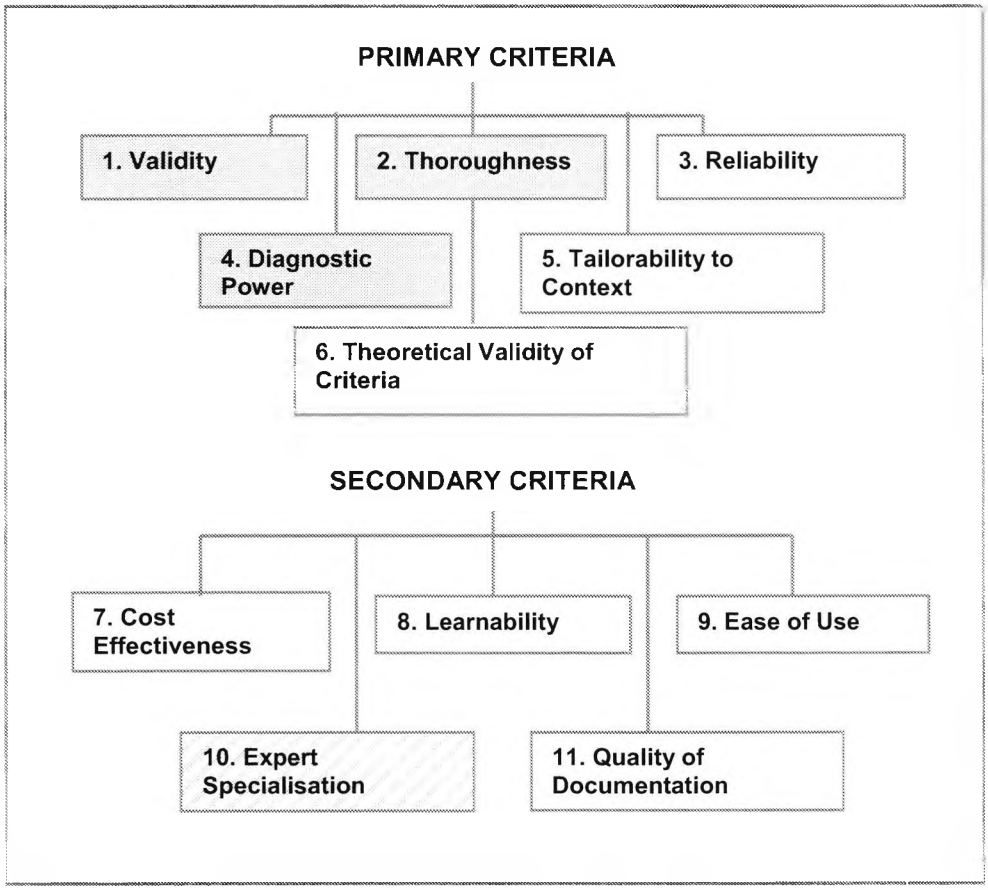


Figure 7.1: EEM Effectiveness Framework and Factors Addressed

The effectiveness criteria were primarily extracted from studies assessing the effectiveness of traditional usability evaluation methods, most of which were adapted to address the specific characteristics of the IMM evaluation process. For example, the measure of *validity* was extended to include two separate stages of problem validation. The first was conducted against user test data (as widely used by researchers) and the second involved instructional experts to validate content and instructional problems, which user test data may not be able to validate. The adaptation of effectiveness measurement criteria were discussed in detail in Chapter 3.

The framework also specifies the factors likely to contribute to the effective or ineffective performance of EEMs. For example, together with defining what constitutes a reliable EEM, the framework also suggests factors which can hinder a method's reliability, such as the ambiguity of the evaluation criteria. Therefore, the framework

provided not only a systematic approach for assessing a method's effectiveness and usefulness, but it also guided the analysis of the method's ineffectiveness and helped reveal the reasons for ineffective performance. Based on this, the EEM's aspects that needed addressing were identified, and actions were then identified to improve the methods' performance.

The main utility of the proposed EEM effectiveness framework is to aid the systematic and comprehensive assessment and comparison of the merit of formative EEMs, particularly those for IMM. As evaluation methods for IMM continue to be developed, the framework can be applied directly by researchers to establish the effectiveness of existing EEMs, and to highlight their strengths and limitations. This can inform efforts to improve the effectiveness of already developed methods, and to enhance the quality of the evaluation results they produce. Ensuring the quality of expert predictions can enable instructional software developers to make more effective decisions regarding how to improve the quality of IMM designs. This can ultimately promote the development of high quality IMM software.

Chapter 2 also argues that existing EEMs for IMM are rarely engineered towards standards of effective performance in terms of the quality of the results produced or meeting the needs of the evaluators who will be using them. The method effectiveness framework can point developers of new EEMs to factors that need to be considered and standards that need to be attained by such methods to ensure both the effectiveness of the evaluation results and the usefulness of the methods they are developing. It can then aid in establishing the effectiveness of the methods once they are developed.

Finally, the proposed framework can be used by IMM developers to compare the relative merit of available EEMs when considering which one to adopt in their usability engineering practices. Some effectiveness criteria require empirical data to measure them (e.g. method validity, reliability, diagnostic power), therefore they need resources which may not be available for benchmarking available EEMs. Other effectiveness criteria, such as theoretical validity, tailorability to context, aspects of thoroughness, aspects of learnability, and quality of documentation, can be assessed without empirical data. Therefore, usability practitioners will be able to use them directly for benchmarking different techniques.

7.2.2 Demonstrating the Effectiveness of Existing EEMs for IMM

A few methods for evaluating IMM software have been developed to date and new ones are continually being developed. It was highlighted in Chapters 1 and 2 that the existing evaluation methods have seldom been the subject of empirical analysis, particularly investigating the validity and the reliability of the predictions made. The unknown effectiveness of these EEMs was identified as a major reason why such methods are not widely used for assessing the usability of IMM environments (Tergan, 1998). Due to the lack of empirical validation, usability practitioners and software developers cannot identify which methods they could use in developing IMM solutions. The insufficient evidence of the cost-effectiveness of such methods and the limited number of successful examples of their use can cause resistance in practitioners when deciding on the most appropriate evaluation techniques to adopt. Furthermore, the lack of investigation into the effectiveness of these EEMs means that their strengths and limitations are not well understood. Therefore, developers do not know at what stages of IMM development the different methods are more appropriate to be used, and researchers are not informed as to what aspects of EEMs need addressing.

The results from the empirical study presented in Chapter 4 promote understanding of the effectiveness and usefulness of three existing EEMs for the evaluation of IMM, by revealing the methods' strengths and limitations according to the eleven criteria defined in Chapter 3. Expert evaluations of a commercial IMM application were conducted using each method. The results produced by the EEMs were validated against user test data of the same software. The results were further validated by two independent instructional specialists. The findings from the empirical study have implications for both the usability evaluation practice and research into developing and improving EEMs for IMM, which are discussed below.

7.2.2.1 Supporting EEM Selection Decisions

The study first highlighted the ability of each EEM to predict valid user problems, the types of problems each method is able to predict, how reliable the predictions are, and how the EEMs compare in cost effectiveness. Based on these results, practical implications were drawn as to the emphasis placed by different methods on different aspects of IMM design. These can inform practitioners at what stages of development such methods can be employed and what other methods can complement them to

provide more thorough assessment of the effectiveness of IMM designs. Furthermore, suggestions were made regarding the type of experts and their qualifications required for conducting usability evaluations using the EEMs in order to achieve better reliability of the predictions. Some concerns were also raised regarding the validity and severity of the problem sets likely to be produced, as well as the amount of resources required to use each EEM, informing usability practitioners of the potential cost benefit of employing such EEMs in their practice. Finally, information regarding the usefulness of these methods was also presented, revealing the experience of the evaluators who applied the EEMs. This can tell usability evaluators how much training they require, how much support they can get from the methods' documentation to assess the quality of IMM interfaces and to identify potential problems.

This information can greatly aid project managers and usability practitioners in deciding what usability evaluation technique/s to use on their projects, at what stage to apply them and the practical constraints that need to taken into consideration, such as training time required. As such decisions can have important implications for the effectiveness of the IMM development process as a whole, and the quality of the IMM produced, having such a comprehensive report on EEMs' performance and usefulness can help practitioners make effective decisions more rapidly.

7.2.2.2 Demonstrating the Limitations of EEMs for IMM

The results of the EEMs evaluation also showed that none of the methods is as effective as IMM developers and usability practitioners would like them to be, and revealed some major limitations of the EEMs. This highlighted aspects of the EEMs that need to be addressed in order to improve their performance and ensure they deliver satisfactory results. The main limitation identified was the unsatisfactory validity of the results produced, due to the large number of false alarms which were predicted, primarily using the IMMC. Furthermore, it was found that none of the methods was very successful at predicting real user problems that exist, particularly those concerning users' comprehension of the subject matter, which are amongst the most severe problems. The main implication of these findings was that the ineffectiveness of the evaluation results can lead to ineffectiveness of the redesign decisions made, as they would be based on invalid or incomplete predictions. Therefore, they highlighted that addressing the limited ability of such EEMs to predict valid user problems, particular comprehension

ones, needs urgent attention. Another major limitation found was that all problem reports were incomplete, particularly because they do not specify the likely effect design faults may have on target users. This can limit the ability of IMM developers to prioritise usability problems. Two of the methods, the MMT and the MMCW, also exhibited poor reliability of usability predictions.

