

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

**COMPARISON OF THE MAIN METHODS FOR EVALUATING
THE USABILITY OF COMPUTER SOFTWARE**

**A thesis presented in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy in Psychology
at Massey University.**

Ronald Derek Henderson

1992

Abstract

The aim of this thesis is to examine the dominant computer software usability evaluation methods. Four evaluation methods (logged data, questionnaire, interview, and verbal protocol analysis) were used to evaluate three different business software types (spreadsheet, word processor, and database) using a between groups design, involving 148 individuals of both genders. When each evaluation method was examined individually, the results tended to support findings from previous research. Comparisons were made to examine the efficiency of each evaluation method, in terms of its ability to highlight usability problems (both between and within the evaluation strategy). Here support for the efficiency of the verbal protocol analysis method was found. The efficiency of using two evaluation methods was also examined, where it was found that no significant improvement was obtained over the verbal protocol analysis used by itself. A comparison addressing the practicality of using these methods was also conducted. It seems that each method has differing strengths and weaknesses depending on the stage of the evaluation. From these results a theory for the effectiveness of evaluation strategies is presented. Suggestions for improving the methods commonly used, are also made. The thesis concludes by discussing the software evaluation domain and its relationship to the wider evaluation context.

Acknowledgements

There have been many individuals who have significantly influenced my thoughts and behaviour throughout the development and refinement of this thesis. Each deserves my heartfelt gratitude.

Special mention must go to my family. In particular, my wife Karen, who has encouraged, supported and helped me throughout the research process. My parents, Derek and June, and my parents-in-law, Ray and Patricia, have also provided immeasurable support and encouragement throughout this period.

Special mention must also go to my supervisors, Doctors Mike Smith, John Podd and Hugo Varela. All have influenced and supported me significantly throughout the development of this thesis. My chief supervisor, Mike, has continually supported and encouraged me throughout both a significant period before undertaking this dissertation, and through the dissertation process. Mike has also been party to innumerable discussions on the nature of applied research, stimulating thought provoking enquiry on my part. John, in turn, has provided thoughtful discussion on the research process, and his insights have been both stimulating and beneficial. Deep gratitude must be expressed towards Hugo, who patiently coaxed me through the pains of learning to program in PASCAL, the virtues of log linear analysis, and trout fishing. I am deeply indebted to all; thank you.

Not least, gratitude must be expressed to each and all of the research subjects who undertook the study and provided the wealth of information that form the basis of this dissertation.

TABLE OF CONTENTS

Abstract	i
Acknowledgements	ii
Table of Contents	iii
List of Tables	vii
List of Figures	xi
List of Appendices	xiii
Chapter One: Introduction	1
1.1. The Software Industry	1
1.2. Problems in the Software Industry	2
1.3. The Reasons for Human Computer Interaction Problems	5
1.4. The Evaluation Process	9
 Chapter Two: Developing and Evaluating the Human-Computer	
Interface	15
2.1. User Based Development and Evaluation Strategies	15
2.2. Models for the Development of Usable Software	17
2.3. The Aims and the Individuals Included in the Evaluation	23
2.4. The Lack of Criterion Development in Software Evaluation	
Methods	27
 Chapter Three: The Usability Construct and the Problems of	
Operational Definitions and Measures	28
3.1. The Usability Construct	28
3.2. Evaluation Relevance, Deficiency, Contamination, and	
Redundancy	32
3.3. Operational Definitions	36
3.4. Analysis of Measures	37
 Chapter Four: Software Evaluation Methods	43
4.1. System Monitoring, or Logged Data	44

