



Real World User Testing:

An Assessment of User Testing

Methodologies in Theory and Practice

Joshue O Connor

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Assistive Technology)

January 2011

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Assistive Technology), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: _____

Date: ***26 January 2011***

1 ABSTRACT

This research sets out to evaluate the practices of the professional usability community when undertaking usability tests of web sites and Rich Internet Applications (RIAs) by examining current user-testing methodologies deployed in user testing with older people and people with disabilities.

The research will attempt to create a snapshot of the current practice that involves a literature review of the topic, some relevant case studies and gives some background as to how some of these methodologies came into being and how they are currently used in professional practice. In the Knowledge Audit presented in this work the aim will be to look at some questions such as:

- 1) In practice do most usability practitioners have an established methodology at all when undertaking user testing?
- 2) If they do not use an established methodology, how do they structure their tests and measure the outcomes?
- 3) If an ad-hoc method is used, can this be considered reliable or un-reliable?
- 4) If an ad-hoc method is used does this approach even have some advantages over more rigid or formal testing methodologies?

This Europe wide research aims to give an overview of current usability practices in diverse domains, such as EU research projects, academia and the commercial world.

Key words: Assistive Technology, Usability, HCI, Accessibility, Universal Design.

ACKNOWLEDGEMENTS

Firstly, I would like to thank my very patient wife, Lorraine McDyer O Connor, for her support and encouragement over the years. Without you, I would still be in a field trying to string my guitar.

All of the people who have helped me complete my work, my tutor Damian Gordon, Brendan Tierney, Colman McMahon, the guys in the Centre for Excellence in Universal Design (CEUD), my colleague Antoinette Fennell (for Excel expertise), Jenny McCann for the photos - and all the usability professionals who kindly took part in the research who were Barbara Schmidt-Belz, Veronika Jermolina, Henry Poskitt, Alastair Campell, Ruairi Galvin, Laurence Veale, Clodagh Kelly, Henrike Gappa, Gaby Nordbrock, and Mark Magennis – to whom I owe a special thanks for his detailed feedback when I was writing up.

Finally, this thesis is dedicated to my Dad, Michael J. O Connor - who would be chuffed to see that I finally have a job.

“Exile ends in Glory.” T. Merton

TABLE OF CONTENTS

1 ABSTRACT	II
-------------------------	-----------

TABLE OF FIGURES	XI
-------------------------------	-----------

1. INTRODUCTION	1
------------------------------	----------

1.1 BACKGROUND	1
1.2 RESEARCH PROBLEM	2
1.3 INTELLECTUAL CHALLENGE	2
1.4 RESEARCH OBJECTIVES	2
1.5 RESEARCH METHODOLOGY	3
1.6 RESOURCES	3
1.7 SCOPE AND LIMITATIONS	3
1.8 ORGANISATION OF THE DISSERTATION	4

2 LITERATURE REVIEW	5
----------------------------------	----------

2.1 INTRODUCTION	5
2.2 SOFTWARE DEVELOPMENT METHODOLOGIES	5
2.2.1 <i>Sequential Methodologies</i>	5
2.2.2 <i>The Software Waterfall Lifecycle Model</i>	5
2.2.3 <i>The Spiral Model</i>	6
2.2.4 <i>The V Model</i>	7
2.2.5 <i>Sawtooth Method</i>	7
2.3 ITERATIVE MODELS	8
2.3.1 <i>The Evolutionary Model (Evo)</i>	8
2.3.2 <i>Agile Models</i>	9
2.3.3 <i>SCRUM</i>	10
2.3.4 <i>Xtreme Programming (XP)</i>	11
2.3.5 <i>Rational Unified Process (RUP)</i>	11
2.4 INTERACTION DESIGN	12
2.4.1 <i>The process of action</i>	12
2.4.2 <i>The ‘Gulf of Evaluation’ and the ‘Gulf of Execution’</i>	14

2.4.3	<i>Functional Content, Semantics and Behaviour</i>	15
2.4.4	<i>Goal Orientated Design and building successful products</i>	16
2.4.5	<i>Goal Vs Task Oriented Design</i>	18
2.4.6	<i>Understanding Mental models</i>	19
2.4.7	<i>Designing for Intermediates</i>	19
2.4.8	<i>The Master Apprentice Model and Contextual Inquiry</i>	21
2.5	COGNITIVE ERGONOMICS	22
2.6	HCI	22
2.7	USER CENTRED DESIGN (UCD)	23
2.7.1	<i>The UCD Process</i>	25
2.8	DEFINING ACCESSIBILITY	27
2.8.1	<i>Why be Accessible?</i>	27
2.8.2	<i>Accessibility: From theory to practice</i>	28
2.8.3	<i>Dealing with Change</i>	28
2.8.4	<i>What Are the Benefits of Accessibility?</i>	28
2.8.5	<i>Assessing Accessibility</i>	30
2.8.6	<i>WCAG 2.0 (Web Content Accessibility Guidelines)</i>	30
2.8.7	<i>Who WCAG is for?</i>	30
2.8.8	<i>Constraints Based Design is not a bad thing</i>	31
2.8.9	<i>Understanding Accessibility</i>	32
2.9	ASSISTIVE TECHNOLOGY AND UNDERSTANDING DISABILITY	33
2.9.1	<i>Blindness</i>	33
2.9.2	<i>Vision Impairment</i>	33
2.9.3	<i>Glaucoma</i>	34
2.9.4	<i>Macular degeneration</i>	34
2.9.5	<i>Retinopathy</i>	35
2.9.6	<i>Detached retina</i>	36
2.10	PHYSICAL DISABILITY	36
2.11	COGNITIVE AND SENSORY DISABILITIES	37
2.12	ASSISTIVE TECHNOLOGY (AT)	37
2.12.1	<i>What is a screen reader?</i>	38
2.12.2	<i>Screen Magnification</i>	39
2.12.3	<i>Switch Access</i>	39
2.12.4	<i>How do switches work?</i>	40

2.12.5	<i>Mouse emulation</i>	41
2.13	UNIVERSAL DESIGN	42
2.13.1	<i>Principle 1: Equitable Use</i>	43
2.13.2	<i>Principle 2: Flexibility in Use</i>	44
2.13.3	<i>Principle 3: Simple and Intuitive Use</i>	45
2.13.4	<i>Principle 4: Perceptible Information</i>	46
2.13.5	<i>Principle 5: Tolerance for Error</i>	47
2.13.6	<i>Principle 6: Low Physical Effort</i>	48
2.13.7	<i>Principle 7: Size and Space for Approach and Use</i>	49
2.13.8	<i>Conclusions</i>	49
3	USER-CENTERED DESIGN (UCD) AND EVALUATION METHODS.....	51
3.1	INTRODUCTION.....	51
3.2	WHAT IS USABILITY?	51
3.3	PARTICIPATORY DESIGN	53
3.4	FOCUS GROUP RESEARCH	53
3.5	SURVEYS	53
3.6	THE COGNITIVE WALKTHROUGH	54
3.7	EXPERT EVALUATIONS.....	54
3.8	USING PERSONAS	55
3.8.1	<i>Building Personas</i>	55
3.8.2	<i>Does using Personas Work?</i>	55
3.8.3	<i>Measuring the effectiveness of using Personas</i>	56
3.9	FIELD STUDIES	57
3.10	CRITERIA FOR EVALUATING OF USABILITY OR USER EVALUATION METHODS (UEM)	57
3.10.1	<i>Can comparison of user evaluation methods be meaningful?</i>	58
3.10.2	<i>Iterative Design Process</i>	59
3.11	USABILITY METHODOLOGIES AND STANDARDS	61
3.11.1	<i>ISO and Usability</i>	61
3.11.2	<i>Usability as a quality objective</i>	61
3.12	CONCLUSIONS	64

4 CASE STUDIES	65
4.1 INTRODUCTION.....	65
4.2 REMOTE USER TESTING	65
4.3 CASE STUDY #1: COMPARATIVE EVALUATION OF USABILITY TESTS.....	66
4.3.1 <i>Test Application</i>	67
4.3.2 <i>Test Output</i>	68
4.3.3 <i>Quantitative Usability Measurements</i>	69
4.3.4 <i>Qualitative Reporting</i>	70
4.3.5 <i>Observations</i>	70
4.3.6 <i>Differences in time and outputs</i>	71
4.3.7 <i>Test Report Comparison</i>	72
4.3.8 <i>Test Results and Observations</i>	75
4.3.9 <i>Test Team Observations</i>	75
4.4 CASE STUDY #2: TESTING THE “5 USER ASSUMPTION”	76
4.4.1 <i>So when is enough, enough?</i>	76
4.4.2 <i>Are 5 users always enough?</i>	79
4.4.3 <i>More users please!</i>	79
4.4.4 <i>To test or not to test?</i>	81
4.5 CASE STUDY #3: EVALUATING THE EVALUATOR EFFECT	82
4.5.1 <i>The role of the User Test Facilitator</i>	82
4.5.2 <i>The Test Sessions</i>	83
4.5.3 <i>Assessment Criteria</i>	83
4.5.4 <i>Test Results</i>	85
4.5.5 <i>Level of Agreement</i>	86
4.5.6 <i>Results</i>	87
4.5.7 <i>When is a usability problem not a problem?</i>	88
4.5.8 <i>Retrospective reporting</i>	88
4.5.9 <i>Analyzing and Communicating Usability Data</i>	88
4.6 CASE STUDY #4: EFFECTIVELY COMMUNICATING THE RESULTS OF USABILITY DATA	89
4.6.1 <i>‘Think aloud’ Studies</i>	91
4.7 CONCLUSIONS	92

5 SECTION 2: KNOWLEDGE AUDIT BACKGROUND. EXPERIMENTATION & EVALUATION 93

5.1	EVALUATING USABILITY TESTING	93
5.1.1	<i>User Testing with People with Disabilities</i>	93
5.1.2	<i>Formal vs. Informal User testing.....</i>	94
5.1.3	<i>Measuring User Testing outputs.....</i>	95
5.1.4	<i>How does User Testing work?</i>	96
5.1.5	<i>The User Environment.....</i>	96
5.1.6	<i>The Observation environment.....</i>	96
5.1.7	<i>Test details</i>	97
5.1.8	<i>Observing a user test</i>	98
5.1.9	<i>Goals of User Testing</i>	99
5.1.10	<i>Limitations of Testing.....</i>	100
5.2	EXPERIMENTATION	101
5.2.1	<i>Basics of a Testing Methodology.....</i>	101
5.2.2	<i>Basic Elements of User Testing</i>	102
5.3	THE EXPLORATORY TEST.....	104
5.3.1	<i>When:</i>	104
5.3.2	<i>Objective:.....</i>	104
5.3.3	<i>Overview of the Methodology.....</i>	105
5.4	ASSESSMENT TEST	106
5.4.1	<i>When:</i>	106
5.4.2	<i>Objective:.....</i>	106
5.4.3	<i>Overview of the Methodology.....</i>	106
5.5	VALIDATION TEST.....	107
5.5.1	<i>When:</i>	107
5.5.2	<i>Objective:.....</i>	107
5.5.3	<i>Overview of the Methodology.....</i>	107
5.6	COMPARISON TEST.....	108
5.6.1	<i>When:</i>	108
5.6.2	<i>Objective:.....</i>	108
5.6.3	<i>Overview of the Methodology.....</i>	108

5.7	EVALUATION.....	109
5.8	CONCLUSION.....	110
6	SECTION 3: KNOWLEDGE AUDIT RESULTS	111
6.1	METHODOLOGY	111
6.2	TYPE OF WORK.....	111
6.2.1	<i>Please describe your role?</i>	<i>111</i>
6.2.2	<i>How would you describe the work that you do?.....</i>	<i>111</i>
6.2.3	EDUCATIONAL/BACKGROUND.....	112
6.3	METHODOLOGIES.....	115
6.3.1	<i>Are you aware of any existing user testing methodologies? If so please outline.</i>	<i>115</i>
6.3.2	<i>Do you use any other usability methods in your projects e.g . Case studies. Focus groups? If so please outline.</i>	<i>116</i>
6.4	SECONDARY SUPPORTIVE METHODS	117
6.4.1	<i>Are you doing user testing/usability analysis? If so please describe.</i>	<i>117</i>
6.5	STANDARDS.....	119
6.6	INFLUENCES	120
6.7	ASSISTIVE TECHNOLOGY (AT).....	121
6.8	USER TESTING PRACTICES.....	123
6.8.1	<i>If you do not adhere to a particular methodology, please outline why? Alternatively, have you created your own methodology that works within the context of your role? If so please outline how you work.</i>	<i>125</i>
6.8.2	<i>If there is any other user testing methods information you feel is relevant, please feel free to add it here, thanks.</i>	<i>126</i>
6.8.3	LAB DETAILS.....	127
6.8.4	Disability Types	130
6.8.5	Numbers in tests.....	131
6.9	OUTCOMES OF USER TESTING	132
6.9.1	<i>What do you feel the main benefits of user testing are?</i>	<i>132</i>
6.9.2	<i>Are the results of user testing incorporated into projects? If so, how?.....</i>	<i>133</i>
6.9.3	<i>Have you ever undertaken user testing more than once in the same project?</i>	<i>134</i>
6.9.4	<i>For multiple user testing sessions was it beneficial, if so how?</i>	<i>134</i>

6.9.5	<i>When testing often in the same project did it reinforce you initial findings or contradict them in any way, or did it shed fresh light?</i>	<i>135</i>
6.9.6	<i>What, in your opinion, are the main deficiencies with user testing? ...</i>	<i>137</i>
6.9.7	<i>Are there aspects of how you undertake user testing that you would like to improve?</i>	<i>137</i>
6.9.8	<i>If there is any other practical user testing information you feel is relevant, please feel free to add it here, thanks.</i>	<i>138</i>
7	CONCLUSION	139
7.1	INTRODUCTION.....	139
7.2	RESEARCH DEFINITION & RESEARCH OVERVIEW	139
7.3	CONTRIBUTIONS TO THE BODY OF KNOWLEDGE	139
7.4	EXPERIMENTATION, EVALUATION AND LIMITATION	142
7.5	FUTURE WORK.....	143
7.6	CONCLUSION	144
	BIBLIOGRAPHY	146
	APPENDIX	155

TABLE OF FIGURES

FIGURE 1: WATERFALL DEVELOPMENT DIAGRAM	6
FIGURE 2: V SOFTWARE MODEL	7
FIGURE 3: SAWTOOTH MODEL	8
FIGURE 4: EVO MODEL	9
FIGURE 5: SCRUM PROCESS	10
FIGURE 6: LIFE CYCLE OF THE XP PROCESS	11
FIGURE 7: RATIONAL UNIFIED PROCESS MODEL	12
FIGURE 8: NORMANS ACTION CYCLE	13
FIGURE 9: DIMENSIONS OF DESIGN	15
FIGURE 10: KELLEY'S TRIANGLE OUTLINES THE THREE PRIMARY QUALITIES IN A HIGH TECHNOLOGY BUSINESS	16
FIGURE 11: COOPERS EXPANSION OF THE KEELEY TRIANGLE	17
FIGURE 12: HOW DO POPULAR TECHNOLOGY VENDORS PERFORM? ⁶	18
FIGURE 13: USER DEMANDS ON SOFTWARE VARY WITH EXPERIENCE	20
FIGURE 14: USER CENTRED DESIGN	24
FIGURE 15: GLAUCOMA, RESIDUAL VISION SAMPLE	34
FIGURE 16: MACULAR DEGENERATION: RESIDUAL VISION SAMPLE	34
FIGURE 17: RETINOPATHY, RESIDUAL VISION SAMPLE	35
FIGURE 18: DETACHED RETINA, RESIDUAL VISION SAMPLE	36
FIGURE 19 SAMPLES OF POPULAR SCREEN READER AND SCREEN MAGNIFICATION APPLICATIONS	38
FIGURE 20: A VARIETY OF SWITCHES	39
FIGURE 21: THE GRID SOFTWARE	40
FIGURE 22: EZ KEYS SOFTWARE	41
FIGURE 23 UEMS USED IN FORMATIVE USABILITY EVALUATION (FROM H. REX HARTSON, TERENCE S. ET AL)	59
FIGURE 24: COMPARISON OF TEST PROCESSES	69
FIGURE 25: PERCENTAGE OF TOTAL KNOWN USABILITY PROBLEMS FOUND IN 100 ANALYSIS SAMPLES	80

FIGURE 26: THE EFFECT OF ADDING USERS ON REDUCING VARIANCE IN THE PERCENTAGE OF KNOWN USABILITY PROBLEMS. EACH POINT IS A SINGLE SET OF RANDOMLY SAMPLED USERS. THE HORIZONTAL LINES SHOW THE MEAN FOR EACH GROUP OF 100.	80
FIGURE 27: HCI PRACTITIONER EXPERIENCE, INITIAL EXPERIENCE WITH BUILDER, AND TIME ANALYSING TAPE	83
FIGURE 28: THE NUMBER OF DETECTED UPTs DEPENDS ON THE NUMBER OF USERS AND THE NUMBER OF EVALUATORS	85
FIGURE 29: PERCENTAGES OF THE UPTs DETECTED BY ONLY 1, ANY 2, ANY 3 AND ALL 4 EVALUATORS	87
FIGURE 30: A USER TEST PARTICIPANT IN THE NCBI CENTRE FOR INCLUSIVE TECHNOLOGY USABILITY LAB	97
FIGURE 31: USER TEST FACILITATOR IN THE NCBI CENTRE FOR INCLUSIVE TECHNOLOGY LAB	98
FIGURE 32: OBSERVING A USER TEST IN THE NCBI CENTRE FOR INCLUSIVE TECHNOLOGY OBSERVATION ROOM	99
FIGURE 33: THE PRODUCT DEVELOPMENT LIFE CYCLE (BASED ON ROBINS ITERATIVE MODEL, 1994).....	103
FIGURE 34 ROLE DESCRIPTION	111
FIGURE 35 GROUPS OF ROLE TYPES #362	112
FIGURE 36 LEVEL OF PROFESSIONAL EXPERIENCE	113
FIGURE 37 THE NUMBER OF USER TESTS PERFORMED	114
FIGURE 38 DESCRIPTION OF MAIN METHODS USED	117
FIGURE 39 AWARENESS OF STANDARDS.....	119
FIGURE 40 OVERVIEW OF AT TESTING/EXPERIENCE OF FACILITATOR	122
FIGURE 41 IS A METHODOLOGY USED IN YOUR WORK? OVERVIEW OF METHODOLOGY USE	123
FIGURE 42 DO YOU OWN OR RUN A USABILITY LAB?	127
FIGURE 43 USE OF VIDEO IN THE LAB	127
FIGURE 44 USE OF DATA ANALYSIS TOOLS	128
FIGURE 45 HOW ARE THE OUTPUTS FROM TESTS ARE USED?	129
FIGURE 46 DO YOU TEST WITH PEOPLE WITH DISABILITIES/OLDER PEOPLE?	129
FIGURE 47 DISABILITY TYPES TESTED WITH	130
FIGURE 48 USER TEST SAMPLE SIZES	131
FIGURE 49 PRACTITIONERS VIEW OF THE BENEFITS OF USER TESTING	132

FIGURE 50 INCORPORATION OF USER TESTING RESULTS	133
FIGURE 51 MAIN DEFICIENCIES IN USER TESTING	136

1. INTRODUCTION

1.1 Background

User Testing is a very useful way to practically assess the usability of a web interface or application (Krug, 2005). A user test is where the participant is given a set of tasks that ideally represent what the user would normally do when using the website by testing the main functionality and features the website or RIA (Rich Internet Application) has (Fraternali et al, 2010).