More importantly, the further analysis of the EEMs' limitations, discussed in detail in Chapter 5, revealed the major factors and characteristics limiting the performance of the methods. For example, the large number of false alarms and poor user problem coverage were attributed to the fact that the expert evaluations focused predominantly on IMM design features and did not enable evaluators to link these features to likely user behaviour or performance. The detailed analysis of the performance limiting characteristics can provide an important part of the conceptualisation of an approach to improving the ability of EEMs for IMM to predict valid user problems. Such a systematic approach was developed in Chapter 5 and validated in Chapter 6. The discussion of the performance limiting characteristic can inform method developers as to what features to avoid when developing new EEMs for IMM.

The empirical study presented in Chapter 4 also demonstrated that the effectiveness and usefulness of EEMs can be tested using the framework proposed in Chapter 3 to identify methods' strengths and limitations. The effectiveness measurement procedures were refined while assessing the performance of the EEMs, and the improved version of the framework was presented in Chapter 3.

7.2.3 Producing a Systematic Approach for Enhancing EEMs' Effectiveness

Chapter 2 argues that existing EEMs for IMM are developed predominantly based on evidence of practical experience rather than on systematic frameworks or theories of effective instructional design. The problem with such a non-scientific approach to developing EEMs is that the validity and reliability of the results they produce cannot be proven (Tergan 1998). It is particularly difficult to prove how decisions regarding the instructional effectiveness of IMM can be derived and how potential learning problems can be identified. The author argues that evaluation methods for educational software need to be based on models encompassing instructional design knowledge which can aid software developers and researchers to effectively study the design of such software.

Jacobson and Spiro (1994) also advocate the need for a theoretically grounded comprehensive evaluation approach to adequately assess the instructional efficacy of educational software.

In the above conceptions, a *hypothetico-deductive approach* to improving certain aspects of the effectiveness of EEMs for IMM was proposed in Chapter 5. It is concerned with enhancing the validity, thoroughness and diagnostic power of expert predictions of the usability of IMM. The approach firstly derives from the empirical evidence of the EEM comparison study described in Chapter 4. Secondly, it is founded in relevant theoretical frameworks of learning with MM. The approach comprises of four components, the significance of which is discussed below.

7.2.3.1 Definition of Effectiveness-limiting and Effectiveness-enabling Characteristics

One major contribution of the research is the definition of the factors found to limit the validity and thoroughness of EEMs for IMM, particularly their validity, thoroughness and diagnostic power. Concrete factors attributed to the poor performance of the three EEMs were identified, and were defined in Chapter 5. They were concerned with the nature of the evaluation process (such as evaluations focusing on design features without considering their effects on users), the knowledge necessary to infer comprehension problems (e.g. lack of considering relevant user characteristics), the reasoning required (e.g. the amount of subjective judgement needed), or the lack of a uniform way for defining usability problems. The effectiveness-limiting characteristics that were identified guide researchers into what aspects of the existing EEMs need to be addressed and help them make effective decisions as to what actions can be taken to improve the methods' validity and/or thoroughness. This analysis also points method developers to aspects they need to avoid in developing new EEMs in order to ensure their effectiveness.

Chapter 5 also identified some aspects of the expert evaluation process, which promoted the effective prediction of valid usability problems. Identifying these aspects can also be of use to method developers, as they reveal features which could enhance the performance of expert predictions.

7.2.3.2 Value of the Taxonomy of User Problems

A taxonomy of user problems was produced based on the user test data generated in the study presented in Chapter 4. The taxonomy not only distinguished between two types of problems – user interaction and comprehension, but also defined the structure of each type and the components they comprise. The taxonomy promotes understanding of the nature and types of valid user problems. Firstly, it can be used by instructional software evaluators to describe the problems experienced by users during user tests in a uniform and complete matter. Secondly, the taxonomy can support the categorisation of usability problems predicted during expert evaluations.

The taxonomy was successfully used to redefine the problems experienced during the user tests and to specify and categorise the problems predicted using the EMMCW. These studies have shown that user and expert problems with the type of IMM evaluated can be classified reliably using the taxonomy. The relationships between the individual components have only been partially established, and the utility of the taxonomy for categorising problems with other types of IMM applications needs to be further investigated.

7.2.3.3 Review of Theories of Learning with IMM

This aspect of the systematic approach emphasised the need to base evaluation techniques for IMM on sound theoretical frameworks of learning with MM, in order to ensure the effectiveness of the predictions made. It also argues the need to consider the target learners and their abilities and characteristics in the usability evaluation process. This can enable a learner-centred approach to assessing the quality of IMM designs, ensuring that learners' needs will be met.

Tergan (1998) provides a useful conception of how instructional design knowledge may be used for the evaluation of the effectiveness of educational software. The author critiques 'craft based' evaluation criteria and advocates the use of criteria based on empirical studies and theories on instructional design. Tergan (1998) notes that a 'craft based' approach is weak, prone to producing results which have unproven validity and reliability, due to the fact that experts in the evaluation of educational software often have misleading assumptions about the effects of instructional design elements on the effectiveness of learning processes. Tergan (1998) suggests that the validity of

evaluation predictions regarding the instructional efficacy of software can be improved if decisions are based on an empirically tested instructional design model or a comprehensive theoretical framework. Reeves and Harmon (1994) also conceptualise that to ensure the validity of evaluations of IMM software, evaluation criteria must be derived from empirical evidence regarding the instructional effectiveness and usability of such software.

In Tergan (1998) and Reeves and Harmon (1994) conceptions, a review of theoretical work and empirical evidence of how people learn with MM was performed. As a result the cognitive processes involved in learning with MM were defined together with how MM representations can utilise cognitive abilities to ensure effective knowledge acquisition.

The theoretical frameworks discussed reveal what cognitive psychology and instructional design knowledge is required to promote effective design of IMM environments, particularly to promote their learning effectiveness. The frameworks can be used as a conceptual basis for deriving design guidelines and evaluation criteria needed to support MMUI design, evaluation and re-design decisions. Thus, developing such support for IMM designers and evaluators can ensure the validity of the design decisions and evaluation predictions made using such criteria.

Chapter 5 also defined what knowledge of the target learner group is required to promote more effective design and evaluation of IMM. The need to consider user characteristics, such as their learning styles, acquired individual approach to learning, and motivation, for ensuring user-centred design of IMM was emphasised. One very important factor that needs to be considered is learners' prior knowledge of the subject matter, as it plays a vital role in determining how comprehensible MM representations are. Building interface design and evaluation support which promotes the consideration of relevant characteristics of the target user group, can help develop IMM which meets the needs of its users.

Based on the above conceptions, deductions were made as to how the effectiveness of expert predictions regarding the quality of IMM can be enhanced. The four hypotheses formulated in Chapter 5 propose concrete aspects which can be implemented in EEMs.

They also specify the likely improvement in the predictive power of such EEMs, which can be achieved by implementing such aspects.

7.2.4 Supporting Effective Expert Evaluation of IMM

A new expert evaluation method for the evaluation of IMM was developed, incorporating the improvements suggested in the hypotheses. The new EEM was called Educational Multimedia Cognitive Walkthrough. The main evaluation tool included in the EMMCW which aims to support expert evaluations is a set of evaluation questions. Their benefits are discussed below.

7.2.4.1 The Value of the Evaluation Questions in the EMMCW for Usability Evaluation

From the theoretical concepts introduced in Chapter 5, a set of evaluation questions were formulated in Chapter 6. The first part of each evaluation question is designed to support the identification of potential problems which can occur during user's interaction with IMM interfaces, by prompting the review of aspects of the design of individual media and combinations of visual and verbal media. The second part of each question is designed to enable the identification and specification of the likely effect interaction problems may have on users' cognitive processes of comprehension of the subject matter. Therefore, the questions promote the discovery of comprehension difficulties by supporting the analysis of the effect of users' interaction with IMM software on their knowledge acquisition processes. As the new EEM supports the identification and rectification of potential comprehension difficulties early in the development of educational software, this can ensure the learning effectiveness of such software. This is a considerable improvement over the existing EEMs for IMM, as no other expert method has been reported to have achieved such a synergy between both.

The design rationale built in the evaluation questions is firstly based on the existing cognitive framework of learning with IMM, presented in Chapter 5. Some evaluation questions also aim to utilise knowledge of learners' prior knowledge in reasoning about the comprehensibility of MM representations. By considering users' prior knowledge the questions promote a learner-centred approach to the assessment of the learning effectiveness of educational software. No existing expert method has proposed such a systematic approach to linking the quality of external representations to learners'

existing level of knowledge in order to discover potential comprehension difficulties. Finally, the evaluation questions are also supported by design principles, which aim to clarify the rationale for the questions, which can decrease their ambiguity and potentially increase the reliability of the predictions made (Tergan, 1998).