4.2. The Questionnaire Method of Software Evaluation	55
4.3. The Interview as a Software Evaluation Method	64
4.4. Verbal Protocol Analysis for Software Evaluation	69
4.5. Other Evaluation Techniques	77
4.6. Comparisons of Evaluation Methods	81
4.7. Composite User Based Methods	84
Chapter Five: The Evaluation Study	88
5.1. Experimental Design	88
5.2. Experimental Tasks	89
5.3. Sample Related Issues	90
5.4. Software	90
5.4.1. Spreadsheet	91
5.4.2. Word Processor	92
5.4.3. Database	92
5.5. Operational Definitions and Evaluation Method Development ..	93
5.5.1. The Logged Data Collection Procedure	93
5.5.2. The Questionnaire	93
5.5.3. Interview Procedure	95
5.5.4. Verbal Protocol Procedure	95
5.6. Analyses	97
5.6.1. Equivalence of Groups: Subjects' Confidence Ratings ..	97
5.6.2. Analysis of the Logged Data	97
5.6.3. Analysis of the Questionnaire Data	98
5.6.4. Analysing the Interview	99
5.6.5. Verbal protocol analysis	100
5.7. The Between Evaluation Methods Comparisons	100
5.7.1. Specific Hypotheses	103
5.8. Practical considerations: The Evaluator's Perspective	103
5.9. Method	104
5.9.1. Subjects	104
5.9.2. Materials	105

5.9.3. Procedure	106
------------------------	-----

Chapter Six: The Evaluations and Discussion of Each Single Evaluation

Method	109
6.1. Equivalence of the Groups	109
6.2. Specific Problems Highlighted by Each Evaluation Method ...	115
6.3. The Logged Data Method	119
6.4. The Questionnaire Method	128
6.5. The Interview as a Software Evaluation Method	139
6.6. The Subsequent Aided Verbal Protocol Analysis Method	142

Chapter Seven: A Comparison of the Evaluation Methods 146

7.1. A Qualitative Examination of the Information Elicited	146
7.1.1. Spreadsheet: The Print Sub-Section	146
7.1.2. Word Processor: The Print Sub-Section	147
7.1.3. Database: The Print Sub-Section	148
7.1.4. An Overview of the Qualitative Information	149
7.2. Usability Problems Highlighted by Evaluation Methods Used Alone	150
7.3. The Incidence of Problem Identification within each Evaluation Group	154
7.4. Problem Identification using Two Evaluation Methods	158
7.5. Practical Psychology	162
7.5.1. Practical Feasibility of the Methods	163

Chapter Eight: General Discussion 167

8.1. Refining the Evaluation Strategies	167
8.1.1. A Theory of the User Based Software Evaluation Process	167
8.1.2. A Post-Hoc Examination of the Evaluation Process Using the Data in the Present Study	169

8.1.3. Specific Hypotheses about the Efficiency of a User Based Evaluation Strategy	172
8.2. Suggestions to Improve the Evaluation Strategies Used in this Study	173
8.3. Suggestions for Future Research	178
8.4. Levels of Criteria: A Reframe of the Evaluative Outcomes	180
8.5. The Law of Diminishing Returns as Applied to the Software Usability Evaluation Process	183
8.6. The Software Evaluation Process and the Psychologist	187
References	189
Appendices	210

LIST OF TABLES

Table 1.1. Revenue Gained Worldwide From the Four Main Categories of Business Software Used on Personal Computers (from DiNucci, 1985)	2
Table 1.2. Seven Step Evaluation Process (from Suchman, 1967) ..	10
Table 1.3. Test Plan for Conducting Ergonomic Evaluations (from Meister, 1986)	13
Table 3.1. Suggested Components of Usability (from Gould, 1987)	31
Table 3.2. Taxonomy of Evaluation Methods (from Maguire and Sweeney, 1989)	38
Table 3.3. Relationship between Evaluation Metrics and Data Capture Methods (from Maguire and Sweeney, 1989)	39
Table 3.4. Some Suggested Measures of the Usability Construct ..	40
Table 4.1. General Methodology for Monitoring and Evaluation of On-Line Information System Usage (from Penniman and Dominick, 1980)	46
Table 4.2. Some of the Suggested Information that can be Recorded Using the Logged Data Approach to System Evaluation	50
Table 4.3. Advantages and Disadvantages of the Questionnaire Approach (from Meister, 1986)	55
Table 4.4. Considerations that Should be Used when Designing a Questionnaire (derived from Bouchard, 1976)	57
Table 4.5. General Characteristics of the Interview (from Sinclair, 1990)	65
Table 4.6. Four Types of Interviews Classified According to Type of Question and Type of Answer Required (from Bouchard, 1976)	66
Table 4.7. Comparative Time Required to Conduct the Evaluation (from Yamagishi and Azuma, 1987)	83