User testing can include diverse user groups such as people with disabilities and older people, and can be used to gain an overview of the issue that effect them when interacting with web content using Assistive Technologies (AT).

User testing came out of the more traditional field of ergonomics, and combines human factors engineering methods (Dumas, 2002), and heuristics with psychology or cognitive ergonomics (Norman. D, 1998). It is also related to advances in iterative software testing, and web development (Gould, Lewis, 1985).

In a user test there is usually a test facilitator who designs the test outline (Script) containing tasks that will be given to the participant during the user test. The test facilitator - ideally - does not intervene in the user tasks or guide the user in any way but merely sets the tasks, observes and takes detailed notes. These notes can be referred to after the test is completed. The test can also be video recorded for further observation, annotation and analysis, and it is common for other interested parties to view a user test remotely either over the Internet or in a separate observation room, if the test situation allows (CFIT Website).

This research is primarily concerned with the testing of websites and applications. Note, that if some of the external research cited refers to the testing of software it can still be considered relevant in this domain. Methodologies have existed in software development since its early inception such as the Waterfall, Spiral, Agile (Software

Methodologies, 2010) and will be looked at later on. Understanding how software development has evolved, helps us to understand how web development and design work has evolved. Looking at the various software development methodologies also illustrates parallel issues when we consider how to include the user as a central part of the Web development process.

1.2 Research problem

This research aims to take a snapshot of current usability professionals practice with regard to how they undertake user testing and feedback the results into the design and development of a website or application. So of the key questions to be addressed in this research are;

- Are the methods they use from an established methodology, do they use their own?
- What is the general level of awareness of these methods and so on?

While interested primarily in testing with people with disabilities, many of the test participants are also from backgrounds where they may test with people with disabilities but not exclusively.

1.3 Intellectual challenge

The challenge of the research is to firstly capture a snap shot of practices “in the wild” and also to get a sense where there may be gaps in what practitioners do. Also the research will help show these gaps, and are the practitioners aware of the failings and inconsistencies (if they exist) in their usability practices.

1.4 Research objectives

This research aims to take a snapshot of current usability professionals practice with regard to how they undertake user testing and feedback the results into the design and development of a website or application. So of the key questions to be addressed in this research are;

- Are the methods they use from an established methodology, do they use their own?
- What is the general level of awareness of these methods and so on?

While interested primarily in testing with people with disabilities, many of the test participants are also from backgrounds where they may test with people with disabilities but not exclusively.

The following objectives have been achieved throughout the dissertation and contributed to the overall outcome:

- Gaining a detailed glimpse of current practice.
- Having a rich source of qualitative data

1.5 Research methodology

The research was undertaken in the form of a Knowledge Audit. This is a qualitative method where respondents answer questions, mostly with descriptive prose. The audit was distributed in MS Word format to the recipients. There were 14 surveys sent out and 10 came back fully completed.

The objective of this research is to capture a snapshot, using the Knowledge audit as a form of social anthropology study. By getting a glimpse of the current state of practice in UCD, we can see what practitioners are actually doing at the moment when undertaking usability analysis and get an overview of the strengths and weaknesses of varying approaches.

1.6 Resources

Mostly used were MS Word, and Excel to record data. Surveys were sent via email to all interested parties. The main resource has been non-technical, such as the participants themselves whose experience in the field has been very valuable in undertaking this research.

1.7 Scope and limitations

This research is primarily concerned with user testing. However, there are several other aspects to a usability professional's toolkit that are also examined in order to give a more complete overview of current usability practice.

Some of these will be touched on in the Knowledge Audit questionnaire and discussed in the second part of this work. This research will take the following approaches, it will incorporate an extensive literature review, with detailed case studies to expand some of the key issues associated with user testing practices, and finally a detailed knowledge audit will be undertaken.

While the research focuses primarily on finding out about current user testing methodologies it is interesting to note that there is secondary information about usability practice, and the knowledge the participants have about people with disabilities and Assistive Technology, that the reader should find interesting. This work should not be considered exhaustive but is hopefully indicative of the current ‘state of the art’ in the world of user testing.

1.8 Organisation of the dissertation

This dissertation is broadly broken into three sections. Firstly, the Literature Review which contains chapters giving a background to the subject in greater levels of detail. This is designed to be comprehensive and informative and the reader will hopefully be able to grasp how many diverse threads are connected to the area of Universal Design and technology as an aid to social inclusion. There is then an overview of User-Centered Design and User Evaluation methods followed by some interesting Case Studies that touch on critical areas of relevance to the research.

This is followed by the Knowledge Audit background section, which covers the rationale and introduction and then finally the research itself.

2 LITERATURE REVIEW

2.1 Introduction

This introductory section will give some background on the many diverse areas that contribute to the domain of user testing and usability evaluation involving people with disabilities. It will cover areas such as HCI, WCAG, Accessibility, Usability, Universal Design and Interaction Design. These are all domains that help to form what is known of today as ‘User Evaluation Methodologies’, which will be looked at in more detail in Chapter 3.

To give this research a clearer sense of context some of the traditional software development methods that are used will be discussed to provide a comparison in terms of rigour and method. This will help to first frame what many software and web development projects are actually like in actual practice (aka “in the wild”) as opposed to in theory. There is also potentially some overlap between some current software development methods and the more progressive methods used to incorporate feedback from usability testing into web application and website testing. This should provide an overview of how traditional software development is undertaken, and how user testing can fit into software development cycles in a practical way.

2.2 Software Development Methodologies

2.2.1 Sequential Methodologies

This family of methodologies and models are called sequential because they staged (one-following the next), and until one stage is complete the next cannot begin, and once a stage is completed it is generally not revisited for better or worse.

2.2.2 The Software Waterfall Lifecycle Model

The Software Waterfall Lifecycle Model (SWLM) is a sequential development method, this means that all of the requirements for each stage are fully completed and

reviewed before the next stage begins. When each stage is completed the project moves, for better or worse, forward. (Usability First, 2010).

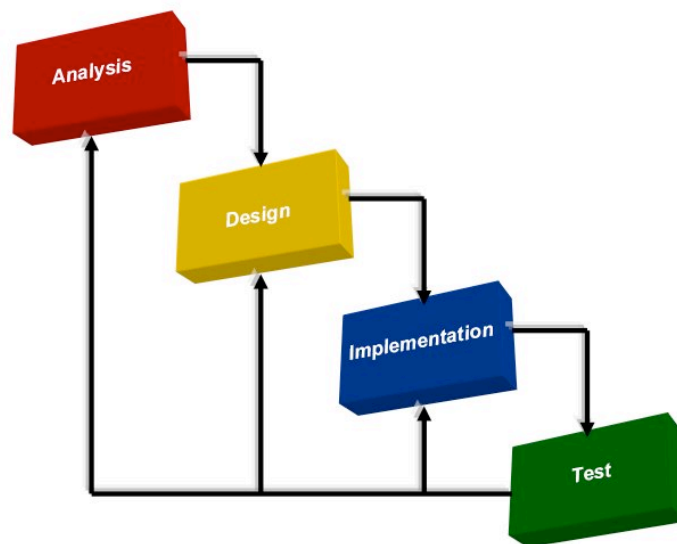


Figure 1: Waterfall development diagram

Traditionally this method has been criticised for having many shortcomings, not least, that it is a very non-responsive way to develop applications or that each of the stages are potentially independent of one another. While still widely used today, it is more common to find it used in conjunction with a more responsive or iterative methodology such as the Agile method and its many offshoots.

2.2.3 The Spiral Model

The Spiral model is one of the earliest models to extend the waterfall's activities into a cycle. Each cycle has four phases as follows:

- The first phase determines objectives, alternatives as well as any constraints.
- Risks are also identified and resolved in the next phase.
- Development and verification takes place in the third phase, and
- Planning for the next cycle takes place in the final phase.

2.2.4 The V Model

Another improvement on the waterfall model is the 'V-Model'. This is similar to the Waterfall but includes a testing phase within each of the steps. This is an improvement on earlier models, as it is a way for system quality to be tested before each stage is signed-off by the client.

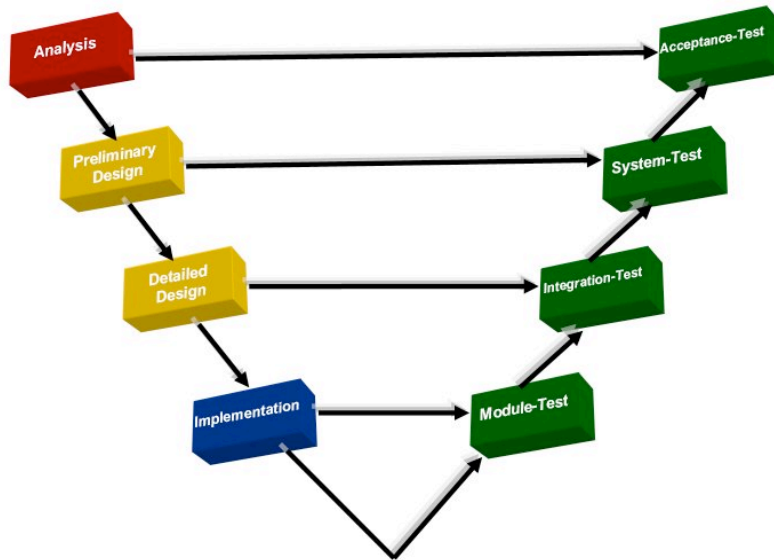


Figure 2: V Software Model

2.2.5 Sawtooth Method

The Sawtooth Model is a more responsive sequential methodology where prototypes are shown to the client for validation before sign off.. Involving the client in this way is not a guarantee of success and can be costly, but it ensures that users are more involved in the development process. It is however a more evolved sequential model and an improvement on the basic waterfall but it is still not responsive or iterative. Another development of this model is the Sharktooth Model.¹

¹ All software model images from <http://www.wittmannolan.de/ptr/cs/slccycles.html>

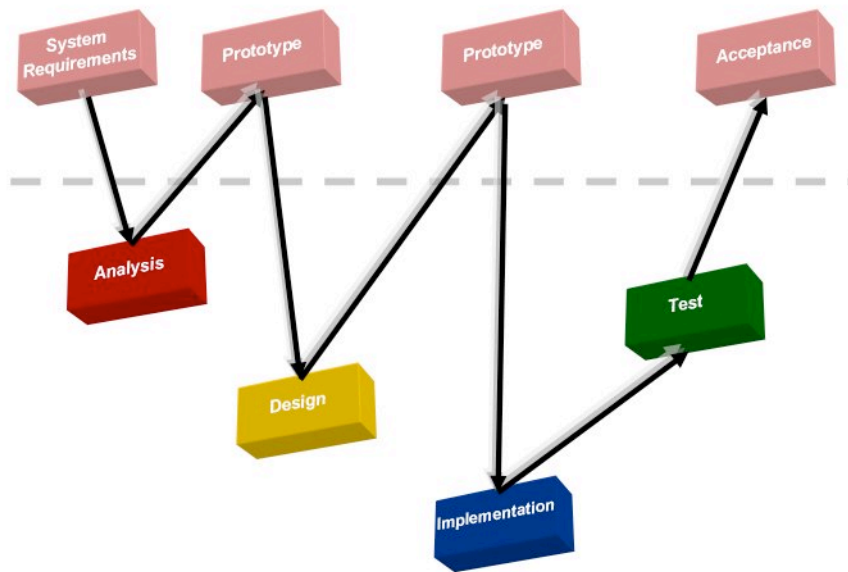


Figure 3: Sawtooth Model

2.3 Iterative Models

This family of methodologies and models are called iterative because this allows the developer to revisit previous stages as necessary, in response to any new data relevant to the project.

2.3.1 The Evolutionary Model (Evo)

The EVO model is suited to smaller projects where each of the project activities are handled in sequence and after each iteration a prototype is produced. This prototype is then used to validate the previous iteration and provides a clear idea of the requirements for the next phase.

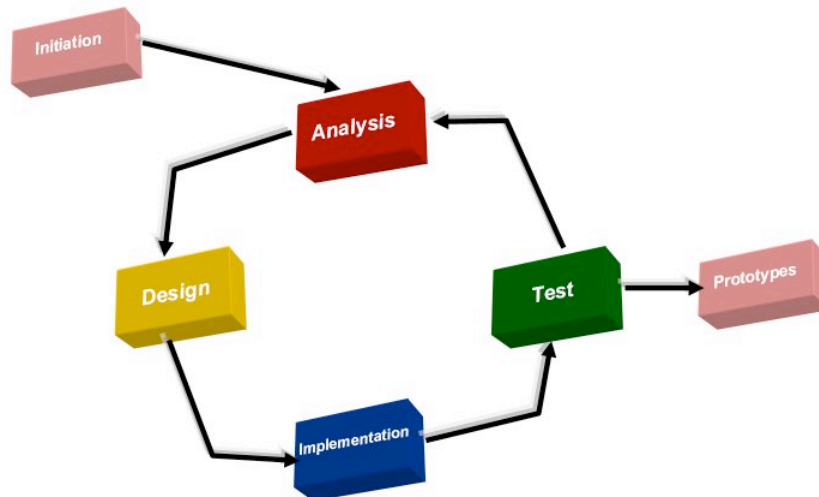


Figure 4: EVO Model

2.3.2 Agile Models

The ‘Agile Movement’ came into being when a group of software developers published the ‘Agile Software Development Manifesto’ in 2001 (Beck et al, 2001; Cockburn 2002).

The work presents a set of values that epitomise the Agile approach:

- **Individuals and interactions** over processes and tools
- **Working software** over comprehensive documentation
- **Customer collaboration** over contract negotiation
- **Responding to change** over following a plan

Thus one of the most influential modern methods of iterative programming was born. The expanded on the above set of values with 12 principles.

There are many common threads between the Agile Method and good usability practice. This may be apparent to some degree from the nature of the four values that encapsulate the spirit of Agile development. For example, if testing with people with disabilities was to take place and the feedback wasn’t favourable about some aspect of the User Interface, if Agile developers were testing they would be more open to adopting a stance towards this new knowledge of “Individuals and Interactions over processes and tools”, and “Responding to change”.

As a core manifesto, the Agile approach could be adopted by the usability community to practically improve the user experience. This idea is explored further in the paper ‘UCD in Agile Projects: Dream Team or Odd Couple?’ (McInerney. P, Maurer. F, 2005). Four of the most popular iterative methods of the Agile Methodology are Scrum, XP, RUP, and Evolution.

2.3.3 SCRUM

SCRUM is a method that consists of common sense practices that can be applied in many situations. SCRUM focuses on how team members should function in a project. Scrum helps to improve the existing development process and identify deficiencies. It has three phases: The ‘Pre-game Phase’, ‘Development Phase’ and ‘Postgame Phase’.