Therefore, the analytic questions can promote more comprehensive usability evaluation of IMM. By systematically exploiting information on cognitive processes of learning with MM and learners' prior knowledge in assessing the effectiveness of MMUIs, the EMMCW questions can be a useful tool for revealing not only design limitations but also what aspects of the design are likely to hinder learners' comprehension. In that way, it overcomes some of the major shortcomings of existing EEMs for IMM.

7.2.4.2 Demonstrating the Effectiveness of the EMMCW

The effectiveness of the developed evaluation method was empirically tested, the results of which were presented in the second half of Chapter 6. Expert usability evaluations using the EMMCW were conducted on the same commercial IMM application, which was evaluated by the three existing EEMs studied in Chapter 4. The results produced using the new EEM were compared to those produced by the other three methods. The problems predicted using the EMMCW were also validated against the results of the user tests conducted of the same software.

Improvement of Validity

The results presented in Chapter 6 showed that using the new EEM an increased number of valid problems were identified in comparison to the existing EEMs. The second part of each evaluation question prompted the identification and specification of the likely impact design features may have on users' interaction and knowledge construction processes for 79% of the expert predictions, which allowed immediate validation of design faults (Cockton and Woolrych, 2001). The main implication of this is that the effect a design fault may have on users' ability to achieve their task can be identified concurrently with the identification of design faults. Based on this analysis, evaluators are in a better position to rule out issues which have little or no effect on users' tasks and comprehension of the subject matter. Therefore, more critical user problems that have such an effect can be specified, potentially reducing the number of false alarms being specified. Identifying predominantly valid problems during expert evaluations

could allow re-design effort to concentrate on rectifying real problems rather than waste effort on false alarms. Thus, the potential of such effort to positively improve the quality of the software by resolving valid problems at formative stages of development could be increased. Consequently, the cost-effectiveness of expert evaluations can also be improved.

Improvement of Thoroughness

The results from the validation study showed that an increased proportion of the user interaction problems were identified during the expert evaluations conducted using the new EEM in comparison to the existing methods studied. More importantly, with the use of the EMMCW, problems with nearly all areas of the subject matter which students had difficulties comprehending were discovered. Therefore, using the new method, not only more valid problems can be identified, but it is possible that the majority of the most critical user problems can be predicted without the need for user tests. As significantly less time and effort is required to plan, conduct and analyse the results of expert evaluations using the EMMCW in comparison to user tests, the new EEM could present developers with a more cost-effective approach to assessing the quality of IMM applications at formative stages of development.

Improvement of Diagnostic Power

Another important improvement achieved is characterised by the fact that 79% of the predicted problems were fully specified compared to a maximum of 32% using the existing EEMs studied. These problems describe not only the design aspects responsible for usability and learning problems, but also the effects of the problems on users' behaviour and performance. The latter can be of great benefit for the usability engineering process, as IMM developers can prioritise the predicted problems based on how critical the effects specified are for the achievement of users' tasks. This can be done without the need to conduct expensive user tests. Developers can then decide which problems to address first, depending on the resources available. This can potentially ensure that the most critical problems are rectified before releasing the software.

Therefore, the results presented in the validation study in Chapter 6 provide evidence that the new EEM could have a marked effect on uncovering valid severe user

problems, which if effectively rectified could potentially increase both the usability and instructional efficacy of IMM software.

Based on these results, the four hypotheses formulated in Chapter 5 were partially accepted, demonstrating that the effectiveness of the existing EEMs can be enhanced based on the conceptual approach proposed in Chapter 5. The results produced using the EMMCW verified that the validity, thoroughness and predictive power of EEMs for IMM can be improved, following the conceptual approach advocated.

7.3 Limitations of Thesis Research

Although the thesis has contributed useful research to the field of usability evaluation of instructional multimedia, there are important limitations of the research which need to be considered.

EEM Effectiveness Framework

Although the effectiveness framework comprises a comprehensive set of measurement criteria, it does not include one important factor - *redesign change effectiveness*. As discussed in Chapter 3, it is important for expert evaluations to be able to generate effective requirements for redesigning user interfaces. In addition, John and Marks (1997) propose studying the effectiveness of UI designs resulting from expert evaluations as a final measure of the effectiveness of the usability results produced using EEMs. Therefore, *redesign change effectiveness* needs to be examined together with the other EEM performance criteria. As noted in Chapter 3, it was not included in the framework as there were no adequate resources to measure it as part of this research.

The following are limitations with the approach to conducting the usability evaluations.

Selecting and Training Experts

During the initial EEM comparison study the MMT and the IMMC were applied by four evaluators, whereas the MMCW was used by two. The reason for this difference is that the creators of the MMCW specified that the method is suitable for multimedia designers and HCI specialists, and did not recommend it being used by instructional experts. The authors of the other two methods explicitly recommend the EEMs to be

used by a variety of experts, including not only MMDs, but also subject-matter and instructional specialists. Such experts were recruited to use these methods, however, the difference in the number of evaluators could explain the variation in the number of problems predicted by each method. Furthermore, in a study involving 19 evaluators who applied Heuristic Evaluation to a voice response system, Nielsen (1994) found that there is an accumulative effect in the number of problems predicted and that the bigger the number of experts, the larger the number of new problems identified. From these findings it appears that involving two to four experts in an evaluation would have constrained the *thoroughness* of the expert predictions. Therefore, if more evaluators had been recruited to use each EEM the *thoroughness* of their predictions might have been better.

Although none of the experts involved in both studies had prior knowledge of the other methods which were used, the MMDs and the HCI specialist had prior experience with usability evaluation and knowledge of traditional evaluation techniques, such as Heuristic Evaluation. Previous research has shown that not only the number of experts but also their expertise is of great importance for the quality of expert evaluations. For example, Nielsen (1994) compared the performance of novice evaluators, single experts (usability experts with no domain knowledge) and double experts (usability experts with domain expertise). This study showed that 15 novice evaluators are needed to identify 75% of the actual user problems, whereas two to three double experts can achieve the same results. As the evaluators who participated in the evaluations conducted in this thesis were primarily novices or single experts, and only two double experts were involved, that could have further contributed to the poor *thoroughness* results shown by the expert evaluations.

Although the experts familiarised themselves with the methods in their own time, they were all novices at the techniques, which meant that they were not performing as well as evaluators experienced with the EEMs. This could have had implications on the time taken to evaluate the application, as well as the actual evaluation results.

Another aspects to be acknowledged is that none of the reviewers had actually participated in the development of the IMM application which was evaluated. Therefore, they did not have extensive knowledge of its design rationale and the wider

context of its development. Although external experts do get contracted to perform usability evaluations, they would usually have the opportunity to discuss the system's requirements and specification with members of the development team, which the reviewers in both studies did not have. Thus, some of the comments they made stemmed from their misunderstandings of the design decisions made by the developers of the application, which were classed as false alarms by the instructional judges, who were familiar with the application and its rationale.

Finally, due to time constraints the experts only familiarised themselves with the Exponential Graphs section of the IMM application. This resulted in the experts in the first study not answering some of the checklist questions, due to lack of knowledge of the entire application. Furthermore, some comments were made about missing functionality which is actually provided in the software, but the reviewers did not have the time to explore it. Such invalid problems were excluded from the problem sets, as acknowledged in Chapter 4. No such comments were made in the second validation study.

Number of Users Involved in the User Tests

Only four users took part in the user tests conducted with the IMM application. As Nielsen and Landauer (1993) suggest, at least 15 users are required to discover all usability problems in a design. Using a statistical formula, the authors have identified that five users would be able to discover 75% of the existing usability problems, and four users will be able to identify around 70%. For this research, only four users could be recruited due to the practical constraints presented in Chapter 4. Due to the limited number of users involved, however, the user problem set may be incomplete. This would have had implications for the *validity* and *thoroughness* results of all EEMs studied, as the number of user problems identified during the user tests was taken as a maximum bound in the calculations of the validity and actual thoroughness of the EEMs. Therefore, had more users been involved in the user tests more user problems would have been identified. Had more user problem being identified through user testing, the *validity* of the EEMs examined could have been improved. At the same time their *actual thoroughness* could have been reduced, as more missed-out user problems could have been identified.

As highlighted in Chapter 4, the four students represented typical users of the Maths application, as they were attending a course in Mathematics requiring the use of the software during tutorials and for revision. However, as Woolrych and Cockton (2001) point out, the chances of getting the right users depends on the distribution of the individual differences between test users, the tool under test and the tasks performed during testing. Therefore, although care was taken to select four representative users, the small sample selected may not have represented the variety of skills and abilities of the students who are likely to use the software.

Environmental conditions of the usability evaluations

In order to keep environmental conditions uniform and prevent any distractions, the expert evaluation sessions were conducted in an isolated environment, where the experts were working on their own, which may not represent a typical usability evaluation situation where evaluators and designers may discuss design considerations.

Methodological limitations of validating and comparing EEMs

The extraction of user problems from the user test sessions captured on video tape was performed only by the researcher. As double-checking with another researcher was not possible, bias could have been introduced and errors may have been made in identifying and specifying user interaction problems. As Jacobsen and John (1998) acknowledge the detection of usability problems is subject to considerable individual variability. The authors point out that no single evaluator can detect all problems in a usability test, as the number of problems revealed depends on both the number of users and the number of evaluators. Therefore, the user problem sets may be partially incomplete or inaccurate due to the fact that only one researcher reviewed the data. As a result, the ratios of actual thoroughness of the EEMs may have been partially distorted.