Table 5.1. Design Used in the Present Study	88
Table 5.2. Interview Questions	96
Table 5.3. General Approach to the Data Analysis for Each Evaluation Method	101
Table 5.4. Cell Sample Sizes Used in the Study	105
Table 6.1. Summary of Group Means, Standard Deviations, and Sample Sizes of the Subjects' Confidence Ratings and Overall Usability, for the Three Software Packages Evaluated	110
Table 6.2. Summary of ANOVAs Used to Test the Subjects' Confidence Ratings and Usability Scores, by Application Domain, and by Evaluation Group Assigned	111
Table 6.3. Means, Standard Deviations and Sample Sizes of the Three Sub-Samples When a Gender Based Breakdown is Undertaken	113
Table 6.4. Summary of One Way ANOVAs Used to Test the Differences Between the Female and Male Self Perceived Confidence Scores and Performance Scores for the Three Sub-Samples	114
Table 6.5. Problems Identified with the User Interface of the Spreadsheet Package and the Evaluation Method Identifying the Problem	116
Table 6.6. Problems Identified with the User Interface of the Word Processing Package and the Evaluation Method Identifying the Problem	117
Table 6.7. Problems Identified with the User Interface of the Database Package and the Evaluation Method Identifying the Problem	118
Table 6.8. Summary t-test Values for a Correlated Mean t-test and a Reduction of Variance t-test for the Three Instances of the Spreadsheet Sub-Tasks	120

Table 6.9. Summary t-test Values for a Correlated Mean t-test and a Reduction of Variance t-test for the Three Instances of the Word Processor Sub-Tasks	121
Table 6.10. Summary t-test Values for a Correlated Mean t-test and a Reduction of Variance t-test for the Three Instances of the Database Sub-Tasks	122
Table 6.11. Short Transcript and Interpretations of an Obtained Log When Using the Logged Data Collection Process	125
Table 6.12. Part of a Short Transcript of a Completed Evaluation Questionnaire	128
Table 6.13. Attributes Rated as Low Using the Questionnaire	129
Table 6.14. Cronbach's Alpha Reliability Coefficients for the Eight Factors of the Questionnaire Used in the Study	131
Table 6.15. Short Portion of an Interview After Transcription	139
Table 6.16. Match Between the Reply in the Interview "Problems" Section and the Interview "Suggestions for Improvement" Section	140
Table 6.17. Mean Ratings, and Relative Rankings of Difficulty for Each of the Separate Sub-Tasks Undertaken	141
Table 6.18. Short Transcript of a Protocol Obtained While Using the Word Processor	144
Table 7.1. Summary of the Number of Problems Identified by Each Evaluation Method Across Each Software Package Evaluated (derived from Tables 6.5 to 6.7)	151
Table 7.2. Hierarchical Log Linear Analysis of the Frequency of Problems Identified by Each Evaluation Method Within Each of the Three Application Domains Evaluated	153
Table 7.3. Comparison Between the Number of Problem Areas Identified by Each Evaluation Method Using the Chi Square Statistic	154
Table 7.4. Percentage of Subjects who Reported Each of the Six Most Frequently Occurring Problems	156