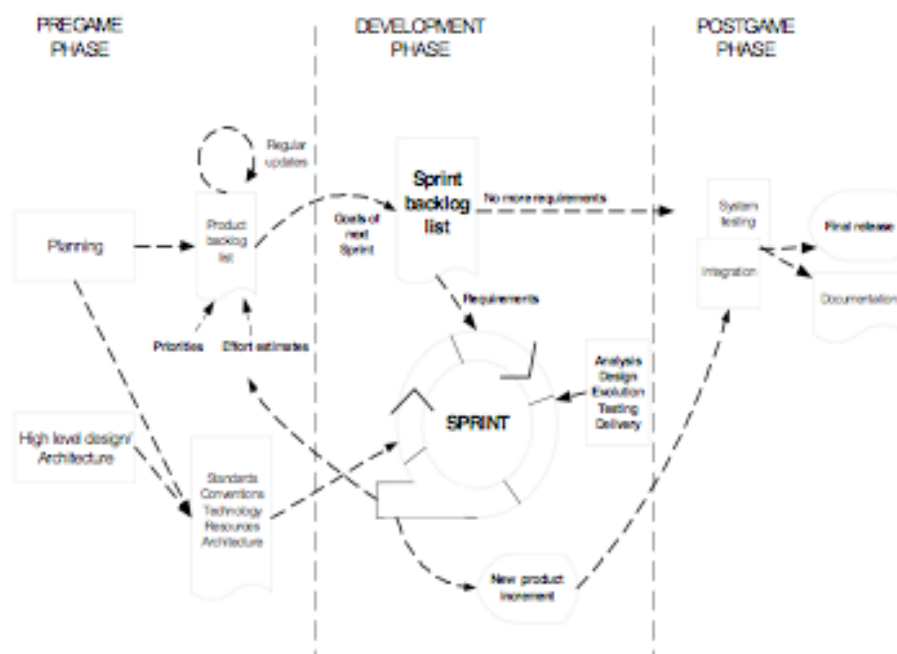


Figure 5: SCRUM Process

Scrum is by itself not sufficient as a full software development methodology and is often used in conjunction with other methods (such as XP).

2.3.4 Xtreme Programming (XP)

Xtreme Programming started as “simply an opportunity to get the job done” (Haungs, 2001). It was a response to the (at the time) long development time and delays associated with sequential development methods. The XP process has five phases: Exploration, Planning, Iterations to Release, Productionising, Maintenance and Death. It has roles and responsibilities for all involved such as the Programmer, Customer, Tester, Tracker, Coach, Consultant and Manager. At the core of this method is the idea that there is no “fixed way” of doing every project and development practices should be modified as needed to suit the need of a project. (Beck, 1996b)

XP regards ongoing changes to requirements as a natural and even desirable aspect of software development and it provides the flexibility to incorporate these changes into project iterations.

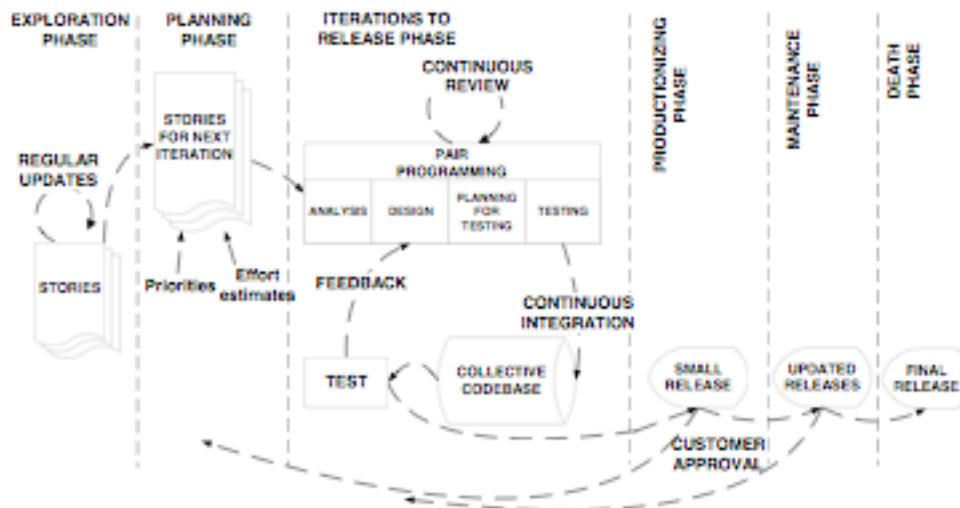


Figure 6: Life Cycle of the XP Process

2.3.5 Rational Unified Process (RUP)

The RUP model also looks at a project in terms of cycles but is much more complex. There are four phases in a cycle:

Inception, elaboration, construction and transition. In each of these phases different issues are dealt with simultaneously. A prototype is produced at the end of each cycle. Phases can be repeated as needed and often many prototypes are produced.

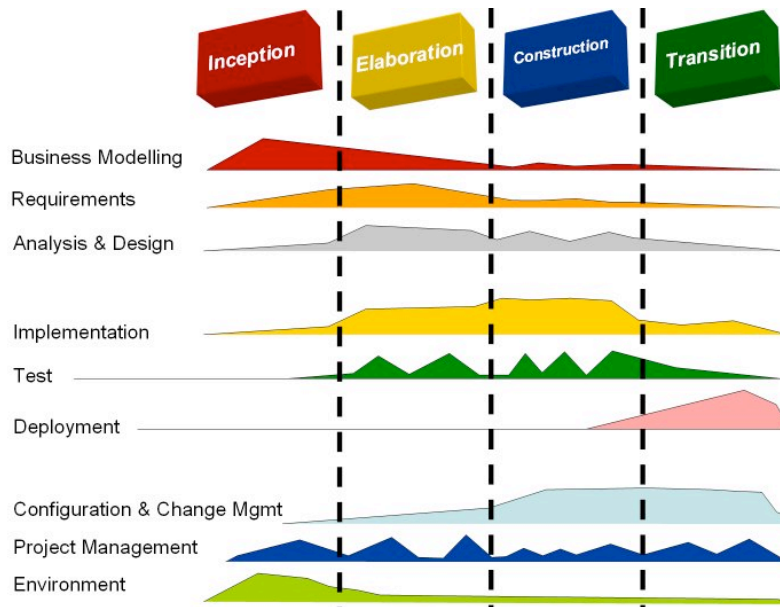


Figure 7: Rational Unified Process Model

2.4 Interaction design

Interaction design is the study of interaction between user and device but really it is all about the design of behaviours (both system and user).

2.4.1 The process of action

Before we look at the psychology of Interaction design in more detail, it is beneficial to look at the process of action. Donald Norman, one of the fathers of modern usability divided the process into four stages. (Norman, 1998)

1. The Goal of the action: This is what the use wishes to achieve.
2. Execution: This is the pushing of the button etc.
3. These actions are done in the world.
4. Evaluation of the results of the action.

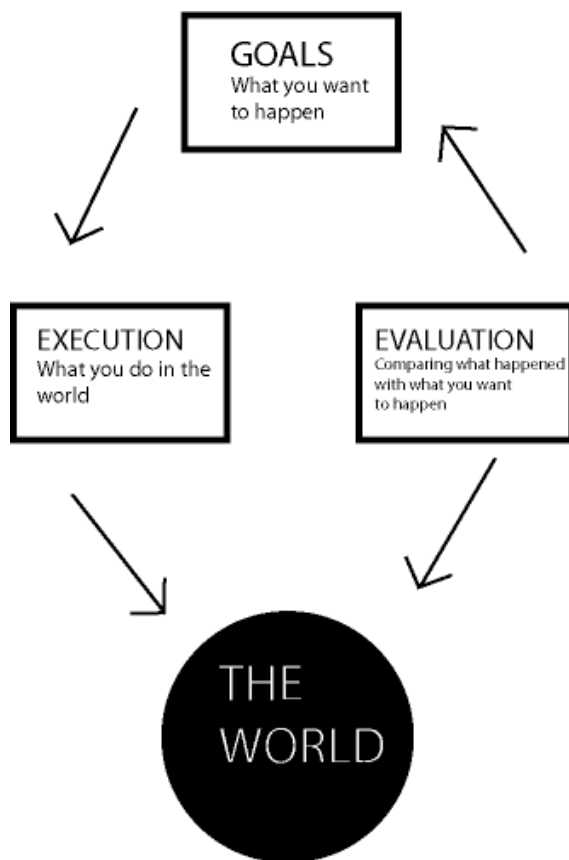


Figure 8: Normans Action Cycle²

Users often do not have very clearly defined goals (though in general when using a system or IT service it is for some specific purpose), however goals are also subject to change.

This is where good design will allow human interaction which does not lead to frustration. A good website or application needs to be designed so people who use it are not penalised when they get an action 'wrong'. A better system is one which helps them to easily identify and fix any problems that do arise.

The user will unconsciously translate their goal into a series of actions, by building in their minds a picture of what they need to do to achieve their desired goal. Finally they will carry out the actions needed. Norman refers to this sequence as the 'Stages of Execution'. The user will then evaluate if they are successful. Norman defined this process of evaluation as:

1. Perception of what happened in the world.
2. Interpret and understanding what happened.
3. Comparing what happened with what the user wanted to happen.

In order for the user to achieve what they wish, the designer needs to understand how the user will perceive the instructions and feedback the system gives them (Norman, 1998).

2.4.2 The ‘Gulf of Evaluation’ and the ‘Gulf of Execution’

Donald Norman also talked about the ‘Gulf of Evaluation’ and the ‘Gulf of Execution’ The **gulf of execution** as defined by Norman is the difference between the intentions of the users and what the system allows them to do or how well the system supports those actions (Norman 1988). This can be understood, for example, as the frustration and confusion felt when any operation or task, that to you is ‘obviously’ done in by pressing button ‘X’ but actually needs you to press ‘Y’ (and possibly activate an obscure modifier key at the same time) so you repeatedly press button ‘X’ with ever growing levels of angst. The net effect is a lot of frustration when ‘your’ button doesn’t do what you want!

This confusion can often come from a very simple process such as changing the time on a clock, or adding a contact to a mobile phone, etc. Often it is down to poor design or misattribution of a function to an object by either the designer or the user.

The gulf of execution can also be measured using the **GOMS** model (**Goals, Operators, Methods and Selection Rules**) bridging the gulf of execution means that the user must form specific intentions such as define steps or actions, undertake those actions, and select the right interface mechanisms. (Card, *et al.*, 1983).

The **gulf of evaluation** is the difficulty of assessing the state of the system and how well the artifact supports the true state of the system (Norman 1991).

2 From “The Design of Everyday Things (Donald Norman)

As Donald Norman says *"The gulf is small when the system provides information about its state in a form that is easy to get, is easy to interpret, and matches the way the person thinks of the system"* (Norman 1988: p. 51).

So if the object/system does not truly represent its state in a way that the user can understand, there is a large gulf of evaluation. In short, the gulf of evaluation and of execution refer to the mismatch between our goals and expectations. Good design reduces these by understanding what the user wants to do, and by helping them to do it in a way that is widely intuitive for the widest range of users.

2.4.3 Functional Content, Semantics and Behaviour

Since people will use computers to interact with an object in a web page or an application, it is important that we not only design web pages and application with a pleasing visual form, but that we also pay attention to what functional content means (semantics) and how it behaves.

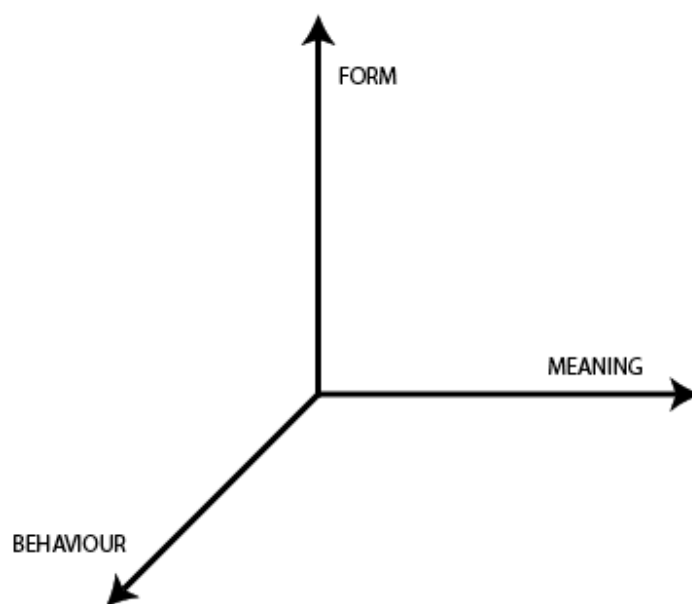


Figure 9: Dimensions of Design³

³ From About Face 2.0: The Essentials of Interaction Design (Cooper, Riemann, 2003)

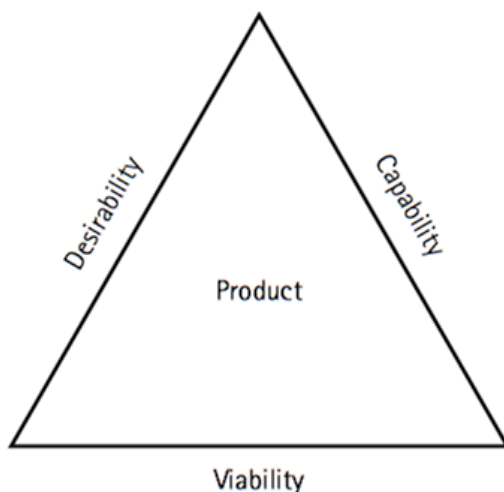
The above figure represents the ‘Dimensions of Design’. Design has traditionally been about form, and meaning but now it is vital to consider behaviour when designing systems for human interaction. (Cooper, Riemann, 2003)

2.4.4 Goal Orientated Design and building successful products

Goal Orientated design is where a users needs and concerns are balanced with engineering and project concerns (such as budget, time constraints etc). Goal Orientated Design comes from understanding what your client wants to achieve, and how a manufacturer can meet these goals.

Usually a company will develop a business model/ plan, then an engineering model and specification. Goal oriented design is similar to these processes and results in a user model and interaction plan. The ‘user plan’ determines how probable it is that a customer will use a product. The ‘business plan’ looks at the economic viability of a product, and the ‘technology plan’ looks at the technical viability of the product and whether it will work or not.

Multiplying these factors together can be used to determine the probability of greater success for a product.



*Figure 10: Kelley's Triangle outlines the three primary qualities in a high technology business*⁴

^{4/5} From About Face 2.0: The Essentials of Interaction Design (Cooper, Riemann, 2003)

Cooper then goes to expand on the original 'Kelley Triangle'.

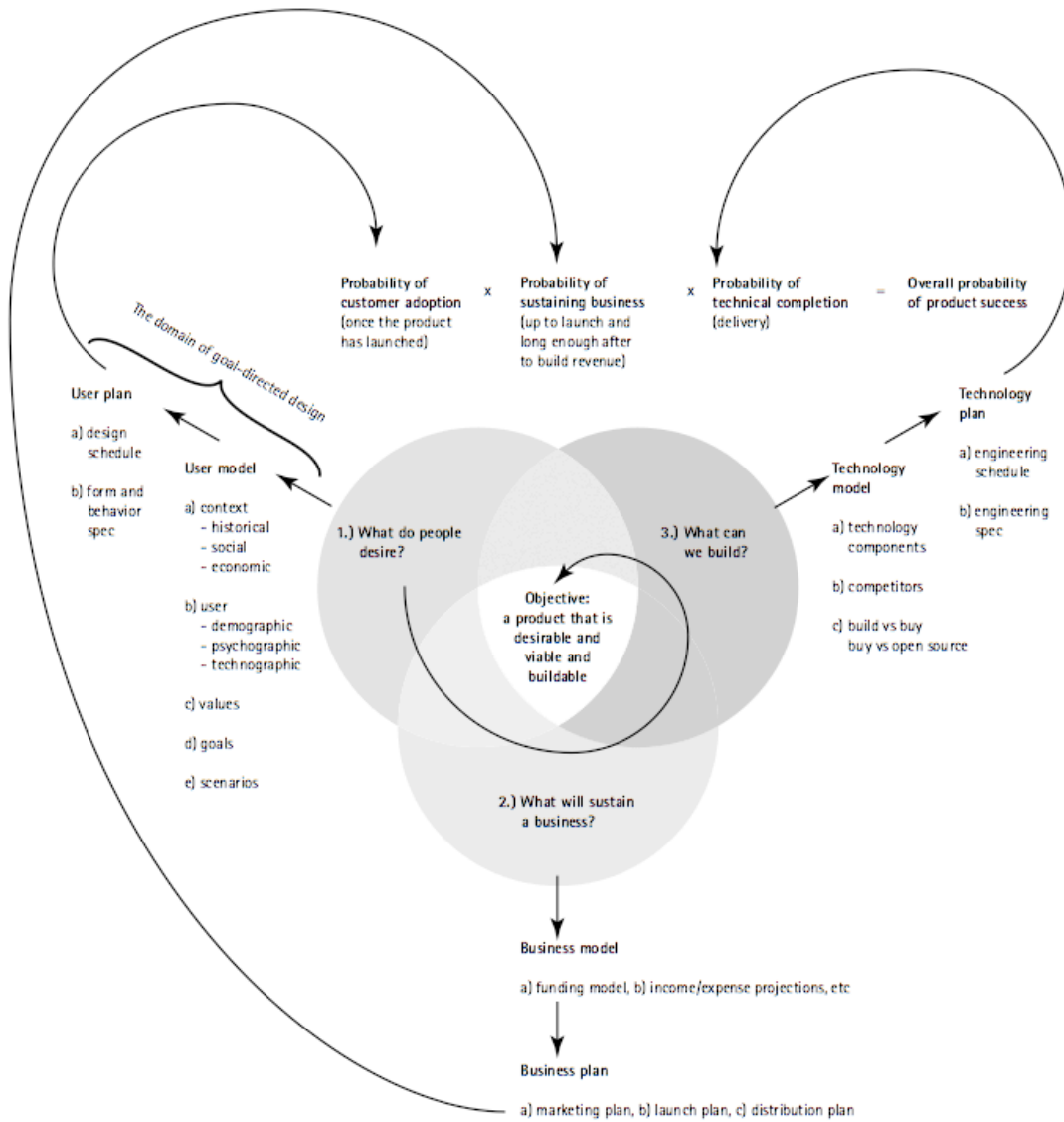
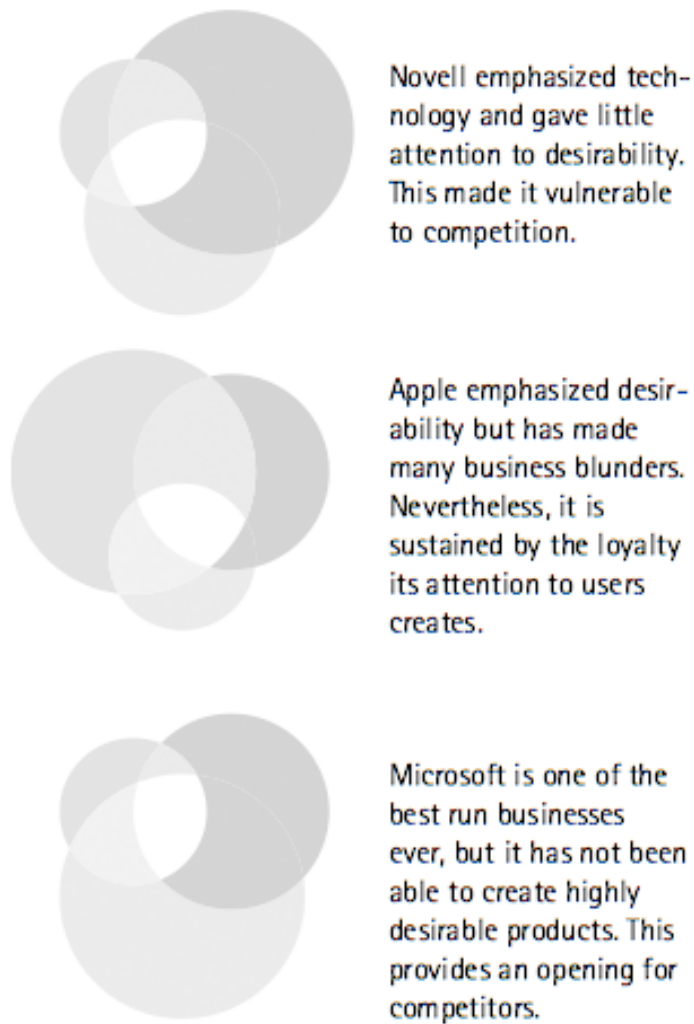


Figure 11: Coopers expansion of the Kelley Triangle⁵

Cooper suggests that the above model can be used to assess how well some of the more popular technology vendors perform.