Cockton and Lavery (1999) propose a systematic analytic framework for improving the outcomes of problem extraction, by distinguishing four phases: isolation of relevant episodes, analysis of relevant difficulties, causal analysis, and recommendation generation. In the proposed SUPEX framework, actions are segmented into episodes, difficulties are then filtered out and coded, the causes of sets of problems are then formulated, and finally recommendations are generated. The authors claim that the framework can ensure the extraction of more reliable user problem sets.

Similarly, the predicted and experienced problem sets were matched only by the researcher, although a set of matching rules and a scale for rating the certainty of each match were used to ensure the uniformity of the match. Lavery, Cockton and Atkinson (1997) acknowledge that the matching of predicted and actual problems requires judgement, and used two experts to perform problem matching to ensure objectivity of the process as no standard report format was used. It was not possible to involve a second expert to perform independent matching of the problems in the first study where no uniform problem report formats were used. Therefore, a certain degree of bias may have been introduced in the problem matching, which would have infringed on the accuracy of the problem comparison.

Finally, when validating the predictions made using the EMMCW, matching was done based on all components in the problem reports, including causes to predicted problems. Problem matching based on problem causes was only done for a small number of the comprehension problems predicted. This was done because the SMEs did not specify the likely outcomes that may occur as a result of certain design faults. They only specified their causes. Such incomplete matches would have marginally increased the ratio of the actual thoroughness of the EMMCW. Therefore, extra measures for ensuring that expert problem reports are complete need to be introduced in order to be able to make more accurate problem matching.

Limitations of the EMMCW

The main limitation of the developed EEM is that it is partial in its present form. The evaluation questions created cover primarily the design of text and still images, and combinations of the two. Evaluation criteria regarding sound and speech were not formulated, as only short animations and no sound are included in the section of the Maths application evaluated.

The use of only one IMM application in the user and expert evaluations limits the generality of the conclusions presented in this thesis. As Gray and Salzman (1998) emphasise, a UEM can perform differently when applied to different types of software applications and these differences need to be studied in order to be able to draw conclusions relevant to a broad range of products. Otherwise 'mono-method bias' will occur. Within the scope of this thesis, evaluations of more than one IMM product were

not possible due to the limited time and resources that were available. Therefore, the conclusions presented in the previous chapters can only reliably be applied to a type of IMM software similar to the one studied.

7.4 Future Directions

There are three major possible directions for further practical and theoretical research. Work has already been done on producing a unified and comprehensive framework for design and evaluation of IMM, which is discussed in this section. A possible practical direction is to extend and further improve the EMMCW method developed as part of this research, and to conduct further studies validating its effectiveness.

7.4.1 Development of a Theoretical Framework of Learning with IMM

Chapters 5 and 6 argue that the successful prediction of comprehension and other learning problems requires a sound theoretical framework of how people learn with IMM. Faraday and Sutcliffe's (1997) framework was criticised for not dealing with aspects of interaction.

Furthermore, in order to improve the predictive power more research is required into how students build mental models of multimedia representations, and how the major factors of IMM design - the content, the instructional approach adopted, the learner and their cognitive processes, the context of use and the multimedia application – all relate to each other.

20 years ago Bates (1981) argued that an adequate theory of instructional media does not exist. The author adds that even the more general theories of learning have rarely formed the basis of experimentation of the effectiveness of instructional media. Theories of learning have further been developed, such as the one proposed by Tergan (1998) which was only presented at its highest level, and was incomplete. Despite these attempts, no coherent theoretical framework has been formulated which can be utilised effectively for the formative evaluation of IMM.

Work on a multidimensional framework for the formative evaluation of IMM was started as part of this thesis, however could not be completed within the time scale available.

The framework is based on the belief that the effective design of IMM user interfaces is the intersection of four main components – the learner and their unique characteristics, the subject matter content, the instructional approach adopted and the context of learning. Thus, it contains five dimensions – the learner, the content, the instructional approach, the context of use and the MM user interface, as can be seen from Figure 7.2. In order to understand the interrelationship between the individual dimensions we first broke them down into their compound attributes. For instance, the *Learner* dimension comprises a number of demographic, affective, cognitive and physiological characteristics of the target audience. After that, the interdependencies between the attributes were presented in the form of rules. Each rule describes the associations between relevant attributes of each dimension for the design of a particular aspect of the user interface. One rule establishes that to assess whether the most appropriate media resources have been selected the information types (*Content* dimension), the learner age, prior knowledge in the domain and their input learning style (*Learner* dimension) have to be considered.

This framework needs further development and validation.

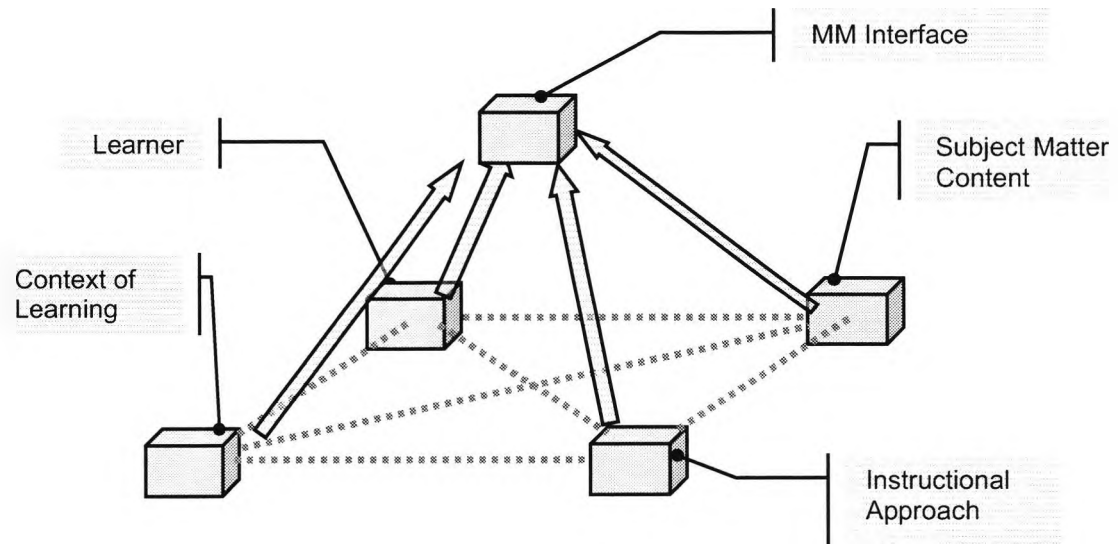


Figure 7.2: A Multidimensional Framework for the Evaluation of IMM

7.4.2 Further Improvement of the EMMCW

There are several areas in which the developed evaluation method could be improved in future work. The first issue of interest is how user activities can be studied more comprehensively. Although the EMMCW requires evaluators to focus on how concepts important for the user learning are represented in the IMM design, they are not encouraged to study specific learning activities. For the assessment of interactive activity-based or problem-based IMM environments, it is important to assess how different learning activities are facilitated by the IMM, and whether they adequately support effective learning (Jonassen and Rohrer-Murphy, 1999). Such an approach would require a list of typical learning activities to be created, and decomposed then into steps and tasks, after which they need to be assessed against the interaction strategies and dialogue options provided by the interactive IMM environment. This then allows a more thorough assessment of the pedagogical effectiveness of interactive IMM.

Secondly, the EMMCW can be extended to include evaluation questions to assess the quality of the design of audio and moving images, and combinations of such media, which are not included in the present version of the method. A complete set of questions would allow a more thorough assessment of IMM applications, which utilise such media resources for representing subject matter content. Further evaluations with other types of IMM products would prevent mono-method bias (Gray and Salzman, 1998).

Finally, the problem report formats which evaluators were asked to use for describing usability problems could be improved. Future problems that occurred with the problem report used in the study presented in Chapter 6 could be resolved by combining the behavioural and performance outcome fields into one field called "outcome", as suggested by Lavery and Cockton (1997). In this way, confusion between the similarity between the two types and the need to separate them can be avoided. Furthermore, an extra field needs to be added to encourage evaluators to provide redesign suggestions of the design faults in order to rectify usability problems. The evaluators were not asked to specify redesign suggestions during the second study as examining the effectiveness of redesign suggestions was outside the focus of the hypotheses and was not feasible, as explained in section 7.3. Apart from the problem report format, it is proposed that evaluators need to be provided with further support in deciding on the most appropriate

redesign solutions. Suggestions of how this could be achieved are presented in the next sub-section.

7.4.3 Further EEM Effectiveness Improvement Studies

The results from the validation of the EMMCW suggested that further research is required in order to improve the prediction rate of learning and comprehension problems, and to explore the type of support evaluators need to be able to infer such problems. Theoretical frameworks such as those proposed in sub-section 7.4.1 can provide conceptual ground for building effective evaluation support. Other instructional theories such as the Dual Coding theory by Clark and Paivio (1991) and the constructivist paradigm of learning (Cunningham, Duffy and Knuth, 1993) can also be used. They might provide evaluators with necessary understanding of how learners build mental models of multimedia representations, and how the major factors of IMM design - the content, the instructional approach adopted, the learner and their cognitive processes, the context of use and the multimedia application – all relate to each other. Furthermore, specific instructions as to how to use the rationale provided in such frameworks for reasoning about the effectiveness of IMM design is also needed.