Table 7.5. Summary of Chi Square Tests of the Relationship Between the Percentage of Individuals Encountering a Problem in the Logged Data Group and the Questionnaire, Interview and Protocol Analysis Groups Respectively	156
Table 7.6. Percentage of the Total Number of Problems Identified Using Two Evaluation Methods, Compared to Each Single Method	159
Table 7.7. Hierarchical Log Linear Analysis of the Frequency of Problems Identified by Each Evaluation Method and Each Combination of Two Methods Within Each of the Three Types of Software	161
Table 7.8. Summary of Practical Aspects	164
Table 8.1. Explicit Hypotheses About the Efficiency of a User- Based Evaluation Method for Identifying Problem Areas Within a Human-Computer User Interface	174
Table 8.2. A Suggested Format for a Questionnaire for Assessing the Human-Computer Interface	176

LIST OF FIGURES

Figure 1.1. Four-Section Aggregations of US Workforce 1860-1980 (from Parker, 1976)	3
Figure 1.2. A Model of Evaluation Research (from Wortman, 1975) .	12
Figure 2.1. Iterative Software Design Process (from Williges, Williges, and Elkerton, 1987).	18
Figure 3.1. The Deficient and Contaminated Evaluation	34
Figure 3.2. Potential Relationships Between the Ultimate Evaluation and Two Actual Evaluations (adapted from Blum and Naylor, 1968).	35
Figure 4.1. A Suggested Methodology for the Monitoring and Evaluation of On-Line Information System Usage (from Penniman and Dominick, 1980)	47
Figure 4.2. The Direct Hardware Tap Method for Collecting On- Line Information System Usage (adapted from Theaker, Phillips, Frost and Love, 1989)	48
Figure 6.1. Proportions of Individuals Who Responded 'Don't Understand the Question,' for the Questions in Factors 1 and 2 Over the Three Software Packages Evaluated	132
Figure 6.2. Proportions of Individuals Who Responded 'Don't Understand the Question,' for the Questions in Factors 3 and 4 Over the Three Software Packages Evaluated	133
Figure 6.3. Proportions of Individuals Who Responded 'Don't Understand the Question,' for the Questions in Factors 5 and 6 Over the Three Software Packages Evaluated	134
Figure 6.4. Proportions of Individuals Who Responded 'Don't Understand the Question,' for the Questions in Factors 7 and 8 Over the Three Software Packages Evaluated	135

Figure 7.1. Percentage of the Total Usability Problems Identified by Each Evaluation Method Over the Three Evaluation Tasks	152
Figure 7.2. Percentage of Individuals Reporting Each of the Six Most Frequently Occurring Problems Examined.	157
Figure 7.3. Average Percentage of the Total Number of Problems Identified Using Two Evaluation Methods, Compared to Using a Single Evaluation Method	160
Figure 7.4. Comparison of Evaluation Methods on a Standardised Continuum of Merit	165
Figure 8.1. Law of Diminishing Returns Applied to the Evaluation Method Used and Issues Identified.	185

LIST OF APPENDICES

Appendix 1. Experimental Subject Contact Procedure	210
Appendix 2. Informed Consent Form Used in the Study	211
Appendix 3. General Overview Information Sheet Used in the Study	212
Appendix 4. Introduction to the Spreadsheet Information Sheet Used in the Study	213
Appendix 5. Spreadsheet Experimental Task Used in the Study . .	215
Appendix 6. Introduction to the Word Processing Information Sheet Used in the Study	217
Appendix 7. Word Processing Experimental Task Used in the Study	219
Appendix 8. Introduction to the Database Information Sheet Used in the Study	221
Appendix 9. Database Experimental Task Used in the Study	223
Appendix 10. Diagrammatic Outline of the Spreadsheet Used in the Study	226
Appendix 11. Diagrammatic Outline of the Word Processor Used in the Study	227
Appendix 12. Diagrammatic Outline of the Database Used in the Study	228
Appendix 13. Base Logged Data Collection Routine	229
Appendix 14. Software Usability Questionnaire Used in the Study .	232
Appendix 15. Usability Rating Sheet Used in the Study	240
Appendix 16. Means, Standard Deviations and Sample Sizes of the Questionnaire Ratings	241