6

Figure 12: How do popular technology vendors perform? ⁶

2.4.5 Goal Vs Task Oriented Design

‘Goals’ and ‘Tasks’ are not the same thing. It is important to outline the difference and understand how determining a suitable approach to interaction design is vital in order to achieve an appropriate result.

As we have mentioned, ‘Goal Orientated Design’ is about understanding the desires of the user and what they wish to achieve. It can be considered to be an ‘end state’ which a task is merely a transient step on the way to achieving a certain goal.

⁶ From About Face 2.0: The Essentials of Interaction Design (Cooper, Riemann, 2003)

Goals are therefore based on motivations while tasks are determined by the technology in use. It is important therefore to fully understand what the users goals are **before** any kind of task analysis etc should take place. This can then have a huge impact on the tasks that the user has to perform. So clarity in the defining user goals will really pay off when trying to understand the best architecture for the particular tasks they may need to undertake to achieve them. Therefore tasks have to be understood in this context. (Cooper, Riemann, 2003)

2.4.6 Understanding Mental models

So how does a designer use their skills to effectively instruct the user how to use a web site, operate a software application, program a digital TV recorder or any other tasks? Kenneth Craik first suggested that people will build “small-scale” mental models of the world in order to help them to understand it, reason etc. (Craik, 1943). These are “models” or ideas people have formed about themselves, the environment and in the current context - technology. These models are then used to design how the interactions between the user and the device take place.

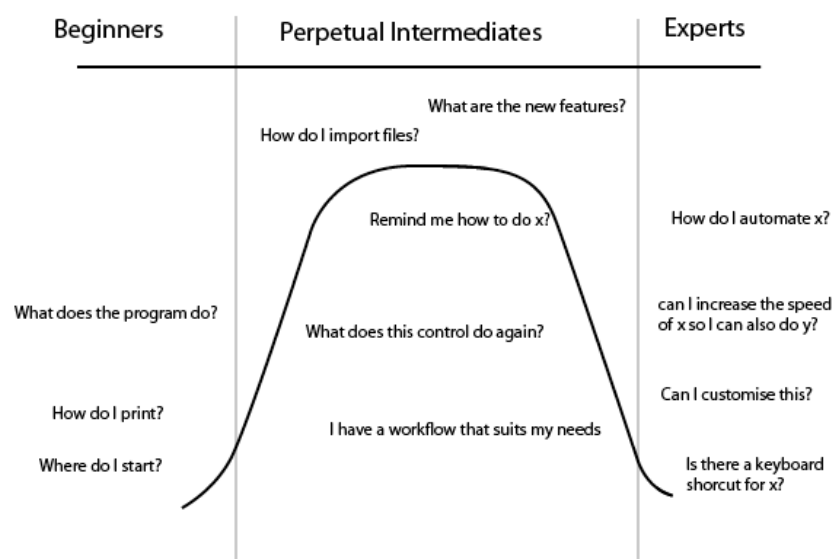
Users often unconsciously develop these models to help them understand how things work . However, perception is highly subjective and often these models can be incorrect, or at least interpreted incorrectly. This is where care and attention must be taken to ensure that the ‘mental model’ that the designer has of how their interface should be used can easily translate to what the users mental model of how the interface ‘should’ work needs in order to have a pleasurable user experience and successfully use the interface.

2.4.7 Designing for Intermediates

When designing a website, or a software application one of the biggest hurdles that developers face is how to build something that can be used by everyone from beginner to expert. Cooper maintains that most users however are not actually beginners, nor are they experts but somewhere in the middle or at the ‘intermediate’ stage. (Cooper, Riemann, 2003)

Cooper further suggests that most beginners do not stay beginners for very long, and with some experience they will quickly become ‘intermediate users’. He suggests that it is therefore best to design primarily for this ‘intermediate stage’ of users. He calls this method ‘Optimising for Intermediates’ and he breaks down this goal into three stages:

- 1) Help beginners to become intermediates as quickly as possible.
- 2) Avoid ‘getting in the way’ of intermediates who wish to become experts
- 3) Keep the ‘perpetual intermediate’ happy as they stay in a middle skill spectrum.



*Figure 13: User demands on software vary with experience*⁷

The tools presented to the user need to reflect their skill level. The interface must satisfy all users needs if it is to be successful and be used over a long period of time and not abandoned in favour of a more suitable application.

⁷ From About Face 2.0: The Essentials of Interaction Design (Cooper, Riemann, 2003)

While it is interesting to note Coopers model of ‘Designing for Intermediates’ it would be interesting research (and beyond the scope of this work) to contrast this with the process of designing for extremes and how this may or may not support intermediate users.

2.4.8 The Master Apprentice Model and Contextual Inquiry

Contextual Inquiry by Beyer and Holtzblatt (quoted in Cooper and Reimann, 2003) is based on the master-apprentice model of learning; observing and asking questions of the user as if they are the master craftsmen and the interviewer is the apprentice.

The four principles for ethnographic interviews are:

- 1) **Context:** Observe users in their own environment or a suitable context to highlight natural behaviours.
- 2) **Partnership:** The interview and observation should take the tone of a collaborative exploration **with** the user, alternating between the observation of work and discussion of its structure and details.
- 3) **Interpretation:** Much of the work of the designer is reading between the line of the facts gathered about the users behaviour, their environment and what they say.
- 4) **Focus:** The designer needs to subtly guide the interview

This process of ‘master-apprentice model’ is very a useful basis for a practical ‘real world’ user testing methodology. In particular for testing with people with disabilities, as the facilitator really does have to be empathic to the situation and circumstance of the position of someone with a disability using the web. They have to look at the world through the eyes of another, and not to dictate how they think things should be, but try to support users needs via good design.

Cooper then suggested improvements to Contextual Inquiry such as:

- 1) **Shortening the interview:** Instead of whole day interviews Beyer and Holtzblatt suggest, Cooper found that interviews of an hour or so were perfectly sufficient to collect relevant user data.

- 2) **Using smaller design teams:** Contextual inquiry also assumed large design teams conducting multiple interviews in parallel, followed by debriefing sessions. Cooper suggests that it may be more beneficial to conduct sequential interviews with the same designers (two or three in each). This means that the entire design team can interact with the user directly and is conducive to more effective data analysis.
- 3) **Identify goals first:** Contextual inquiry is very task focused. Cooper suggested that ethnographic interviews first identify and prioritize user goals first before determining tasks. Coopers' suggestions are very positive, as by defining the goals first the tasks can therefore change to suit the goal.
- 4) **Looking beyond business contexts:** Going beyond the corporate product environment into the consumer domain.

2.5 *Cognitive Ergonomics*

Cognitive ergonomics is a field that aims to understand and enhance the processes that underlying interaction between people, their environment and the systems that they use. It came from an overlap between Human Computer Interaction (HCI) and traditional workplace ergonomics. (Long, 1987)

Cognitive Ergonomics is a mix of psychology, ergonomics and human factors. It is very much a subset of HCI. In fact many of the disciplines that we are examining here all have their roots in HCI.

2.6 *HCI*

“Human-computer interaction is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them.”

HCI is the grandfather of usability and accessibility. It has its roots in many diverse fields such as computer graphics, operating systems, human factors, ergonomics,

industrial engineering, cognitive psychology, and the systems part of computer science. (HCI Origins, 1996)

From a computer science perspective, the focus is on **interaction** and **specifically on interaction between one or more humans and one or more computational machines**.

2.7 User Centred Design (UCD)

Our current models or definitions of usability have historically come from what engineers used to refer to as “human factors engineering” (Mark S. Sanders, Ernest J. McCormick, 2002) and ergonomics (D Meister, 1999).

It could be argued that this growing need for effectively considering the needs of the user in the design process was more to reduce ‘human error’ or accidents in the workplace - than out of a genuine need to improve the user experience.

The International Standards Organisation has also defined “Human centred design processes for interactive systems” as “ *[...] an approach to interactive system development that focuses specifically on making systems usable. It is a multi-disciplinary activity.*” (ISO 13407, 1999) In UCD, all “*development proceeds with the user as the centre of focus.*” (Rubin, 1994)

Rubin graphically illustrates the User-Centred Design Process as follows:

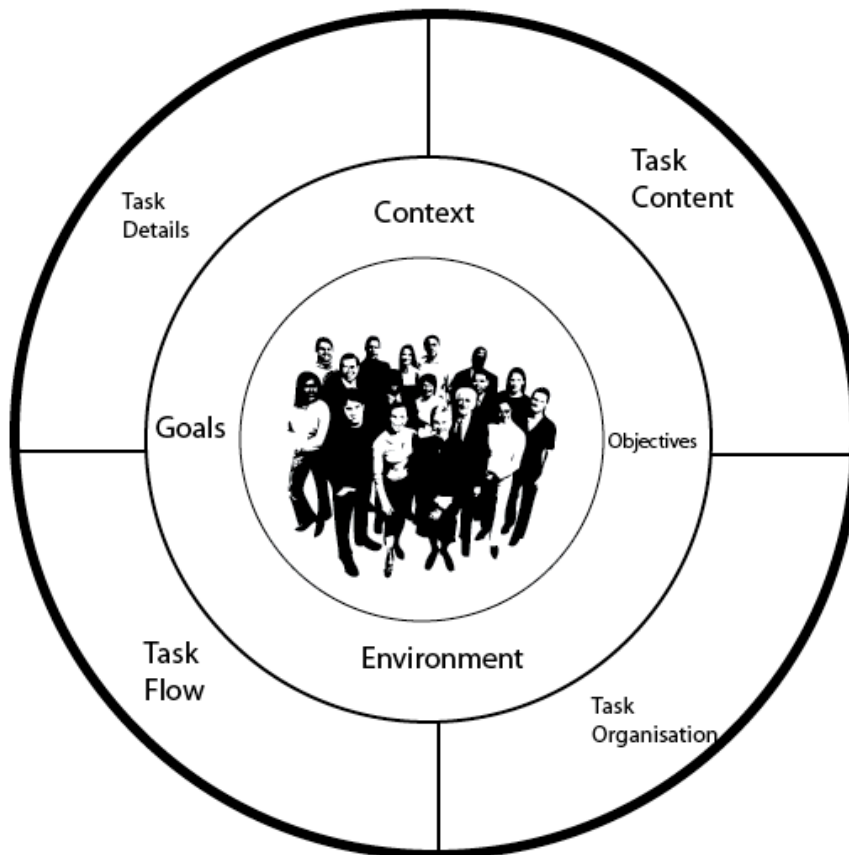


Figure 14: User Centred Design⁸

- * The users are in the centre of a double circle.
- * The inner ring contains: Context; Objectives; Environment and Goals.
- * The outer ring contains: Task Detail; Task Content; Task Organization and Task Flow.

In ‘Designing for Usability: Key Principles and What Designers Think’, the authors Gould and Lewis list three principles of a User Centered Design. (Gould, Lewis, 1985)

- 1) **An early focus on users and tasks:** Gould and Lewis advocated direct contact between the design team and users throughout the development cycle. This goes beyond the idea of merely identifying users or even just developing user personas. Rubin suggests that while this is very sound principle, care must be

⁸ From About Face 2.0: The Essentials of Interaction Design (Cooper, Riemann, 2003)

taken that the interaction is structured and there is a systematic method of collecting user data.

- 2) **Empirical measurement of product usage:** There should be behavioural measurement of ‘ease of learning’ and ‘ease of use’ very early in the design process, throughout the development and the testing of prototypes with real users.
- 3) **Iterative design whereby a product is designed, modified, and tested repeatedly:** Gould and Lewis advocate true iterative design allowing for the complete overhaul and rethinking of a design where required due to relevant user data through the earlier testing of conceptual models. It is important that designers are prepared to take this kind of step or else the influence of iterative design is merely cosmetic. True iterative design allows the design to be truly ‘shaped’ throughout the process.

While the principles that Gould and Lewis suggest are quite old (circa 1985), they are still very relevant in today’s world of RIAs and interactive web content. They form the basis of an excellent design methodology, and are deceptively simple but far-reaching in their implication if adhered to in the design and development of a web application.

2.7.1 The UCD Process

In her excellent book giving an overview of the UCD process ‘Just Ask: Integrating Accessibility Throughout Design’ (Shawn Lawton Henry, 2007) the UCD process is defined as “[...] *a user interface design process that focuses on usability goals, user characteristics, environment, tasks, and workflow in the design of an interface. UCD follows a series of well-defined methods and techniques for analysis, design, and evaluation of mainstream hardware, software, and web interfaces. The UCD process is an iterative process, where design and evaluation steps are built in from the first stage of projects, through implementation.*”

Henry then gives an example of some of the UCD process phases and steps. She suggests that User-Centred Design can be broken into three main phases: Analysis, Design, and Evaluation.

The Analysis Phase typically includes steps such as:

1. Vision, goals, objectives.
2. User analysis.
3. Task analysis.
4. Information architecture analysis.
5. Workflow analysis.

The Design Phase typically includes:

1. Conceptual/mental Model, Metaphors, Design Concepts.
2. Navigation design
3. Storyboards, wire frames.
4. Detailed design.
5. Paper or other low-fidelity prototypes.
6. Medium-fidelity prototypes, for example, online mock-ups.
7. Functional, high-fidelity prototypes.

Evaluation uses techniques such as:

1. Design walkthroughs, cognitive walkthroughs.
2. Heuristic evaluations.
3. Guidelines reviews.
4. Usability testing: low fidelity through high fidelity; informal through formal.

Henry then states that *“UCD is a process for designing usable products, and user interface accessibility can be approached as a subset of usability. It follows then that designers can use UCD to design products that are accessible. In practice, accessible design techniques do fit well into established UCD processes.”* And she outlines how accessibility fits into UCD:

- Business and usability goals include meeting accessibility requirements.
- Understanding user characteristics includes users with various disabilities.
- Environmental aspects for a mobile device include hands-free operation.
- Workflow scenarios include use of an assistive technology.
- Usability testing includes participants with disabilities.

2.8 Defining Accessibility

There are several definitions of accessibility. The International Standards Organization (ISO) defines accessibility as:

"The usability of a product, service, environment or facility by people with the widest range of capabilities" (ISO TC 16071, 2003)

If we apply this definition to the Web it refers to the interfaces that can be used by the widest possible audience; ensuring that there are no users who are left out when trying to use them. That's great; however note that this definition does not mention blind users or other people with disabilities at all, yet it talks about general **usability** (which we will look at in Section 3).

To clarify further, accessibility can in fact be grouped as a subset of **usability**. This does not mean that it is inferior in any way, but that it can be considered to be originally a child of another discipline, although at this stage the child has grown!

2.8.1 Why be Accessible?

The W3C in its 'Introduction to Web Accessibility', defines accessibility as:

"Web accessibility means that people with disabilities can use the Web. More specifically, Web accessibility means that people with disabilities can perceive, understand, navigate, and interact with the Web, and that they can contribute to the Web. Web accessibility also benefits others, including older people with changing abilities due to aging." (WAI Introduction to Accessibility, 2010)

Some definitions of accessibility specifically talk about people with disabilities and others don't. The details of exactly *how* to support the needs of diverse user groups like people with disabilities are very important and I therefore prefer the definition that specifically mentions people with disabilities.