To further improve the prediction rate of behavioural outcomes, patterns of how cognitive breakdowns influence the likely user behaviour can be identified to further support the prediction process. As discussed in Chapter 5, the user's behaviour following cognitive breakdowns is dependent on the severity of the breakdown, which influences the user response strategy. Thus, these factors need to be considered by usability evaluators when determining the likely user behaviour following a breakdown.

To further enhance the prediction rate of performance outcomes evaluators need to be encouraged to consider the relationship between behavioural and performance outcomes together. As outlined in Chapter 5, behavioural and performance outcomes are very tightly dependent on each other, i.e. what the user will do as a result of a breakdown will determine whether they will be able to recover and still achieve their goals or not. Therefore, there is strong dependence between the two, and the severity of the first directly influences the outcome of the second.

Finally, as discussed in Section 7.3, an effective EEM should be able to generate requirements for effective redesign of the application (Dutt, Johnson and Johnson, 1994). Furthermore, Mayo and Hartson (1993) outline a set of desirable characteristics for EEMs, one of which is *Design Support*. The authors specify that to promote more direct design support, expert evaluation techniques should be able to produce design-redesign requirements. It is proposed that instructional and MM design guidelines can be integrated into the method to provide such support. The effect of implementing such guidelines on the generation of effective redesign solutions then needs to be studied. Such a study would involve implementing the redesign suggestions into the design of IMM and measuring the effectiveness of the redesigned software with representative users, including its pedagogical effectiveness.

List of Abbreviations Used in the Thesis

EMMCW	- Education Multimedia Cognitive Walkthrough
EM	- Evaluation Method
HCI	- Human-Computer Interaction
ICT	- Information and Communication Technologies
IMM	- Instructional Multimedia
IMMC	- Interactive Multimedia Checklist
MM	- Multimedia
MMCW	- Multimedia Cognitive Walkthrough
MMD	- Multimedia Designer
MILS	- Multimedia Interactive Learning System
MMT	- Multimedia Taxonomy
MMUI	- Multimedia User Interface
PD	- Problem Description
SME	- Subject-matter Expert
UEM	- Usability Evaluation Method

Glossary

Actual thoroughness - measures what proportion of the real problems that exist in an application have been identified by expert evaluations.

Actual validity - measures the ability of an evaluation method to predict valid user problems.

Analytic evaluation method - see expert evaluation method.

Comparative thoroughness - measures the types of problems identified by different evaluation methods to establish the comprehensiveness of the focus of each method in comparison to the others.

Comprehension problem - a difficulty experienced by users which occurs as a result of the Instructional Multimedia (IMM) application making a demand on the users' cognitive processes that hampers their knowledge construction processes.

Cost effectiveness - measures the ratio of the time spent using each evaluation method to the severity of the problems identified by experts minus the number of false alarms identified.

Diagnostic power - a measure of the ability of an expert evaluation method to support the identification of cause-and-effect relationships between intrinsic design features and their outcomes on users' behaviour and performance.

Direct context consideration - is a property of the measure of Tailorability to Context. Occurs when the evaluation criteria or other aspects of an EEM make an explicit reference to particular aspects of the learning context, and indicate what the implications of certain design features could be for the potential usage of the application.

Ease of use of an EEM - assesses how easy to use an evaluation method is for those applying it in a particular evaluation situation.

Empirical evaluation method - see user evaluation method.

Experienced problem - a problem identified during user tests.

Expert evaluation method (EEM) - an evaluation method which involves experts to reason about the effectiveness of a particular application.

Expert specialisation - measures whether an evaluation method is adequately tailored towards different specialists and whether it takes into account the various perspectives such specialists bring into the evaluation can.

False alarm - a problem predicted during expert evaluations which is not experienced by users and is not considered valid by instructional judges.

Falsity - measures the proportion of false alarms predicted during an expert evaluation.

Fully specified problem (FS) - describes the nature of a design problem, its cause and its effect on users' behaviour or performance.

Incompleteness - measures the proportion of actual user problems missed out by expert evaluations.

Indirect context consideration - is a property of the measure of Tailorability to Context. Occurs when the evaluation criteria or other aspects of an EEM prompt experts to look into factors in the learning context, but leave it to their expertise to judge what implications these factors may have for learners and the achievement of their goals.

Instructional Multimedia (IMM) - is the combination of text, audio, still and moving images delivered via an electronic system to support learners in achieving their learning goals

Instructional problem - a problem related to the subject matter or the instructional approach implemented in an IMM design, which was identified by experts during expert evaluations.

Invalid problem - a problem identified as invalid by instructional judges.

Learnability of an EEM - assesses the amount of training required to achieve a reasonable level of efficient and effective error prediction during expert evaluations. It is measured by the familiarisation time and the subjective assessment of the ease of acquisition of the EEM by evaluators.

Learning problem - a problem experienced by users or identified by experts during evaluations of education software, which is concerned with the ease of acquiring new knowledge, the content of or the instructional approach adopted in the education application.

Multimedia (MM) - the combination of a variety of representational media, such as text, graphics, sound, speech, animation and video to deliver information to the user via an electronic system.

No context consideration - is a property of the measure of Tailorability to Context. Occurs when the evaluation criteria or other aspects of an EEM do not refer to any aspects of the context in which the design would be used.

Predicted problem - a problem identified during an expert evaluation.

Quality of documentation - a measure of how well documented an EEM is.

Redesign suggestion - a problem description which only suggests a redesign solution without specifying what the usability or learning problem is.

Reliability - measures whether evaluators consistently provide the same results under the same evaluation conditions.

Theoretical validity - measures the extent to which evaluation criteria of an EEM are grounded in empirical evidence and relevant theoretical work.

Under-specified problem, no cause specified (US-NC) - a problem stating that there is a design problem without relating it to a particular design feature/s that may cause it.

Under-specified problem, no effect specified (US-NE) - a problem describing the cause without specifying the potential effect a design problem may have on users.

Usability evaluation method (UEM) - an evaluation method aiming to assess the usability of software applications. It can either involve various kinds of reviews and inspections of different aspects of user interface design, or is based on formal analysis of user interfaces or performance prediction.

User evaluation method - an evaluation method which involves potential users to provide feedback and identify usability problems, and evaluators to reason about the causes of the problems and devise redesign suggestions.

User interaction problem - a difficulty occurring during the user dialogue with the user interface, which typically has an immediate effect on their behaviour or performance, which hinders or prevents them from successfully achieving a part of their task.

Valid problem - a problem predicted during expert evaluations which was experienced by users and was considered as a valid instructional problem by independent judges although not experienced by users.

References

Ainsworth, S. (1999) The functions of multiple representations. *Computers & Education,* 33, pp 131-152.

Alexander, S. and Hedberg, J.G. (1994) Evaluating technology-based learning: Which model? In K. Beattie, C. McNaught and S. Wills (Eds.) *Interactive Multimedia in Education: Designing for Change in Teaching and Learning*. Holland: Elsevier Science B.V.

Bailey, B. (1999) Heuristics Evaluations. *User Interface Design Update Newsletter,* May 1999. Available at: http:/www.humanfactors.com/library/may992.htm

Bailey, R.W., Allan, R.W. and Raiello, P. (1992) Usability testing vs. heuristic evaluation: A head-to-head comparison. In *Proceedings of the Human Factors Society 36th Annual Meeting*. Santa Monica, CA, pp 409-413.

Barker, P. and King, T. (1993) Evaluating Interactive Multimedia Courseware – A Methodology. *Computers in Education*, 21(4), pp 307-319.

Bastien, J.M.C. and Scapin, D.L. (1995) Evaluating a User Interface With Ergonomic Criteria. *International Journal of Human-Computer Interaction*, 7, pp 105-121.

Bates, T. (1981) Towards a Better Research Framework for Evaluating the Effectiveness of Educational Media. *British Journal of Educational Technology,* 12 (3), October 1981, pp 215- 233.

Bell, D. & Johnson, P. (1993) Multimedia support for cooperative working. In *Proceedings of the ERCIM Workshop on Multimedia Systems.* Nancy, France 2-4th.

Bloom, B.S., Engelhart, M.D., Furst, E.J., Hill, W.H. and Krathwohl, D.R. (1956) *Taxonomy of Educational Objectives*. New York: David McKay Company, Inc.

Byrum, D.C. (1992) Formative Evaluation of Computer Courseware: An Experimental Comparison of Two Methods. *Journal of Educational Computing Research*, 8(1), pp 69-80.

Campbell, D.T. and Stanley, J.C. (1966) *Experimental and quasi-experimental design for research.* Chicago: Rand McNally.

Chan, T.S. and Ahern, T.C. (1999) The Importance of Motivation: Integrating Flow Theory into Instructional Design. In *Proceedings of SITE '99*, AACE Publications, pp 78-82.

Cheng, P.C. -H. (1999) Unlocking conceptual learning in mathematics and science with effective representational systems. *Computers & Education,* 33, pp 109-130.