While the first definition talks about **universality**, the fact is that people with disabilities have very specific needs that often have specific solutions that may not easily fall under the umbrella of universality.

2.8.2 Accessibility: From theory to practice

Leaving the strengths and weaknesses of theoretical definitions aside, in practice it is important to truly understand the diverse modes of interaction of your audience. So how can the average designer do this? How can the often seemingly arcane or esoteric recommendations of the WCAG be grounded in real world practice? Are there practical workable methodologies that designers and developers can apply to their projects?

User testing is a fantastic way to do this and a remarkable tool to bridge the gap between the designers and their end users and as shall be re-iterated often in this work, acts as a powerful way of moving from a theoretical to practical understanding of how design decisions impact on the user. (WAI Introduction to WCAG, 2010)

2.8.3 Dealing with Change

In many ways the discipline of accessibility encompasses our ability, as designers, to deal with change and to cope in a positive way with diversity. There are natural changes that many of us will go through such as failing sight and other physical and mental changes, as we get older. Therefore our ability to perform certain tasks and the equipment we need may also change.

Understanding accessibility involves expanding our ability to deal with these changes. The success of designers' efforts often depends on how they can accommodate diverse user requirements.

2.8.4 What Are the Benefits of Accessibility?

There are some substantial benefits of accessible web design and development:

It makes good business sense: Building accessible websites can actually increase the amount of revenue a business can turn over by ensuring that no one is excluded from their website. There are some well-documented case studies that outline the business benefits of accessibility such as:

Legal & General Group - doubled visitor numbers, cut maintenance costs by two-thirds, increased natural search traffic by 50%.

Tesco - £35 thousand GBP to build website, £13 million GBP per year in resultant revenue.

CNET - 30% increase in traffic from Google after CNET started providing transcripts. *"We saw a significant increase in SEO referrals when we launched an HTML version of our site, the major component of which was our transcripts."* - Justin Eckhouse, CNET, 2009. (W3C Business Case Examples, 2009)

These benefits can be generally categorised into:

1) Social Factors: Increased Web accessibility provides equal opportunities for people with disabilities by removing barriers to communication and interaction.

2) Technical Factors: Increased Web accessibility improves interoperability, quality, reducing site development and maintenance time, reducing server load, enabling content on different configurations, and being prepared for advanced web technologies.

3) Financial Factors: Greater accessibility improves search engine optimization (SEO); enhances direct cost savings due to easier maintenance etc.

4) Legal and Policy Factors: Increased Web accessibility addresses requirements for Web accessibility from governments and other organizations in the form of laws, policies, regulations, standards.

5) Better design: Graphic designers often design for themselves. This is not always the case, but is often true. As a result the Web is littered with sites that use tiny text that can't be resized, illegible fonts, and bad colour contrast. This often renders the site content unreadable to many. (WAI Business Case for Accessibility, 2010)

So by considering the diverse needs of users, for example, people with vision impairments who need good colour contrast and resizable text, the designers should,

for example, change their designs style to accommodate these user's needs. A good design principle is that 'form should follow function'. This is a simple but effective rule of thumb that is unfortunately often at worst ignored or just forgotten as the design process progresses.

2.8.5 Assessing Accessibility

So far we have looked at various evaluation methods and tools that are at the disposal of the accessibility or usability professional.

2.8.6 WCAG 2.0 (Web Content Accessibility Guidelines)

The Web Content Accessibility Guidelines (WCAG) documents explain how to make Web content accessible to people with disabilities. Web content generally refers to the information in a Web page or Web application, including text, images, forms, sounds, and such.

WCAG is part of a series of accessibility guidelines, including the Authoring Tool Accessibility Guidelines (ATAG) and the User Agent Accessibility Guidelines (UAAG). Essential Components of Web Accessibility explains the relationship between the different guidelines.

2.8.7 Who WCAG is for?

WCAG is primarily intended for:

- * Web content developers (Page authors, site designers, etc.)
- * Web authoring tool developers
- * Web accessibility evaluation tool developers
- * Others who want or need a technical standard for Web accessibility

WCAG and related resources are also intended to meet the needs of many different audiences, including people who are new to Web accessibility, policy makers, managers, and others. WCAG 1.0 had various priority checkpoints; WCAG 2.0 has 'success criteria'. While WCAG 1.0 was organized around a set of guidelines, WCAG 2.0 is organized around four principles.

These are four simple principles and are grouped under the acronym **POUR**.

Principle 1) Content must be Perceivable (P):

This refers to all content including Multimedia, Video and Audio.

- 1.1 Provide a text alternative for all non-text content.
- 1.2 Synchronized alternatives for Multimedia (Captioned Video, Audio Descriptions etc)
- 1.3 Information and Structure must be separate from presentation.
- 1.4 Make it easy to distinguish foreground information from background. (Good Colour contrast)

Principle 2) Interface elements must be Operable (O):

- 2.1 All functionality must be operable via the keyboard.
- 2.2 Users must control limits on their reading or interaction.
- 2.3 Users must be able to avoid content that can cause seizures due to photosensitivity.
- 2.4 Provide mechanisms for users to find content, orientate themselves and navigate through it.
- 2.5 Help users avoid mistakes and make it easier to correct mistakes when they do occur.

Principle 3) Content and controls must be Understandable (U):

- 3.1 Make text content readable and understandable.
- 3.2 Make the placement and functionality of content predictable.

Principle 4) Content should be robust enough to work with current and future technologies (R):

- 4.1 Support compatibility with current and future user agents.
- 4.2 Ensure that content is accessible or provide accessible alternatives. (WAI Introduction to WCAG, 2010)

2.8.8 Constraints Based Design is not a bad thing

Accessibility brings some important fundamental design issues back into sharp focus for designers if they are to factor in the diverse needs of people with disabilities.

This means that they must carefully consider how they layout content and how the underlying semantic structure of a website, or the functionality of a complex dynamic control could be used by someone with a disability.

The effect of considering these ‘constraints’ means that the designer must only use what is effective for the task at hand. This can actually be a marvellous approach resulting in a certain considered economy in the design process.

So "Accessibility is not anti-design". Actually it helps to ground design in best practice and allows the developer to create more future proof, interoperable applications and web sites.

2.8.9 Understanding Accessibility

Good accessible websites are actually by-product of good design. Good design comes from understanding firstly:

- 1) What you are designing for?
- 2) The purpose of your site.
- 3) Understanding your audience needs.
- 4) What they will really wish to do when using your site?

As stated previously, accessibility can seem slightly abstract and esoteric at first. It can be difficult for many developers grasp it fully as a practical discipline. This is understandable. There are some aspects of accessibility that are initially easier to understand than others but gradually as knowledge of best practice deepens developers can quickly grasp that accessibility is rather practical.

Accessibility is an ever-evolving line, a continuum. However, for users of AT it would be reasonable to state that there are some core issues for each user group that don't really change - even if the technology does. For example, blind users need to be able to access equivalent content that describes to them what a particular image is all about, people with limited physical mobility really appreciate not having a lot of useless links to tab through and so on.

Actually, Accessibility is in many ways a ‘quality’ issue and good accessible interfaces, applications and websites are therefore a happy by-product of good design and development practices.

2.9 Assistive Technology and Understanding Disability

2.9.1 Blindness

There are many different degrees of blindness. For a person to be considered blind it does not mean that they cannot see anything at all. A blind user may be able to make out some degrees of light and dark, shapes and other forms. Other blind people however may not see anything at all.

2.9.2 Vision Impairment

There are also a very broad range of vision impairments. What follow are some photographic samples that aim to simulate some of the more common vision impairments such as glaucoma and macular degeneration.

2.9.3 Glaucoma



Figure 15: Glaucoma, residual vision sample

A person with glaucoma may experience loss of their peripheral or side-vision. In the early stages glaucoma causes a subtle loss of contrast, which can lead to difficulties seeing things around the environment or using a computer.

2.9.4 Macular degeneration



Figure 16: Macular degeneration: Residual vision sample

This condition is quite common amongst older people and causes a loss of vision in the centre of the eye. Reading, writing and up-close work can become very difficult. There can also be a problem recognising colours.

2.9.5 Retinopathy



Figure 17: Retinopathy, residual vision sample

This condition causes a partial blurring of vision or patchy loss of vision and can be brought on by advanced diabetes. The persons near vision may be reduce they may have difficulty with up close reading.

2.9.6 Detached retina

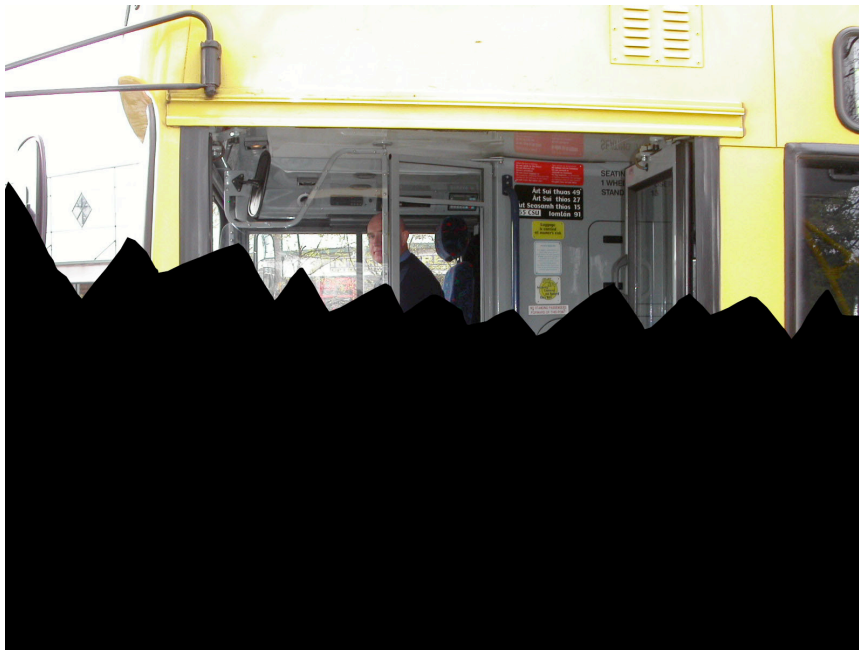


Figure 18: Detached retina, residual vision sample

A detached retina can result in a loss of vision where the retina has been damaged. A detached retina may appear like a dark shadow over part of the eye or the person may experience bright flashes of light or showers of dark spots.

2.10 Physical Disability

There are many kinds of physical disability. Some can be quite extreme and others not so. Physical disability can manifest in such a broad range of ways for many reasons. People can be born with physical disabilities or acquire them later on in life due to accident or old age.

Common mobility problems include tremors, shakes, becoming easily exhausted or experiencing difficulty in movement. Many people with physical disabilities cannot use a mouse at all and therefore have great difficulty if websites are not keyboard accessible. In fact, ensuring your websites are keyboard accessible is probably one of the single greatest things you can do to help users with physical disabilities.

2.11 Cognitive and Sensory Disabilities

Of all disability types users with cognitive and sensory disabilities are probably the hardest to accommodate. It is such a new field, particularly in its relation to the web, that methods of accommodating this user groups' needs are still being developed. In short, it is hard to find definitive evidence of what does and doesn't work for this group of web users. (O Connor, 2007)

2.12 Assistive Technology (AT)

While it is not vital that a usability professional who user tests with people with disabilities has a very in depth knowledge of how Assistive Technology works, it is desirable. Particularly when the it is important to understand how the technology works in order to be able to make technical recommendations as to how to designing and code Web interfaces for people with disabilities. While this is not always the case, as plenty of improvements can be made following outputs of user testing involving people with disabilities, but it can certainly help as some AT (such as screen readers) can be very complex and difficult to understand. The following aims to give an overview of the kind of AT that could be used in a user test by someone with a disability.

There are many kinds of assistive technology (AT) and there are also many definitions. I like this one from the US National Multiple Sclerosis Society:

“A term used to describe all of the tools, products, and devices, from the simplest to the most complex, that can make a particular function easier or possible to perform.”
(AT definition, 2010)

Note that it doesn't mention disability at all, and this is important. Most people don't think of their spectacles or our TV remote controls as assistive technology, but they actually are. The idea of technology that can be used by many different people regardless of ability, is appealing and also technology that is not just used by people with disabilities but by the ordinary users who doesn't think of themselves as being disabled. This is good design 'enabling'. These definitions take us into the realm of **Universal Design**, which we will look at later.

For a fun introduction to AT watch the AT boogie video by Jeff Moyer with animation by Haik Hoisington. (AT Animation, 2010) The following is a brief introduction to what various AT is all about and maybe help to shed some light on how people with disabilities use AT.

2.12.1 What is a screen reader?

A screen reader is text to speech software that literally reads out the contents of the screen to a user, whether it's a webpage or a structured MS Word document or even a tagged PDF. Screen readers can also interact very well with the operating system of the computer itself and can give a blind user a very deep level of interaction allowing the performance of complex system administration tasks. Screen readers are mainly used by blind and visually impaired people but screen readers can also be used by other groups, such as people with dyslexia. (Word, 2010), (PDF, 2010)

There are many different screen readers available like JAWS, Window-Eyes, the free open source Linux screen reader ORCA and the free NVDA as well as the constantly improving 'VoiceOver' which comes already bundled with Mac OS X. Some screen readers like JAWS can be very expensive. (JAWS, WinEyes, ORCA, NVDA, VoiceOver, 2010)

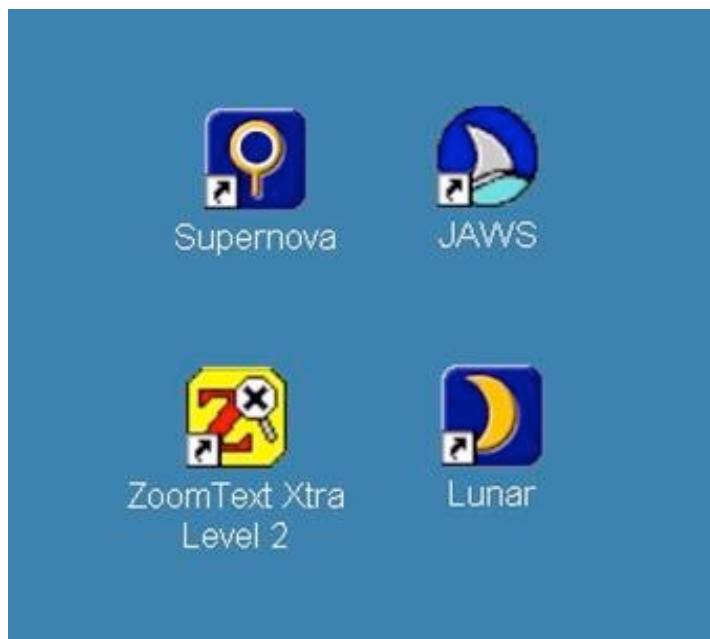


Figure 19 Samples of Popular Screen Reader and Screen magnification applications

2.12.2 Screen Magnification

Screen Magnification software allows the user to literally view their desktop or web browser at an increased rate of magnification. This feature is already a part of the Windows operating system and Mac OS X. The difference between a dedicated package and the feature in your operating system is one of quality and clarity and this is obviously really important for users with poor vision.

When you use the magnification features of your operating system you can get artifacts and blurred text whereas a screen magnification package like Supernova or Zoom text will redraw the screen at a high resolution and they have other features that provide high quality anti-aliasing so the re-drawn text is sharper and clearer. (ZoomText, SuperNova, 2010)

2.12.3 Switch Access

Enhanced informational design is also good for users with very limited physical mobility or movement. Users with physical disabilities often use a device called a **switch** to interact with their computer and access the web.



Figure 20: A variety of Switches

A switch is often a single large button designed so that the user can easily press it with the least amount of effort on their part. There are also switches that are controlled, not by pressing them but by blowing into them, or by wobbling them and a host of other forms of tactile interaction designed to suit the ability of the user.

Some users will use a combination of two or more of these switches each of which can be set to perform a different task or represent a certain input. This can greatly increase the user's power and speed of interaction with the computer or web interface. However, some users with very limited movement may successfully use only one button to interact, browse the web, type emails and other documents or play games.

2.12.4 How do switches work?

Switches are usually used in conjunction with **scanning software** applications such as the Grid, Clicker and EZKeys that are used by people who may have had a stroke or who have other physical disabilities that can result in limited or uncontrolled movement. (Grid, Clicker, EZKeys, 2010)

These scanning packages work by dividing the screen up into a grid type layout and highlighting the content of the grid one square at a time. This temporary highlighting happens in a linear fashion and is referred to as scanning. When the user wishes to select the content of the square they then press the switch button.

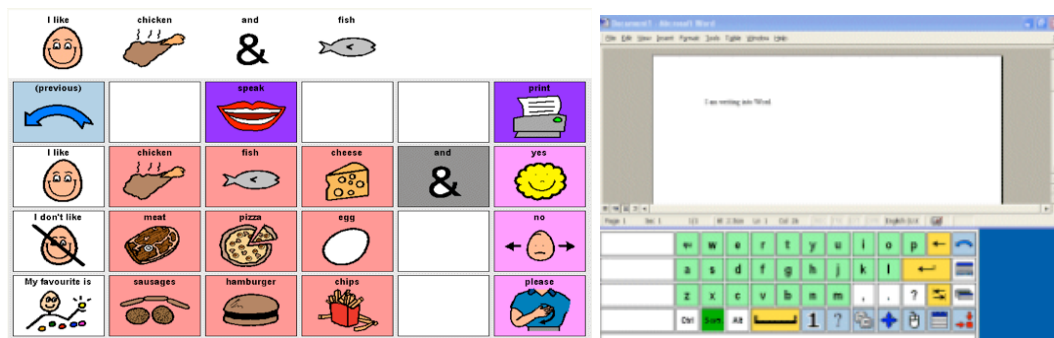


Figure 21: The Grid Software

This combination of single or multiple switch and grid type software is very empowering technology for many people with disabilities; enabling them to use their computers, communicate with family and friends via email and surf the web.