Chi, M.T.H, Feltovich, P.J. and Glaser, R. (1981) Categorisation and representation of physics problems by experts and novices. *Cognitive Science*, 5, pp 121-152.

Clark, R.E. (1994) Media Will Never Influence Learning. *Journal of Educational Technology, Research and Development,* 42(2), pp 21-29.

Clark, J.M. and Paivio, A. (1991) Dual Coding Theory and Education. *Educational Psychology Review*, 3, pp 149-210.

Cockton, G. and Lavery, D. (1999). A Framework for Usability Problem Extraction. In M. A. Sasse and C. Johnson (Eds.), *Proceedings of Interact '99*, IOS Press: The Netherlands, pp. 344-352.

Cockton, G. and Woolrych, A. (2001) Understanding Inspection Methods: Lessons from an Assessment of Heuristic Evaluation. In A. Blandford, J. Vanderdouckt and P. Gray (Eds.), *People and Computers XV - Interaction Without Frontiers,* Springer, pp 171-191.

Comer, P.G. and Geissler, C. (1998) A Methodology for Software Evaluation. In *Proceedings of 9th International Conference of the Society for Information Technology and Teacher Education (SITE '98),* 10-14 March 1998, Washington, DC, USA, volume 1.

Craig, Mehrens and Clarizio (1975) *Contemporary Educational Psychology: Concepts/Issues/ Applications.* John Wiley & Sons, Inc.

Crane, V. and Isaacson, S. (1990) In Search of a Methodology: Evaluating "SCOOP", the WGBH Teletex Magazine for High School Students. In B.N. Flagg (ed.) *Formative Evaluation for Educational Technologies,* Lawrence Erlbaum Associates: Hillsdale, New Jersey, pp 113- 160.

Crozat, S., Hu, O. and Trigano, P. (1999) A Method for Evaluating Multimedia Learning Software. In *Proceedings of IEEE Multimedia Systems 1999*, June 7-11, Florance, Italy, vol. 1, pp 714-719.

Cunningham, D.J., Duffy, T.M. and Knuth, R.A. (1993) The Texbook of the Future. In C. McKnight, A. Dillon and J. Richardson (eds.) *Hypertext: a psychological perspective,* Ellis Horwood, pp 19- 49.

Desurvire, H.W. (1994) Faster, Cheaper!! Are Usability Inspection Methods as Effective as Empirical Testing? In *Usability Inspection Methods,* J. Nielsen and R.L. Mack (eds.). John Wiley and Sons, Inc, pp. 173-202.

Desurvire, H., Lawrence, D. and Atwood, M. (1991) Empiricism versus judgement: Comparing user interface evaluation methods on a new telephone-based interface. *ACM SIGCHI Bulletin*, 23, 4 (October). Cambridge: Cambridge University Press, pp 58-59.

Desurvire, H. W., Kondziela, J.M. and Atwood, M. E. (1992) What is gained and lost when using evaluation methods other than empirical testing. In A. Monk, D. Diaper, and M.D. Harrison (Eds), *People and Computers VII*, Cambridge University Press: Cambridge, UK, pp 89-102.

Dimitrova, M. and Sutcliffe, A. (1999) Designing Instructional Multimedia Applications: Key Practices and Design Patters, In *Proceedings of ED-MEDIA '99*, AACE Publications, Seattle, Washington, USA, June 19-24, pp 358-363.

Doubleday, A., Ryan, M., Springett, M. and Sutcliffe, A. (1997) A comparison of usability techniques for evaluating design. In *Proceedings of Designing Interactive Systems: Processes, Practices, Methods and Techniques Conference '97*, 18-20 August 1997, Amsterdam, The Netherlands, pp 101-109.

Draper, S.W., Brown, M.I., Henderson F.P. and McAteer, E. (1996) Integrative Evaluation: An Emerging Role for Classroom Studies of CAL. *Computers and Education*, 26(1-3), pp 17-32.

Dutt, A., Johnson, H. and Johnson, P. (1994) Evaluating Evaluation Methods. In G. Cockton, S.W. Draper and G.R.S. Weir (Eds.), *Proceedings of HCI '94, People and Computers IX*, Glasgow, pp 109-121.

Erricsson, K.A. and Simon, H.A. (1984) *Protocol Analysis*. Cambridge, Massachusetts: Massachusetts Institute of Technology Press.

Faraday, P. (1997) *Theory Based Design and Evaluation of Multimedia Presentation Interfaces*. PhD Thesis, March 1997.

Faraday, P. and Sutcliffe A. (1996) An Empirical Study of Attending and Comprehending Multimedia Presentations. In *Proceedings of ACM Multimedia '96,* Boston, USA, November 18-22, pp 265-275.

Faraday, P. and Sutcliffe A. (1997a) Multimedia: Design for the 'Moment'. In *Proceedings of ACM Multimedia '97,* Seattle, WA, USA, November 9-13, pp 183-192.

Faraday, P. and Sutcliffe, A. (1997b) Evaluating Multimedia Presentations. *The New Review of Hypermedia and Multimedia,* 3, pp 7-37.

Fisher, K.M. and Lipson, J.I. (1986) Twenty Questions about Student Errors. *Journal of Research in Science Teaching,* 23 (9), pp 783-803.

Fitzpatrick, R. and Dix, A. (1999) A Process of Appraising Commercial Usability Evaluation Methods, In *Proceedings of International Conference on Human-Computer Interaction (HCI International) '99, v.1, Ergonomics and User Interfaces,* pp. 1068-1072.

Flagg, B.N. (1990) *Formative Evaluation for Educational Technologies*. Hillsdale, NJ: Lawrence Erlbaum, 1990.

Foshay, R. (1999) "But Does It Work?" 6 Ways to Evaluate Technology. In *Proceedings of the 11th World Conference on Educational Multimedia, Hypermedia and Telecommunications (ED-MEDIA '99),* pp 822-827.

Georganas, N.D. (1997) Multimedia Applications Development: Experiences. *Multimedia Tools and Applications, 4*(3), pp 313-332.

Giouvanakis, T., Samaras, H. and Tarabanis, K. (2001) Designing a Pedagogically Sound Web-Based Interface: The Critical Role of Prior Knowledge. In *Proceedings of the 13th Annual World Conference on Educational Multimedia, Hypermedia and Telecommunications (Ed-Media) 2001,* Tampere, Finland, 25-30 June 2001, pp 592-597.

Gray, W.D. and Salzman, M.C. (1998) Damaged Merchandise? A Review of Experiments that Compare Usability Evaluation Methods. *Human-Computer Interaction*, 13, pp 203-261.

Hayden, M. and Speedy, G. (1995) Evaluation of the 1995 National Teaching Development Grants. Project commissioned by the Committee for the Advancement of University Teaching. Lismore, Australia: Southern Cross University.

Heller, R.S. (1996) Evaluating Interactive Multimedia. In K. Shields (Ed.), *Thomson Multimedia Resource,* Thomson Technology Services Group (Wadsworth Imprint), pp 65-87.

Heller, R.S. and Martin, C.D. (1999a) Multimedia Taxonomy for Design and Evaluation. In B. Furht (Ed.), *Handbook of Multimedia Computing*, CRC Press, pp 3-16.

Heller, R.S. and Martin, C.D. (1999b) Using a Taxonomy to Rationalise Multimedia Development. In *Proceedings of IEEE Multimedia Systems '99*, Firenze, Italy, 7-11 June.

Heller, R.S. and Martin, C.D. (1999c) Using a Theoretical Multimedia Taxonomy Framework. In *Proceedings of ED-MEDIA 1999*, June 1999, USA, AACE Press, pp 415-420.

Henderson, R., Podd, J., Smith, M. and Varela-Alvarez, H. (1995) An examination of four user-based software evaluation methods. *Interacting with Computers*, v. 7 (4), pp. 412-432.

Hertzum, M. and Jacobsen, N.E. (1999) The Evaluator Effect during First-Time Use of the Cognitive Walkthrough Technique. In *Proceedings of HCI International, Ergonomics and User Interfaces*, v.1, pp 1063 – 1067.

Honebein, P., Duffy, T.M. and Fishman, B. (1993) Constructivism and the design of learning environments: Content and authentic activities for learning. In T.M. Duffy, J. Lowyck and D. Jonassen (Eds.) *Designing environments for constructivist learning.* Heidelberg: Springer-Verlag.

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

Jacobsen, N.E. and John, B.E. (1998) The Evaluator Effect in Usability Studies: Problem Detection and Severity Judgements. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, pp 1336-1340.

Jacobson, M.J. and Spiro, R.J. (1994) A framework of context-analysis of technology-based learning environments. *Journal of Computing in Higher Education,* 2 (5), pp 3-32.

Jeffries, R., Miller, J.R., Wharton C. and Uyeda, K.M. (1991) User Interface Evaluation in the Real World: A Comparison of Four Techniques. In *Proceeding of Computer-Human Interaction '91*, 27 April- 2 May 1991, New Orleans, Lousiana, S.P. Robertson, G.M. Olson and J.S. Olson (eds.), ACM Press, New York, NY, pp 119-124.

Jeffries, R. (1994) Usability Problems Reports: Helping evaluators communicate effectively with developers. In J. Nielsen and R.L. Mack (Eds.) *Usability Inspection Methods.* John Wiley and Sons, New York, pp 273-294.