2.12.5 Mouse emulation

Another scanning type application that operates slightly differently is EZ Keys XP. EZ Keys XP provides complete mouse emulation using alternative inputs, such as a keyboard or even a slight movement of the eye using switch activation. It has several access modes, including standard keyboard, expanded keyboard, joystick, single and multiple switch scanning. (O Connor, 2007)

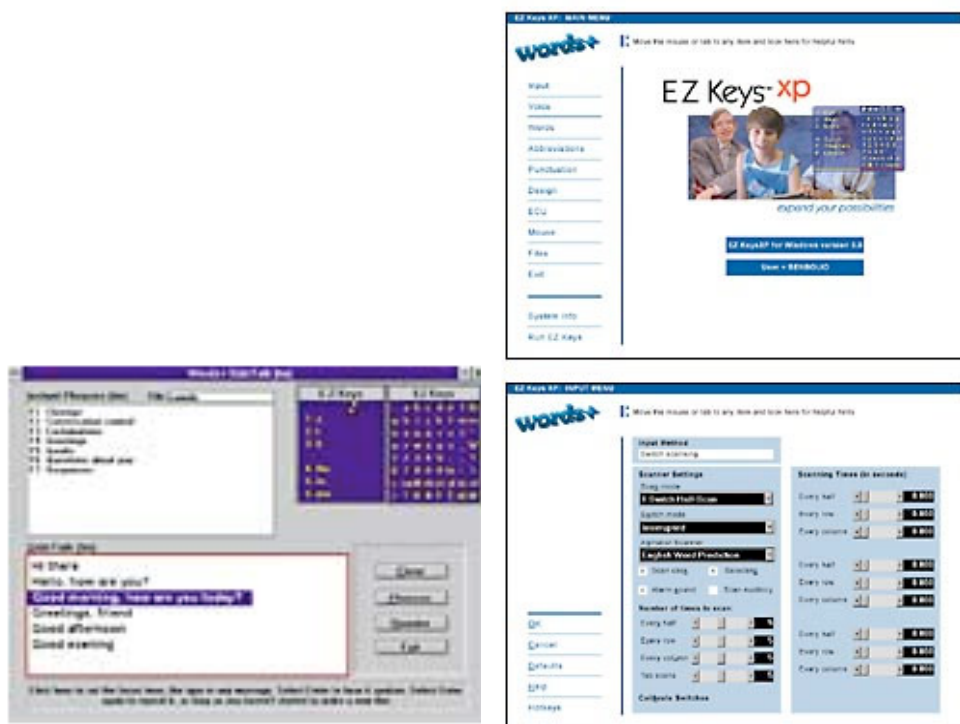


Figure 22: EZ Keys Software

2.13 Universal Design

One of the most exciting developments - in terms of design for inclusion - in recent times has been 'Universal Design'. Universal Design as a term can be understood to be interchangeable with 'Design For All'.

Universal Design can be defined as:

"The design of products and environments to be usable by all people, to the greatest extent possible, without the need for adaptation or specialized design."

The 7 Principles of Universal Design were developed in 1997 by a working group of architects, product designers, engineers and environmental design researchers, led by the late Ronald Mace in the North Carolina State University. (Principles of UD, 2010)

The purpose of the Principles is to guide the design of environments, products and communications. According to the Centre for Universal Design in NCSU, the Principles "may be applied to evaluate existing designs, guide the design process and educate both designers and consumers about the characteristics of more usable products and environments." (CEUD UD Principles, 2010)

2.13.1 Principle 1: Equitable Use

The design is useful and marketable to people with diverse abilities.



Guidelines:

- 1a. Provide the same means of use for all users: identical whenever possible; equivalent when not.
- 1b. Avoid segregating or stigmatising any users.
- 1c. Provisions for privacy, security, and safety should be equally available to all users.
- 1d. Make the design appealing to all users.

2.13.2 Principle 2: Flexibility in Use

The design accommodates a wide range of individual preferences and abilities.

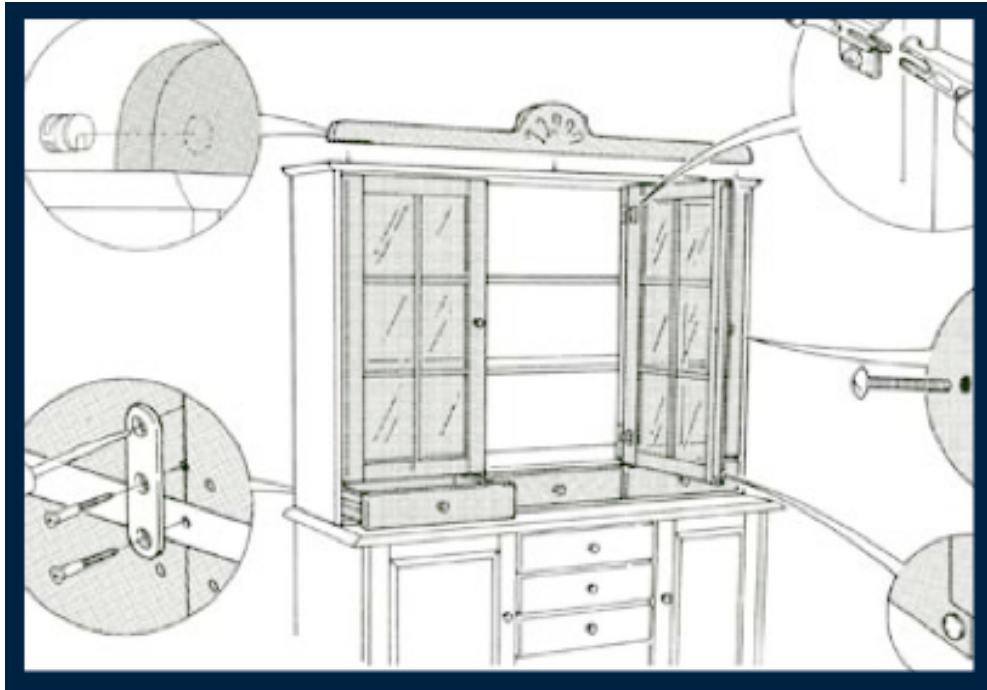


Guidelines:

- 2a. Provide choice in methods of use.
- 2b. Accommodate right- or left-handed access and use.
- 2c. Facilitate the user's accuracy and precision.
- 2d. Provide adaptability to the users pace.

2.13.3 Principle 3: Simple and Intuitive Use

Use of the design is easy to understand, regardless of the user's experience, knowledge, language skills, or current concentration level.



Guidelines:

- 3a. Eliminate unnecessary complexity.
- 3b. Be consistent with user expectations and intuition.
- 3c. Accommodate a wide range of literacy and language skills.
- 3d. Arrange information consistent with its importance.
- 3e. Provide effective prompting and feedback during and after task completion.

2.13.4 Principle 4: Perceptible Information

The design communicates necessary information effectively to the user, regardless of ambient conditions or the user's sensory abilities.



Guidelines:

- 4a. Use different modes (pictorial, verbal, tactile) for redundant presentation of essential information.
- 4b. Provide adequate contrast between essential information and its surroundings.
- 4c. Maximize legibility of essential information.
- 4d. Differentiate elements in ways that can be described (i.e. make it easy to give instructions or directions).
- 4e. Provide compatibility with a variety of techniques or devices used by people with sensory limitations.

2.13.5 Principle 5: Tolerance for Error

The design minimizes hazards and the adverse consequences of accidental or unintended actions.



Guidelines:

- 5a. Arrange elements to minimize hazards and errors: most used elements, most accessible; hazardous elements eliminated, isolated, or shielded.
- 5b. Provide warnings of hazards and errors.
- 5c. Provide fail safe features.
- 5d. Discourage unconscious action in tasks that require vigilance.

2.13.6 Principle 6: Low Physical Effort

The design can be used efficiently and comfortably and with a minimum of fatigue.



Guidelines:

- 6a. Allow user to maintain a neutral body position.
- 6b. Use reasonable operating forces.
- 6c. Minimize repetitive actions.
- 6d. Minimize sustained physical effort.

2.13.7 Principle 7: Size and Space for Approach and Use

Appropriate size and space is provided for approach, reach, manipulation, and use regardless of user's body size, posture, or mobility.



Guidelines:

- 7a. Provide a clear line of sight to important elements for any seated or standing user.
- 7b. Make reach to all components comfortable for any seated or standing user.
- 7c. Accommodate variations in hand and grip size.
- 7d. Provide adequate space for the use of assistive devices or personal assistance.⁹

2.13.8 Conclusions

This chapter has given us an overview of Assistive Technology, WCAG, Accessibility, Interaction Design, Universal Design and a brief look at Software development methods.

⁹ All images on Universal Design © Copyright 1997 NC State University, Center for Universal Design, College of Design

It also looked at and how they may relate to inclusive design as well as some of the issues that face users with vision impairments. This chapter gives a sense of how diverse the field of inclusive design is and the many factors that have to be considered.

Also some of the benefits of designing accessible websites and applications were mentioned such as improvements to customer satisfaction and increases sales. These benefits are happy by products of inclusive or Universally Designed products and services. The next chapter will build on what has been covered here and explore User-Centred Design and User Evaluation Methods in more detail.

3 USER-CENTERED DESIGN (UCD) AND EVALUATION METHODS

3.1 Introduction

In the previous chapter we gave an introductory overview of what UCD is and outlined some related principles and practical processes used to achieve it. This chapter covers more user evaluation methods. Some of these may be considered to be ‘secondary’ or supportive evaluation methods whereas user testing with people with disabilities (or indeed any user type) could be considered a ‘primary’ evaluation method due to its rather immediate and in some ways, intimate, level of user involvement.

This chapter will cover areas like participatory design, using focus groups and surveys for research, expert evaluation, use of personas, how to evaluate the effectiveness of these methods and usability and standards. Firstly, if accessibility is mostly about people with disabilities, what is usability?

3.2 What is Usability?

Usability looks at the quality of the user experience and attempts to understand how to improve it. Usability as a discipline attempts to determine how successfully a user can complete a task and how satisfying a device or interface may be to use. This can be for groups of users such as vision impaired or blind, older people and but also for users without disabilities.

In terms of a definitions of usability there are several:

“A measure of how easy it is for a user to complete a task. In the context of Web pages this concerns how easy it is for a user to find the information they require from a given Web site.” (HCI Glossary 1, 2010)

This definition is very much focused on the user being able to complete a specific task, which is obviously very important.

“The ease with which a system can be learnt or used. A figure of merit or qualitative judgment of ease of use or learning. Some methods of assessing usability may also express usability as a quantitative index.” (HCI Glossary 2, 2010)

This second definition is interesting as it mentions how easily the system can be ‘learnt’ by the user. A good rule of thumb in user interface design is if you have to provide instructions on how to perform particular tasks, it's already too complicated! The user should just ideally intuitively ‘get it’. This is, of course, in some situations impossible. The user won’t just ‘get’ how to fly a plane for example.

“The effectiveness, efficiency, and satisfaction with which specified users can achieve specified goals in a particular environment. Synonymous with ‘ease of use’.” (HCI Glossary 3, 2010)

This third definition is one of the most interesting as it goes beyond dryly looking at the ‘tasks’ the user needs to do and mentions the level of satisfaction the user will feel when they use a web interface. This takes the usability definition to a higher level by looking at the quality of the user experience and not merely a task-based approach. This is where user testing is very useful as it is a fantastic way of assessing the quality of the user experience.

Donald Norman, one of the fathers of usability, has this to say on his website:

“I caution that logical analysis is not a good way to predict people's behavior (nor are focus groups or surveys): observation is the key. I caution that the time frame for adoption of new technologies is measured in decades, not the months everyone would prefer. And I help formulate new products and services. For both products and services I'm a champion of beauty, pleasure and fun, coupled with behavioral and functional effectiveness.” (HCI Glossary 3, 2010)

Usability is about looking at how usable, intuitive, user friendly and simply satisfying an interface is to use. As a discipline, it also examines the psychology of user interaction. It is an attempt to understand how users perceive the instructions that they receive from looking at or interacting with a user interface or device.

While accessibility and usability are two different fields, there is a very strong relationship between the two. The following techniques are often used in the preparatory phase of a project and if sufficient care is taken to use these techniques well, then these requirements gathering and prototyping phases can really help to avoid very serious mistakes in a UI design further on in the project.

3.3 Participatory Design

This is a technique where there may be one or more end users on the design team itself. The user is put at the heart of the process by having their knowledge, skill set and emotional responses tapped by the designers. They may be however consumed by the process itself and gradually lose their own impartiality as a user thus diluting the effectiveness of their feedback and involvement. (Rubin, 1994)

3.4 Focus Group Research

Focus group research aims to evaluate the project's basic concepts at an early stage in the development process. It can be used to identify and confirm the characteristics of the user, and also to validate the projected effectiveness of the product. It usually involves multiple participants.

The concepts to be explored can be presented to the group as paper and pencil drawings, storyboard, PowerPoint presentations, 3D prototypes and models etc. The idea is to identify how acceptable the concepts are and in what ways they can be improved. Focus groups can be used to explore the user's feelings in great depth. (Rubin, 1994)

3.5 Surveys

Surveys are often used to try to understand a user's preference about an existing product or a potential product. In this domain, the survey is in some ways a more superficial way of collecting data than the focus group but it is still useful in particular to draw a potential picture of the views of a larger population. They can be used at any time but are often used at the beginning of a product development cycle.

Thorough survey design is very important and a great deal of thought must go into survey design in order to ensure that questions are clear and unambiguous in order to get the best use from the returned data.

3.6 The Cognitive Walkthrough

The cognitive walkthrough is a common technique for evaluating the design of a user interface, with special attention to how well the interface supports ‘exploratory learning,’ i.e., first-time use without formal training. This evaluation can be performed by the system designer, in the early stages of design before empirical user testing is possible.

Early versions of the walkthrough method relied on a detailed series of questions, to be answered on paper or electronic forms. These could take the form of ‘Paper and Pencil Evaluations’ which are a very useful way of finding out about user preference for certain attributes of a user interface, such as organisation and layout of menu or other controls.

‘Paper and Pencil Evaluations’ are very useful in that designers can find out critical information very quickly and inexpensively and get some real feedback about how intuitive a user interface may be before any development work has taken place. This technique can be used as often as necessary and can be used in conjunction with or instead of prototyping software such as Serena Prototype Composer or Axure. (Riemann.J, Franzke. M, Redmiles. D, (1995), (Rubin, 1994) (Serena, 2010), (Axure, 2010)

3.7 Expert Evaluations

This is where a usability specialist who has little to do with the project is brought in to assess its usability. Usability principles are used to assess the quality of the system and any potential problems it may have. This may be performed in conjunction with an accessibility audit of the system to see how usable it would be by people with disabilities using Assistive Technology.

Outputs from this kind of expert evaluation would be a usability and/or accessibility audit. (Rubin, 1994)

3.8 Using Personas

In some instances there is no ability to user test at all. It just may not be logistically possible, so this is where using personas can be useful. A Persona is like a distilled archetype of a certain user group's qualities and attributes. These attributes are models of the various qualities a user experience professional thinks may epitomise a certain user group – such as blind people. They therefore build a persona around them.

Persona use aims to simulate what the experience of using a website may be like for this group of users. If various personas are accurate, then the simulation of their experience will hopefully be also. Personas can be used as a basis to justify the modification of an application design around the perceived needs of the persona.

3.8.1 Building Personas

Personas are created from the gathered research about a target group; this can be from surveys, interviews and so on. It is possible build imaginary personas that represent an average user. These various groups can include older people, young people, blind users, and so on. A good persona does come from real world feedback that has been gathered from real users.

3.8.2 Does using Personas Work?

While personas are in wide use, there is very little empirical evidence to support the claim that using personas is actually beneficial to improving the quality of the user interface design (Cooper and Riemann, 2003). In a very interesting field study the effectiveness of using personas was investigated. This took the form of an experiment conducted over a period of 5 weeks using students from the National College of Art and Design in Dublin. The results showed that, through using personas, designs with superior usability characteristics were produced. The results also indicated that using personas provides a significant advantage during the research and conceptualisation stages of the design process (supporting previously unfounded claims).

The study also investigated the effects of using different presentation methods to present personas and concluded that photographs worked better than illustrations, and that visual storyboards were more effective in presenting task scenarios than text only versions (Frank Long, 2009).

3.8.3 Measuring the effectiveness of using Personas

Long's study produced objective evidence to support the key claims made by Cooper et al for using personas in the product design process. Using Personas seemed to strengthen the focus by designers onto the end user, their tasks, goals and motivation. Personas make the needs of the end-user more explicit and thereby can direct decision-making within design teams more towards those needs. The study also suggests that using personas can improve communication between teams and facilitate more constructive and user-focused design discussion.