Jih and Reeves (1992) Mental models: a research focus for interactive learning systems. *Educational Technology, Research and Development*, 40(3), pp 39-53.

John, B.E. and Marks, S.J. (1997) Tracking the Effectiveness of Usability Evaluation Methods. *Behaviour and Information Technology*, 16(4/5), pp 188-202.

Johnson-Laird, P.N. (1983) *Mental Models*. Cambridge, MA: Harvard University Press.

Jonassen, D. (1998) Designing Constructivist Learning Environments. In C.M. Reigeluth (ed.), *Instructional Theories and Models*, 2nd Edition. Mahwah, NJ: Lawrence Erlbaum.

Jonassen, D.H. and Rohrer-Murphy, L. (1999) Activity Theory as a Framework for Designing Constuctivist Learning Environments. *Journal for Educational Technology, Research and Development,* 47(1), pp 61-79.

Kalyuga, S., Chandler, P. and Sweller, J. (1998) Levels of experience and instructional design. *Human Factors*, 40, pp 1-17.

Kalyuga, S., Chandler, P. and Sweller, J. (2000) Incorporating Learner Experience Into the Design of Multimedia Instruction. *Journal of Educational Psychology*, 92(1), pp 126-136.

Karat, C.-M., Campell, R. and Fiegel, T. (1992) Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of CHI '92*, pp 397-404.

Kennedy, G., Petrovic, T. and Keppell, M. (1998) The Development of Multimedia Evaluation Criteria and a Program of Evaluation for Computer Aided Learning. In R.M. Corderoy (ed.) *Proceedings of the fifteenth annual conference of Australian Society for Computers in Tertiary Education (ASCILITE '98)*, University of Wollongong, Australia, pp 407-415,.

Kennedy, G.E. (1999) Defining the Dimensions of a Formative Evaluation Program: A Multi-Method, Multi-Perspective Approach to the Evaluation of Multimedia. In *Proceedings of ED-MEDIA '99*, AACE Publications, pp 604-609.

Kirsh, D. (1997) Interactivity and multimedia interfaces. *Instructional Science*, 25, pp 79-96.

Kirwan, B. (1992) Human error identification in human reliability assessment. Part 2: Detailed comparison of techniques. *Applied Ergonomics*, 23, pp 371-381.

Koedinger, K.R. and Anderson, J.R. (1990) Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science,* 14, pp 511-550.

Kozma, R.B. (1991) Learning with Media. *Review of Educational Research,* 61(2), Summer 1991, pp 179-211.

Kozma, R.B. (1994) Will Media Influence Learning? Reframing the Debate. *Journal of Educational Technology, Research and Development,* 42(2), pp 7-19.

Large, A., Beheshti, J., Breuleux, A. and Renaud, A. (1994) Multimedia and Comprehension: A Cognitive Study. *Journal of the American Society for Information Science*, 45 (7), pp 515-528.

Large, A., Beheshti, J., Breuleux, A. and Renaud, A. (1995) Multimedia and Comprehension: The Relationship among Text, Animation, and Captions. *Journal of the American Society for Information Science*, 46(5), pp 340-347.

Large, A., Beheshti, J., Breuleux, A. and Renaud, A. (1996) Effect of Animation in Enhancing Descriptive and Procedural Texts in a multimedia Learning Environment. *Journal of the American Society for Information Science,* 47(6), pp 437-448.

Laurillard, D. (1993) *Rethinking Usability Teaching: a framework for effective use of educational technology.* London: Routledge.

Lavery, D. and Cockton, G. (1995). A Pilot Study of Early Usability Evaluation Methods for Software Visualisation. FIDE Technical Report FIDE/95/141, University of Glasgow.

Lavery, D. and Cockton, G. (1996). Iterative Development of Early Usability Evaluation Methods for Software Visualisations. In W.D. Gray and D. A. Boehm-Davis (Eds), *Proceedings of the 6th Workshop of Empirical Studies of Programmers*, Ablex, pp 275-276.

Lavery, D. and Cockton, G. (1997) Representing Predicted and Actual Usability Problems. In H. Johnson, P. Johnson, and E. O'Neill (Eds.), *Proceedings of International Workshop on Representations in Interactive Software Development,* Queen Mary and Westfield College, University of London, pp 97-108.

Lavery, D., Cockton, G. and Atkinson, M. P. (1997) Comparison of Evaluation Methods Using Structured Usability Problem Reports, *Behaviour and Information Technology*, 16 (4/5), pp 246-266.

Lee, A.Y., Gillan, D.J., Upchurch, E.E., Melton, J.S. and Harrison, C.L. (1995) Is multimedia-based training effective? Yes and no. In *Proceedings of International*

Conference on Computer-Human Interaction (CHI '95), Denver, Colorado, USA, May 7-11, pp 111-112.

Leedy, P.D. (1993) *Practical Research: Planning and Design.* Fifth Edition. New York: MacMillan Publishing.

Linn, M.D. (1992) The art of Multimedia and the State of Education. *Educational Researcher*, 1992, 21(1), pp 30-32.

Mack, R. and Montaniz, F. (1994) Observing, Predicting and Analyzing Usability Problems. In J. Nielsen and R.L. Mack (Eds.) *Usability Inspection Methods.* John Wiley and Sons, New York, pp 295-339.

Mayer, R.E. (1997) Multimedia learning: Are we asking the right questions. *Educational Psychologist,* 32, pp 1-19.

Mayer, R.E. and Anderson, R.B. (1992) The Instructive Animation: Helping Students Build Connections Between Words and Practices in Multimedia Learning. *Journal of Educational Psychology*, 84(4), pp 444-452.

Mayer, R.E. and Moreno, R. (1998) A Split-Attention Effect in Multimedia Learning: Evidence for Dual Processing Systems in Working Memory. *Journal of Educational Psychology,* 90 (2), pp 312-320.

Mayer, R.E., Moreno, R., Boire, M. and Vagge, S. (1999) Maximizing Constructivist Learning From Multimedia Communications by Minimizing Cognitive Load. *Journal of Educational Psychology,* 91 (4), pp 638-643.

Mayo, K.A. and Hartson, H.R. (1993) Synthesis-Oriented Situational Analysis in User Interface Design. In *Proceedings of East-West International Conference on Human-Computer Interaction '93 (EWHCI '93),* v. 2, pp 54-68.

Moreno, R. and Mayer, R.E. (1999) Cognitive Principles of Multimedia Learning: The Role of Modality and Contiguity. *Journal of Educational Psychology,* 91(2), pp 358-368.

Morris, J.M., Owen, G.S. and Fraser, M.D. (1994) Practical Issues in Multimedia User Interface Design for Computer-Based Instruction. In S. Reisman (Ed.), *Multimedia Computing: Preparing for the 21st Century.* IDEA Group Publishing, pp 225-284.

MUMMS (1996) Measuring the Usability of Multi-Media Systems. The MUMMS Questionnaire. Human Factors Research Group, University College Cork, Ireland. 1996, Ver. 2.

Najjar, L.J. (1996) Multimedia Information and Learning. *Journal of Educational Multimedia and Hypermedia,* v.5, pp 129-150.

Nielsen, J. (1993) *Usability Engineering.* Academic Press, Inc: Boston, USA.

Nielsen, J. (1994) *Usability Engineering.* Expanded Edition, Morgan Kaufmann, San Francisco.

Nielsen, J. (1999) Severity Ratings for Usability Problems, Available at: http://www.useit.com/papers/heuristic/severityrating.html

Nielsen, J. and Landauer, T.K. (1993) A Mathematical Model of the Finding of Usability Problems. In *Proceedings of ACM INTERCHI '93 Conference,* Amsterdam, The Netherlands, 24-29 April 1993, pp 206-213.

Nielsen, J. and Phillips, V.L. (1993) Estimating the Relative Usability of Two Interfaces: Heuristic, Formal, and Empirical Methods Compared. In *Proceedings of INTERCHI '93*, pp 214-221.

Norman, D.A. and Spohrer, J.C. (1996) Learner-Centered Education. *Communications of the ACM,* 39 (4), April 1996, pp 25-27.

Northrup, P.T. (1995) Concurrent formative evaluation: Guidelines and implications for multimedia designers. *Educational Technology*, 35, Nov/Dec 1995, pp 24-31.

Oppenheim, A.N. (1992) *Questionnaire Design, Interviewing and Attitude Measurement*. Pinter Publishers: London, UK.

Passig, D. and Levin, H. (2001) The interaction between gender, age, and multimedia interface design. *Education & Information Technologies*, 6(4), December 2001, pp 241-250.

Peterson, M.G. (1998) Towards Usability Evaluation of Multimedia Applications. Available at: http://info.acm.org/crossroads/xrds4-4/usability.html

Pham, B. (1998) Quality evaluation of educational multimedia systems. *Australian Journal of Educational Technology*, 1998, 14(2), pp 107 – 121.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. and Carey, T. (1994) *Human Computer Interaction.* Addison-Wesley Publishing Company.

Preece, J. and Shneiderman, B. (1995) Survival of the Fittest: The Evolution of Multimedia User Interfaces. *ACM Computing Surveys,* 27 (4), December 1995, pp 558-559.

Reeves, T.C. (1992) Evaluating Interactive Multimedia. *Educational Technology*, May 1992, pp 47-53.

Reeves, T.C. (1993) Evaluating Interactive Multimedia. In D.M. Gayeski (Ed.), *Multimedia for Learning: Development, Application, Evaluation.* Educational Technology Publications Inc., New Jersey, 1993, pp 97-112.

Reeves, T.C. and Harmon, S.W. (1994) Systematic Evaluation Procedures for Interactive Multimedia for Education and Training. In S. Reisman (Ed.) *Multimedia Computing: Preparing for the 21st Century*, IDEA Group Publishing, pp 472-505.

Reiser, R.A. and Kegelman, H.W. (1994) Evaluating instructional software: A review and critique of current methods. *Educational Technology, Research and Development*, 42(3), pp 63-69.

Rogers, Y. and Scaife, M. (1999) How can Interactive Multimedia Facilitate Learning? In *Proceedings of 1st International Workshop on Intelligence and Multimodality in Multimedia Interfaces*, AAAI Press, Menlo Park, CA, 1999, pp .

Rowley, D.E. and Phoades, D.G. (1992) The Cognitive Jogthrough: A Fast-Paced User Interface Evaluation Procedure. In *Proceedings of ACM CHI '92,* Monterey, California, USA, 3-7 May, 1992, pp 289-395.

Salomon, G. (1978) *Interaction of media, cognition, and learning.* San Francisco: Jossey-Bass.

Saroyan, A. and Geis, G.L. (1988) An analysis of guidelines for expert reviewers. *Instructional Science*, 17, pp 101-128.

Scaife, M. and Rogers, Y. (1996) External Cognition: how do graphical representations work? *International Journal of Human-Computer Studies,* 45, pp 185-213.

Schelin, E. (2001) A New Model for Education. *E-Learning Magazine Online,* May 2001. Available at:
http://www.elearningmag.com/issues/may01/onlineuniversity.asp

Schueckler, L.M. and Shuell, T.J. (1989) A Comparison of Software Evaluation Forms and Reviews, *Journal of Educational Computing Research,* 5(1), pp 17-33.

Sears, A. (1997) Heuristic Walkthroughs: Finding the Problems Without the Noise. *International Journal of Human-Computer Interaction*, 9(3), pp 213-234.

Sears, A. and Hess, D.J. (1999) Cognitive Walkthrough: Understanding the effect of task description detail on evaluator performance. *International Journal of Human-Computer Interaction,* 11(3), pp 185-200.

Shackel, B. (1990) Human Factors and Usability. In J. Preece and L. Keller (Eds) *Human-Computer Interaction*, Hemel Hempstead: Prentice-Hall.

Shneiderman, B. (1998) *Designing the User Interface: Strategies for Effective Human-Computer Interaction.* Addison Wesley, 3rd Edition.

Shuell, T. (1988) The role of the student in learning from instruction. *Contemporary Educational Psychology*, 13, pp 276-295.

Sims, R. (1997) Interactivity: A Forgotten Art? *Computers in Human Behaviour*, 13(2), pp 157-180.

Sims, R. (1999) Interactivity on stage: Strategies for learner-designer communication. *Australian Journal of Educational Technology*, 15(3), pp 257-272.

Soloway, E. (1998) The Need: Moving Beyond Ease of Use To Supporting Learning. *Human Factors in Computing Systems: Learner Centered Design Workshop, 1998*, ACM SIGCHI, Los Angeles, CA, USA.

Spencer, K. (1991) Modes, media and methods: the search for educational effectiveness. *British Journal of Educational Technology*, 22(1), pp 12-22.

Squires, D. (1999) Usability and Educational Software Design: Special Issue of Interacting with Computers. *Interacting with Computers*, 11, pp 463-466.

Squires, D. and McDougall, A. (1996) Software evaluation: a situated approach. *Journal of Computer Assisted Learning*, 12, pp 146-161.

Squires, D. and Preece, J. (1996) Usability and Learning: Evaluating the Potential of Educational Software. *Computers and Education*, 27(1), pp 15-22.

Squires, D. and Preece, J. (1999) Predicting quality in educational software: Evaluating for learning, usability and the synergy between them. *Interacting with Computers*, 11, pp 467-483.

Stanton, N.A. and Stevenage, S.V. (1998) Learning to predict human error: issues of acceptability, reliability and validity. *Ergonomics*, 41(11), pp 1737-1756.

Stead, R. (1990) Problems with learning from computer-based simulations: a case study in economics. *British Journal of Educational Technology*, 21 (2), pp 106-117.

Strauss, A.L. (1993) *Qualitative Analysis for Social Scientists.* Cambridge University Press: Cambridge.

Sumner, T. and Taylor, J. (1998) New Media, New Practices: Experiences in Open Learning Course Design. In *Proceedings of Chi '98,* April 18-23, Los Angeles, pp 432-439.

Sweller, J. (1999) *Instructional Design*. Melborne, Australia: ACER Press.

Tergan, S. (1998) Checklists for the Evaluation of Educational Software: Critical Review and Prospects. *Innovations in Education and Training International*, February 1998, 35 (1), pp 9-20.

Thornton, D. and Phillips, R. (1997) Evaluation. In *The Developer's Handbook to Interactive Multimedia: A Practical Guide for Educational Applications*. R. Phillips. Kogan Page, London, pp127 – 146.

Verdi, M.P., Kulhavy, R.W., Stock, W.A., Rittschof, K.A. and Johnson, J.T. (1996) Brief Research Report: Text Learning Using Scientific Diagrams: Implications for Classroom Use. *Contemporary Educational Psychology*, 21, pp 487-499.

Virzi, R.A. (1992) Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, 34, pp 457-468.

Vossen, P., Maguire, M., Graham, R. and Heim, J. (1997) Design Guide for Multimedia. *Telematic Applications Project IE 2016 Report,* 2nd Edition, June 1997.

Wild, M. & Quinn, C. (1998). Implications of Educational Theory for the Design of Instructional Multimedia. *British Journal of Educational Technology*, 29(1), pp 73-82.

Wharton, C., Rieman, J., Lewis, C. and Polson, P. (1994) The cognitive walkthrough method: a practitioner's guide. In J. Nielsen and R.L. Mack (Eds.): *Usability Inspection Methods*. New York: John Wiley, pp 105-140.

Woolrych, A. and Cockton, G. (2000) Assessing Heuristic Evaluation: Mind the Quality, not just Percentages. In S. Turner and P. Turner (Eds.), *Proceedings of HCI 2000*, 2, pp 35-36.

Woolrych, A. and Cockton, G. (2001) Why and When Five Test Users aren't Enough. In J. Vanderdonckt, A. Blandford and A. Derycke (Eds.), *Proceedings of Joint AFIHM-BCS Conference on Human-Computer Interaction 2001*, Volume 2, pp 105-108.

Woolrych, A. and Cockton, G. (2002) Testing a Conjecture based on the DR-AR Model of Usability Inspection Method Effectiveness. In *Proceedings of Human Computer Interaction 2002,* H. Sharp, P. Chalk, J. LePeuple, and J. Rosbottom (Eds), Volume 2, pp 30-33.

Yildiz, R. and Atkins, M. (1993) Evaluating Multimedia Applications. *Computers in Education*, 21(1/2), pp 133-139.

Zhang, Z., Basili, V. and Shneiderman, B. (1998) An Empirical Study of Perspective-Based Usability Inspection. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting*, pp 1346-1350.

Zhang, Z., Basili, V. and Shneiderman, B. (1999) Perspective-based Usability Inspection: An Empirical Validation of Efficacy. *Empirical Software Engineering,* 4 (1), March 1999, pp 43-69.

Bibliography

Dimitrova, M. (1999) Towards a Pattern Language for Instructional Multimedia Interface Design. In *Proceedings of INTERACT '99*, IFIP, Edinburgh, Scotland, UK, August 30 - September 3, pp 676-678.

Dimitrova, M. and Sutcliffe, A. (1999) Designing Instructional Multimedia Applications: Key Practices and Design Patterns. In *Proceedings of ED-MEDIA '99*, AACE Publications, Seattle, WA, USA, June 19-14, pp 358-363.

Sutcliffe, A. and Dimitrova, M. (1999) Patterns, Claims, and Multimedia. In *Proceedings of INTERACT '99*, IFIP, Edinburgh, Scotland, UK, August 30- September 3, pp 329-335.

Dimitrova, M. (2000) Usability Evaluation of Instructional Multimedia User Interfaces. In *Proceedings of HCI 2000*, volume 2, BCS, Sunderland, UK, September 5-8, pp137-139.

Dimitrova, M., Sharp H. and Wilson, S. (2001a) Are Experts Able to Predict Learner Problems during Usability Evaluations? In *Proceedings of Ed-Media 2001*, June 25-30, Tampere, Finland, pp 419-424.

Dimitrova, M., Sharp H. and Wilson, S. (2001b) Why Do Experts Predict Invalid Problems? An investigation into the Validity of Expert Evaluations of Instructional Multimedia Software. In *Proceedings of HCI International 2001*, August 5-10, New Orleans, USA, pp1247-1251.