Students using personas produced designs with better usability attributes than the students that did not use personas - thereby answering one of Chapman and Milham's key concerns about persona usage and outputs. (Chapman, C.N and Milham, R.P , 2006)

Chapman and Milham were concerned about the claim that the use of Personas was effective at all. In fact they suggested that they could indeed be harmful and lead to skewed and incorrect conclusions, and were therefore unreliable. They asked, *"How many users are represented by this persona?"*, *"Is this persona relevant for a group?"*, *"Are personas a valid method at all (and how can this be verified)?"* (Chapman, C.N and Milham, R.P, 2006)

Long also found that using illustrations instead of photographs of the persona seem to reduce effectiveness and empathy towards the illustrated persona. Also use of a storyboard task scenario was more effective than the text version and facilitated more detailed design solutions.

Long finally concluded that using personas offers several benefits for user-centred design in product development, enhancing the possibility of incorporating user-centred features at the product specification stage and provided some objective evidence that using personas does work. (Frank Long, 2009)

3.9 Field Studies

This is where a product or interface is tested in its natural setting. This could be an office, home or any other realistic environment that will reflect how the product will be used. This is usually conducted late in the product cycle and is not used as an indicator of significant issues with the product or interface but as a way of refining it. (Rubin, 1994)

3.10 Criteria for evaluating of Usability or User Evaluation Methods (UEM)

Traditionally user testing in the lab has been the stable of assessing the quality of the user experience for the end user. Other User-based evaluation methods included verbal protocols (Ericsson & Simon, 1984), critical incident reporting (del Galdo, Williges, Williges, & Wixon 1987) and user satisfaction ratings (Chin, Diehl, & Norman, 1988). Usability was often just bolted on at the end of the development process, so other expert evaluation methods came into vogue in the 80s and 90s.

These include:

- Guideline Reviews based on interaction design guidelines such as those by Smith and Mosier (1986)
- Heuristic Evaluation (Nielsen & Molich, 1990)
- Cognitive Walkthroughs (Lewis, Polson, Wharton, & Rieman, 1990; Wharton, Bradford, Jeffries, & Franzke, 1992)
- Usability Walkthroughs (Bias, 1991)
- Formal Usability Inspections (Kahn & Prail, 1994)
- Heuristic Walkthroughs (Sears, 1997)

3.10.1 Can comparison of user evaluation methods be meaningful?

It can be very difficult for usability professionals to get a true picture of the User Evaluation method (UEM) that is best suited to any given project. Whether it is user testing, focus groups, prototyping and so on. Some of the main issues are:

- 1) The ‘evaluator effect’. This is where different test facilitators or evaluators come up with varying results for the same data set.
- 2) Lack of scientific rigour when applying usability evaluation techniques. This has had the net effect of greatly diluting the reliability of much user data that is collected during a user test or other usability evaluation method.
- 3) There is a general lack of appropriate standards or metrics that can be used to compare evaluation methods.

In an attempt to measure the effectiveness of incorporating ‘real’ user data into a project and thereby choosing a suitable UEM two methods were compared (Chattratchart, J. Brodie, J. 2004). They were Nielsen’s Heuristic Evaluation (HE) and HE-Plus, which is a modified extension of HE.

The main difference in the HE-Plus method is that the evaluators are given a list of the common problems identified with the given product or interface being evaluated. Apart from that they are identical.

The criteria used for assessment were “thoroughness, validity and effectiveness” (Hartson et al) and the outcome was the HE-Plus method was shown to be more effective. The outcome of this research undertaken by Chattratchart and Brodie cannot be considered to be a truly useful way of assessing the value of diverse user evaluation methodologies. They have merely taken exactly the same methodology and by having made some additional a priori knowledge available, which by the very nature of its availability would certainly have improved the test outputs, they have then stated that the HE-Plus method is a better methodology.

While this may be useful as an academic exercise - or be indicative of the need for future research - the methodologies were not diverse enough for the outcomes to be considered noteworthy.

3.10.2 Iterative Design Process

Much is made in usability circles of the importance of a responsive or iterative design process, however, according to there is little agreement of exactly how to achieve this (H.Rex Hartson, Terence S. et al). In principal, these steps encapsulate the essence of the iterative design process and ideal would be to include user involvement as early on as possible in each of these stages. With the results of each usability test of the initial, prototype, and final design (Kies, Williges, and Rosson 1998) stages being fed into the each consecutive stage – thus producing a product that has considered real world feedback from user into its very core.

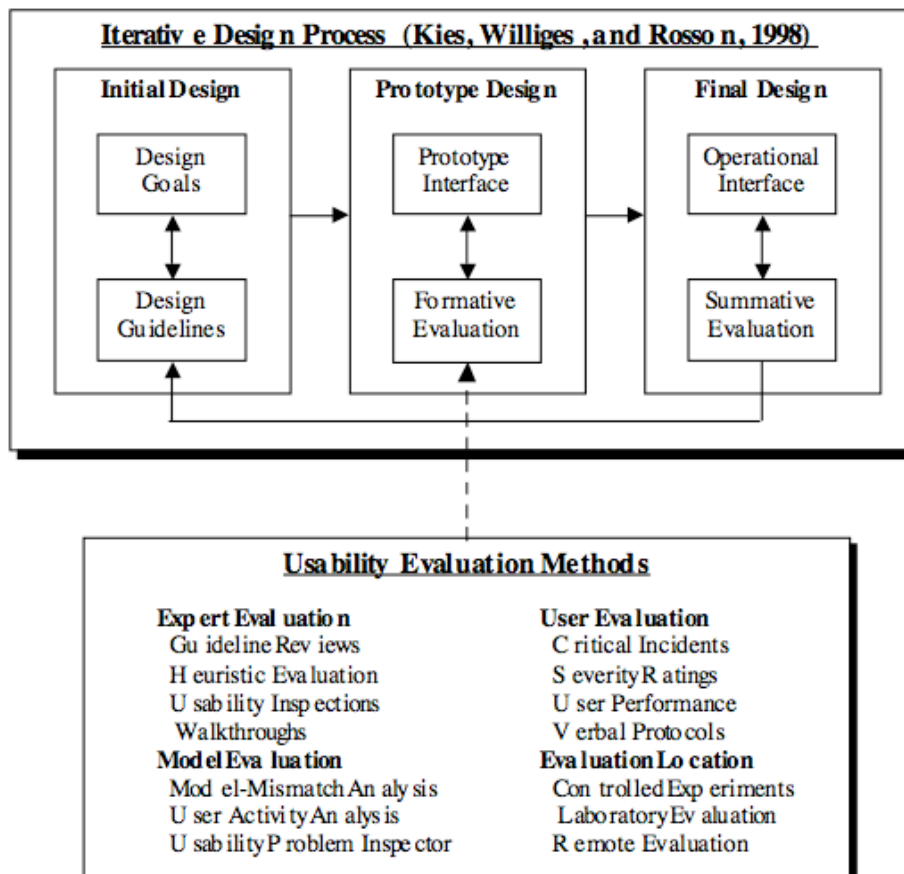


Figure 23 UEMs used in Formative Usability Evaluation (from H.Rex Hartson, Terence S. et al)

The term usability evaluation method (UEM) is used (by H. Rex Hartson, Terence S. et al) to refer to any method or technique used to perform usability evaluation of an interaction design at any stage of its development. They use this term to include lab-based usability testing with users, heuristic and other expert-based usability inspection methods, model-based analytic methods, expert evaluation, and remote evaluation of interactive software after deployment in the field. While this research is primarily concerned with user testing, examining other forms of evaluation in use in the wild today is very important for a true sense of context.

In their paper “Criteria for Evaluating Usability Evaluation Methods” (H. Rex Hartson, Terence S. Andre, and Robert C. Williges, 2001) assert that there is a consensus in the usability community that:

- Usability is seated in the interaction design
- An iterative evaluation-centred process is essential for developing high usability in interaction designs
- Usability, or at least usability indicators, can be viewed as quantitative and measurable
- A class of usability techniques called UEMs have emerged to carry out essential usability evaluation and measurement activities.

However, they state that there is still little agreement on the various methods of user evaluation methods (UEM) and there is a genuine lack of knowledge amongst usability professionals as to the merits and weaknesses of UEMs.

They note that this may come from a lack of:

- Standard criteria for comparison
- Standard definitions, measures, and metrics on which to base the criteria
- Stable, standard processes for UEM evaluation and comparison.

Lund (1998) then suggested that there is a lack of standard usability metrics and this deficiency is exacerbating the situation. This need for metrics in usability evaluation suggests there is a missing framework that is needed to measure, record, and compare usability data to ensure maximum effectiveness for both practitioners and consumers of usability data.

3.11 Usability Methodologies and Standards

3.11.1 ISO and Usability

In research into ‘Human-Computer Interaction Standards’ (Bevan, 1995) outlined how work on international standards for HCI *“has not been about precise specification, but instead has concentrated on the principles which need to be applied in order to produce an interface which meets user and task needs.”*

Bevan suggests that these standards broadly fall into two categories. One is a *“top-down” approach “[...] concerned with usability as a broad quality objective: the ability to use a product for its intended purpose”*.

The other is a product-oriented ‘bottom-up’ view *“ [...] concerned with aspects of the interface which make a system easier to use”. The broad quality view originates from human factors, and standards of this type are applicable in the broad context of design and quality objectives. The product-oriented view “[...] relates more closely to the needs of the interface designer and the role of usability in software engineering”* (Bevan, 1995).

3.11.2 Usability as a quality objective

As a high-level quality objective, ISO 9241-11 defines usability 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.” (Bevan, 1995)

Standards of this type can be used to support the following activities:

- Specification of overall quality and usability requirements and evaluation against these
- Requirements (ISO 9241-11 and ISO/IEC 14598-1)
- Incorporation of usability into a quality system (ISO 9241-11)
- Incorporation of usability into the design process (ISO/IEC 13407)

In the product-oriented view (which concentrates more on the design of specific attributes, and relates more closely to the needs of the interface designer) usability is seen as one relatively independent contribution to software quality and is defined in ISO/IEC 9126 as:

“A set of attributes of software which bear on the effort needed for use and on the individual assessment of such use by a stated or implied set of users.”

Other standards that deal with usability in terms of attributes, which must be designed into a software product to make it easy to use:

- ISO 9241: Ergonomics requirements for office work with visual display terminals:
- Part 10, 12-17: Dialogue design
- ISO/IEC 10741-1: Dialogue interaction - Cursor control for text editing
- ISO/IEC 11581: Icon symbols and functions
- ISO/IEC 9126: Software product evaluation - Quality characteristics and guidelines for their use.

These standards can be used in the following ways:

- To specify details of appearance and behaviour.
- To provide detailed guidance on the design of user interfaces.
- To provide criteria for the evaluation of user interfaces.

ISO 9241-11 can be used to help understand the context in which the above attributes may be needed. (Bevan, 1995)

Note that at time of writing the ISO series covering usability including ISO 13407 is being redrafted as ISO 9241-210 ‘Human-centered design for interactive systems’’. This draft makes significant references to accessibility. [ISO 9241-210, 2011]

There are also 3 new ISO guidelines from the 9241 series covering various aspects of accessibility published in 2008:

- ISO 9241-20:2008 Accessibility guidelines for information/communication technology (ICT) equipment and services.
- ISO 9241-171:2008 Guidance on software accessibility.
- ISO 9241-151:2008 Guidance on World Wide Web user interfaces.

It is interesting to note how the following definitions of usability vary depending on context, so how can we assess which is an ideal methodology? Is it a case of one-size fits all? Do some methodologies suit more formal or informal settings? These questions will be looked at in the Knowledge Audit section of this research.

Usability testing is a way of gathering empirical data on how users perform representative tasks. The kind of testing that is done is roughly divided into two types, formal and informal.

Formal user testing is closer to true ‘science’ or true ‘experiment’ – in order to confirm or to refute a hypothesis. The second approach is less formal and uses an iterative cycle of tests that are intended to expose usability problems and gradually shape and improve the product or interface in question.

Informal testing is the main testing type of interest in this research as it reflects more ‘real world’ usability analysis in terms of the context of the testing (commercial projects etc) and the budgetary and time restraints that they have.

3.12 Conclusions

This chapter aimed to introduce some User Evaluation methods and demonstrate some of the difficulties in comparing them definitively. Many diverse processes used in the field of UCD were covered and some interesting research was referenced that aimed to critically assess the use of these methods, such as Personas. Also the efforts of international Standardisation bodies like ISO in the field of UCD were introduced.

The issue of trying to get a definitive picture of the strengths and weaknesses of current practice in the field of user testing and usability analysis is core to this work. In the next chapter the important issue of trying to measure and understand the quality of evaluation methodologies is explored in greater detail.

4 CASE STUDIES

4.1 Introduction

This chapter looks at how current practices came to be by exploring previous research that evaluates user testing and usability methodologies in general. By taking a critical look at how the relatively new field of usability testing for the web was defined in the 1990s by the most influential leaders in the field this section explores how some of the assumptions about the best way to conduct user testing and usability studies came to be - via some interesting case studies.

The case study approach helps illustrate the complexity of the diverse elements such as the human elements and the technology elements that are involved in effective user testing.

This section also deals with User Testing methods in a detailed way, a more general introduction to user testing and methodologies is given in ‘Section 2: Knowledge Audit Background’.

4.2 Remote User Testing

Remote user testing is an automated technique where test participants access the site being assessed for usability issues from their preferred location using their normal browser (Hartson, H. R., Castillo, J.). Tasks can be presented to the user in a small browse window at the top of the screen that can be used to capture their input. Remote testing can be used as an alternative, where cost is an issue (due to staff, limited number of test participants etc).

So how does remote testing compare with lab based testing? Some research shows that remote testing yields very similar results and provides a rich set of data. (Hartson, H. R., Castillo, J.)

Remote testing has some advantages of lab based testing:

- 1) Remote testing uncovers more usability issues. This could be due to the often much larger sample sizes involved in remote testing.
- 2) User feedback from remote testing can be quite detailed.
- 3) Lab based testing, due to the small sample sizes involved (typically 5-8) may not give a clear indication of all usability issues and as a methodology may be flawed. A small number of test participants make it difficult to draw substantial conclusions from subjective user experience.
- 4) Remote testing captures the feedback of a more diverse range of users.

Both techniques (remote and lab based) showed up critical usability issues and seem to in general, capture similar info about site usability. Remote testing however, may lack a certain dynamic that is inherent in facilitated user testing, but this may be ok for some kinds of testing and could help produce some statistically significant results and help with qualitative metrics that could be used to back up design decisions made on the basis of usability analysis.

4.3 Case Study #1: Comparative Evaluation of Usability Tests

In a field study by Molich *et al.* (1999), which looked at how four separate usability labs would undertake a similar job, some significant results were uncovered. In general the main type of user testing is where a usability lab tests an application but what if several usability labs test the same application and the outputs are analysed? The idea was a simple comparative evaluation that attempted to highlight the differing approaches and distil their various strengths and weaknesses.

The purpose of the exercise was to:

- Demonstrate the variety of some of the many different approaches to professional usability testing that are being applied commercially today.
- Show each team the comparative strengths and weaknesses of its approach to conducting and reporting usability tests.

- Provide an informed basis for an ongoing discussion of whether professional usability testing is an art or a mature discipline that turns out reproducible results.
- This paper compares process, reporting and results. The paper discusses the difference between usability testing and good usability testing.

The following usability labs participated in the evaluation study:

- HFRG (Human Factors Research Group), University College Cork (Ireland)
- National Physical Laboratory (UK)
- Rockwell Software (USA)
- Sun Microsystems, JavaSoft Division (USA)

They will be referred to as teams A,B,C,D.

4.3.1 Test Application

The application to be tested was the English language version of Task Timer for Windows, version 2 (TTW). TTW is a calendar program written by the Danish software house DSI for the Danish company Time/system. Task Timer is rather like a forerunner of Microsoft Outlook.

The usability test scenario was as follows:

Time/system® is a Danish company that manufactures and distributes paper calendars. In the fall of 1994 Time/system released Task Timer for Windows version 2 as a computer version of the paper calendar.

The primary user group for TTW is professional office workers; typically lower and middle level managers and their secretaries. Time/system also offers the demo version of TTW freely to anyone at hardware and software exhibitions, conferences, and "events", e.g. Microsoft presentations. Time/system hopes that the demo version will catch the interest of people who pick it up by chance.

TTW is intended for users who have a basic knowledge of Windows. Familiarity with the paper version of the calendar or with other electronic calendars is not required.

Time/system is planning to send out version 3 of TTW in April 1998. However, their

sales staff have heard negative comments about users' initial experience with the program, and TTW faces stiff competition from other programs, like Microsoft Schedule.

They have therefore asked you to perform a cheap usability test involving e.g. five typical users to test the usability of the software for new users of the product.

Task Timer for Windows is a diary, task and project management program for individuals and work groups. To reduce cost, you have agreed with Time/system to focus mainly on the diary and address book functions for individuals. In other words: Do not test task management, project management, networking functions, etc.

Each lab was asked to use its standard usability report format with one exception: The name of the company should not be directly or indirectly apparent from the report. Therefore, the usability labs are referred to as Team A, B, C, and D. In addition, each usability lab was asked to report in an addendum:

- Deviations from its standard usability test procedure.
- Resources used for the test (person hours).
- Comments on how realistic the exercise had been.

The labs did not communicate during the test period.

4.3.2 Test Output

The desired output from the user tests completed by each of the usability labs was a detailed report. Each of the labs test facilitators who participated was also interviewed after the test. It is not necessary to get into the details of the test itself – it is sufficient to set the context and then look at the outcomes of the research.

Table 1 presents comparative data about the usability test processes applied by the teams.

Team	A	B	C	D
1. Total person hours used for the test by the usability professionals. Test participants' time is not included. Equal to the sum of the following rows 2-4.	26	70	24	84
2. Time used for planning and usability context analysis	9	10	6	28
3. Time used for recruiting test participants and testing. Test participants' time is not included.	12	20	8	21
4. Time used for analysis of results and reporting	5	40	10	35
5. Number of usability professionals involved	2	2	1	3
6. Number of tests	18	5	4	5
7. Approximate length of each usability test in minutes	4 to 32	120	120	60
8. Profiles of test participants reported	No	No	Yes	Yes
9. Number of scenarios/tasks used in test	5	11	5	4
10. Detailed scenario descriptions provided (see also table 5)	Yes	Yes	Partly	Yes
11. Quantitative assessment of user interface provided	Yes	No	No	Yes
12. Results of heuristic evaluation performed by usability professional included in report	No	No	Yes	No

Figure 24: Comparison of test processes ¹⁰

Both teams B and D spent a great deal more time working on the test? Did it yield a greater qualitative result?

4.3.3 Quantitative Usability Measurements

In general the purpose of a usability test report can be to enable developers to make informed decisions about whether a piece of software should be released or revised.

The Software Usability Measurement Index (SUMI) questionnaire is an industry-standard evolution questionnaire for asserting the value of software for end-users. Team A and D provided quantitative assessments of the usability of TTW, which are useful for this purpose. Both measurements are based on the SUMI questionnaire and can thus be compared – see figure 3.

- The SUMI questionnaire provides numeric assessments on the following scales:
- **Efficiency:** degree to which user feels he gets his work done well with the system
- **Affect:** degree to which user feels the system is enjoyable and stress-free to use

- **Helpfulness:** degree to which user feels the system helps him along
- **Control:** degree to which user feels in control of the system, rather than vice versa
- **Learnability:** degree to which user feels he can learn new operations with the system

There is also a Global usability score, which is a combination of items from each of the above scales.

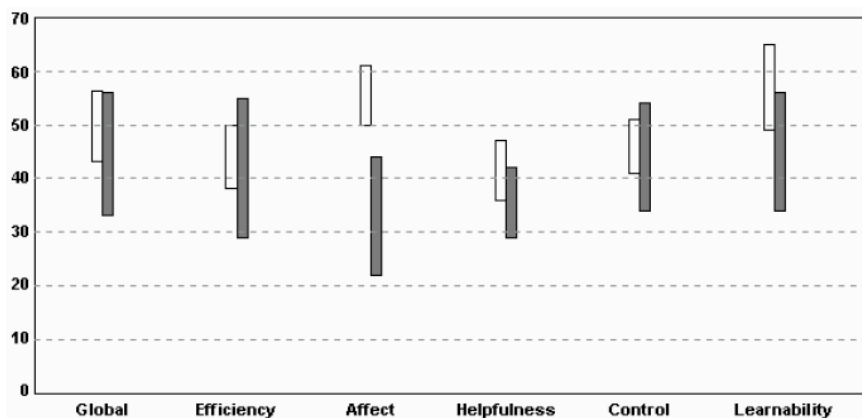


Figure 3. Comparison of quantitative usability measurements. Each column represents the 95% confidence interval around the median (the median is not shown) – that is, the range within which we are 95% certain that the true median of the user population lies. White columns represent results from Team A, grey columns represent results from Team D. It should be noted that Team D used far fewer usability test participants than did Team A, which probably explains the larger 95% confidence intervals shown in Figure 3. Comparison between Team A and Team D results using SUMI must also take account that Team D included the software installation task and Team A did not. Since both teams reported the number of users involved and the tasks evaluated it is possible to make a meaningful interpretation of the differences in the two profiles.

11

4.3.4 Qualitative Reporting

The user testers report has traditionally been the main way that developers can find out how best to improve the design of their User Interface. In this case Teams, B, C, and D provided reports.

4.3.5 Observations

What is highly significant is that in the paper, the authors outline the differing methodologies used by the test facilitators. The methodologies used by teams B, C and D were very similar. They recruited 4-5 users and set them real world tasks, observed the user undertake the tasks and note any difficulties etc. Team A, did not do this at all.

¹⁰ Chart from Comparative Evaluation of Usability Test (Molich et al)

¹¹ Chart from Comparative Evaluation of Usability Test (Molich et al)

They formed two groups of a total of 19 users, they were all set tasks but they were not observed in real time but were given questionnaires to fill out after the test. They were asked what they did and did not like, what were their favourite and least favourite aspects of the application and so on. This had the net effect that the test facilitators had far less feedback than the other groups.

Is this because the sample size was not vastly larger? In the previous paper didn't they report that they had more feedback because of the larger sample size using the questionnaire system?

The SUMI results from teams A and D were very similar. As these results were obtained by completely independent teams, this result is very interesting and could be put down to the 'evaluator effect'.

The amount of time taken by the various groups and their corresponding outputs is also interesting as there were wide ranges in the time reported by the teams: from 26 hours by Team A and 24 hours by Team C to 70 hours by Team B and 84 hours by Team D.

4.3.6 Differences in time and outputs

- Both team A and team C produced high quality outputs in dramatically less time than teams B and D. This included a high quality 25 page report by Team C in only 10 hours.
- Team A took 26 hours to test 18 users. It apparently took 40 minutes per user to administer a 20 min task and 3 questionnaires, which took 10 min to complete. It then took an average of 3 minutes to analyse each questionnaire, and two hours to produce a 38-page report.
- Team A commented that the use of standardised tools and software is the key to their increased efficiency. However, they also add that this increased efficiency comes at a cost: less attention can be given to specific diagnostics of poor interface features.

4.3.7 Test Report Comparison

Team	A	B	C	D
1. Number of pages in report, excluding blank pages	38	18	25	66
2. Number of pages in main report, excluding blank pages and appendices	8	18	12	22
3. Number reports submitted	1	1	1	2
4. Length of executive summary in pages. An executive summary is recommended in [1] and [2]	½	Not provided	2 (Entitled "Human Interface Targets")	1½
5. Number of screen shots provided to illustrate problems	0	0	11	0
6. Quotations from test participants provided Recommended in [1]	Yes, in appendix	A few	Included in detailed 8-page log	No

Table 2. Comparison of usability test reports.

12

The authors note that there are some important considerations in writing and delivering a test report:

- 1) A usability report that is ignored by the developers is useless.
- 2) The report must be short to be effective. In this case study, the reports from team A and D contain a lot of detailed information about the SUMI method. This information has wisely been put into appendices or into a separate report. Nevertheless, the total size of these reports is considerable; there is also a risk that developers will not even look at it sufficiently closely to realise that they do not have to read it all.
- 3) Team A comments: "By putting this information in an appendix, we are quite clearly saying to our readers: "You don't have to read this unless you want to." We cannot help readers who don't even open a report... yet you will notice for the Team A report the first thing you see as you open the front page is 'Summary of Findings and Recommendations' "
- 4) The report must be easy to understand. All reports lived up to this requirement with one exception. The reports that used the SUMI method contained some statistical information that can be hard to understand. Also, the SUMI method compares the usability of TTW to a large body of programs whose usability serve as a standard reference.

- 5) This process is complicated, and may be difficult to sell to skeptical developers. Team D disagrees: *"Actually, we find quite the contrary is true - developers usually love SUMI results and can't wait to get hold of them! A score of above or below 50 is easy to interpret as a simple pass/fail criterion for usability! We usually have to invest quite a bit of effort in dissuading developers and managers from taking SUMI results at face value and jumping to conclusions."*
- 6) Team A also disagrees: *"The quantitative data that Team D and Team A have shown in the body of their report is the absolute minimum necessary to allow a reader to understand what is going on. There is a graph, and there are verbal conclusions. Everyone can see from the graph that most of the profiles are 'below the line' (i.e. 50). I think you are making too many negative assumptions about the consumers of this report. In my experience, sales staff are extremely receptive to these kinds of statistics, and they are well used to market survey results."*
- 7) The report should be attractive and well laid out.

¹² Chart from Comparative Evaluation of Usability Test (Molich et al)

Team	A	B	C	D
1. Number of reported problems	4	98	25	35
2. Number of reported problems that include specific recommendations for improving the interface	0	24	6	35
3. Number of reported problems that were encountered by one user only	0	4	2	8
4. Number of reported problems that deal exclusively with aesthetics (choice of colors, etc.)	0	0	5	1
5. Problems classified by severity Recommended in [1]	All four problems are severe	No	No	No
6. Number of positive findings reported. Recommended in [1]	1	4	3	0
7. Number of reported suggestions from test participants for improving the interface	0	2	5	0
8. Number of program errors reported	0	1	0	0
9. Indication of how many users encountered each problem Recommended in [1]	Yes	No	No	No

Table 3. Comparison of usability test results.

	Team A	Team B	Team C	Team D
Team A	-	2	1	1
Team B	2	-	3	8
Team C	1	3	-	5
Team D	1	8	5	-

Table 4. Problems found by more than one team. The table shows for instance that eight problems were found by both team B and D. Only one problem was found by all four teams. Another problem was found by three teams, namely B, C and D. Eleven problems were found by two teams.















	Installation	Log-in	Familiarization with TTW	Basic Calendar	Advanced Calendar (Recurring appointment, group appointment)	Basic Address and Telephone Book
Team A						
Team B						
Team C						
Team D						

Table 5. Scenarios used by the teams. In the familiarization scenario the user is asked to take a few minutes to explore TTW.

4.3.8 Test Results and Observations

What is significant is that the overlap between the problems that each group found is very small. Meaning different teams found different problems with often only as small a number of 8 being similar say for example between the 98 problems identified by team B and the 35 problems found by team D.

The researchers suggest that the discrepancies between the teams may be as a result of variances in methodology. Because of the slight differences in the scenarios that each team undertook, and they suggest that Team B did not test the installation procedure and this may account for some of the lack of overlap. This might not be whole story since team B had the greatest reported usability bugs.

The low number of overlap with issues may indicate that small sample sizes are just not sufficiently significant for reflecting the number of true usability issues in an application? Or is it just that all user tests will throw up different results that can be hard to anticipate via accessibility testing or any other standardised way of assessing a user interface? Or was it merely that the software itself had many, many usability problems that the results seem orthogonal?

4.3.9 Test Team Observations

In a post-test questionnaire the teams that took part made some very interesting observations about their experience. Both Teams B and C noted that the lack of the developer involvement in the process was very much contrary to what they would consider best practice in the usability. They both noted that they could not stand beside how efficient or useful it was to do this kind of user testing or assess if it would meet their clients needs.

This research and its conclusions are interesting on many levels. A very serious question that arises from this is the use and effectiveness of the report writing process itself. Is report writing even worthwhile?

4.4 Case Study #2: Testing the “5 user assumption”

Many of the user tests that are facilitated by usability professionals, involve 4 – 8 users, and very occasionally up to 12. Is this number sufficient? Some user tests have only 4 users, but they have turned up enough usability issues to make undertaking the test in the first place certainly worthwhile, but is this always the case?

For some, the suggestion is that you only need 5 test participants to uncover the majority of usability issues in a web site or application (Nielsen, 1993; Virzi, 1992) but there is also equally convincing advice that challenge that assumption such as “*Why Five Users Aren’t Enough* (Woolrych & Cockton, 2001) and “*Eight is Not Enough*’ (Perfetti & Landesman, 2002).

The answer may lie somewhere in the middle, and it is worth noting that context is also very important. By context this refers to:

- 1) The level of experience the user testers have with their technology.
- 2) Their level of comfort with the web in general and their overall digital literacy – all of these things play a part and are determining factors in the quality of the outputs that you can hope to gain from user testing regardless of the numbers involved.

But let us look at the first claim that was made by Jakob Nielsen (2005) and his suggestion that you could capture up to 80% of all usability issues with just 5 users, and check if this is actually a sufficient number?

4.4.1 So when is enough, enough?

In a 2003 paper by Laura Faulkner of the University of Texas (Faulkner, L., 2003), she examines how these conclusions were arrived at in the first place. She tests the ‘*5 User Assumption*’ by conducting a series of rigorous tests of her own. In her study, 60 users were tested and random sets of 5 or more were sampled from the whole, to demonstrate the risks of using only 5 participants and the benefits of using more. Some

of the randomly selected sets of 5 participants found 99% of the problems; other sets found only 55%. With 10 users, the lowest percentage of problems revealed by any one set was increased to 80%, and with 20 users, to 95%.

Faulkner explores how the initial mantra of the 5-person user test came to be. She outlines the following nodes on the path:

The first sources:

- 1) Secondary analyses of other testers' data by Nielsen (1993) and
- 2) The “law-of-diminishing-returns” arguments made by Virzi (1992).

In general, both argued for a looser approach to user testing, which was attractive to usability professionals at the time. In that it would certainly make their lives easier in terms of data analysis, test logistics and so on.

However, the practical application of actually doing user tests on a day-to-day basis does seem to challenge the assumption that 80% of user testing issues can somehow be found by small samples of user testers. For example, in one study (Spool & Schroeder, 2001) the first 5 users revealed only 35% of the usability problems contained in a website. What is interesting, and provides some empirical evidence to challenge Nielsen's assumption both the 13th and 15th users tested revealed at least one new *severe* problem that could easily have been missed with a small sample of 5, and halting the test prematurely (Spool, J., & Schroeder, W., 2001).

In Faulkner's research she also found another study where the team tested 18 users; each new user, including those in test sessions 6–18, found “more than five new obstacles” (Perfetti & Landesman, 2002).

The idea that the five person user test was sufficient came from a combination of both Virzi and Nielsen's work, Faulkner examines the methods they used and how they came to that conclusion – and she finds the methodologies to be fundamentally flawed. Firstly, she states that Virzi's (1992) essential finding was that 5 users would uncover approximately 80% of the usability problems in a product and that even only 3 users would reveal most of the more severe problems. She states that he calculated these

various sample sizes “against the number of errors revealed by 12 users in the first study and by 20 in the second and third studies”.

While Jakob Nielsen had also been writing advocating the idea that 5 test users are sufficient in usability testing (Landauer & Nielsen, 1993; Nielsen, 1993) (Nielsen, 2000). He based his initial calculations of user error rates on data from 13 studies. Faulkner then outlines what could be a potentially fatal error in the calculation when she states that:

“In calculating the confidence intervals, he (Nielsen) uses the z distribution, which is appropriate for large sample sizes, rather than the t distribution, which is appropriate for small sample sizes. Using z inflates the power of his predictions; for instance, what he calculates as a confidence interval of $\pm 24\%$ would actually be $\pm 32\%$ (Grosvenor, 1999). Woolrych and Cockton (2001), in their detailed deconstruction of Landauer and Nielsen’s (1993) formula, confirmed the potential for over predicting the reliability of small-sample usability test results, demonstrating the inflated fixed value recommended by Landauer and Nielsen for the probability that any user will find any problem.” * (Grosvenor, L. 1999) (Woolrych, A., & Cockton, G. (2001).

*Emphasis by the author.

So this seems to be an indication that the 5-user test model is based on assumptions that have no real empirical value, and could be actually counterproductive. While both Nielsen (1993) and Virzi (1992) were upfront about the limitations of their 5-user recommendations. Virzi indicated that “[s]ubjects should be run until the number of new problems uncovered drops to an acceptable level” (p. 467), this was still not enough to halt the adoption of the idea that 5 people were enough for most user testing situations.

The idea gained traction due to a lack of a rigid method and leaving the details of when and where to apply this method up to the best judgement of the user test facilitator. On some levels, this is fine, and it could be demonstrated that 5 people will capture many usability errors on many projects, but it cannot be relied on. It could be suggested that luck will have more of a part to play in the assessment process than good science.

4.4.2 Are 5 users always enough?

So did Faulkner's trials shed any light on the Nielsen's hypothesis? Yes, she found that her observations compared favourably to Nielsen (1993) and Virzi (1992), as the average percentage of problem areas found in 100 trials of 5 users was 85%, with an *SD* of 9.3 and a 95% confidence interval of $\pm 18.5\%$. This seems to back up the hypothesis that tests with five users will find up to 80% of all usability problems. However, could this be a once-off co-incidence with more longitudinal studies needed to truly test if this is the case?

What stands against the Nielsen hypothesis is that Faulkner found the percentage of problem areas discovered by any one set of 5 users ranged from 55% to nearly 100%. This is a rather large variation, and dilutes the claim of 80% somewhat, or it may be more accurate to say that it sheds an interesting light on the claim, and stating that the 5 person user test would have a certain probability of uncovering consistently covering 60% of usability issues would be erring on the side of a more realistic hypothesis. It must also be noted that any claim of 'usability discovery' should only be applied where at least 90% of the usability issues that do exist, are known in advance.

4.4.3 More users please!

Faulkner also discovered that adding users increased the minimum percentage of problems identified. This is generally known with any statistical analysis that increasing the sample population will increase the confidence in the hypothesis and provide more statistically sound results.

She found that:

- Groups of 10 found 95% of the problems (*SD* = 3.2; 95% confidence interval = ± 6.4).
- Table 2 shows that groups of 5 found as few as 55% of the problems, whereas no group of 20 found fewer than 95%.
- Even more dramatic was the reduction in variance when users were added.

- Figure 1 illustrates the increased reliability of the results when 5, 10, and 15 users were added to the original sets of 5.

No. Users	Minimum % Found	Mean % Found	<i>SD</i>	<i>SE</i>
5	55	85.55	9.2957	.9295
10	82	94.686	3.2187	.3218
15	90	97.050	2.1207	.2121
20	95	98.4	1.6080	.1608
30	97	99.0	1.1343	.1464
40	98	99.6	0.8141	.1051
50	98	100	0	0

Figure 25: Percentage of Total known usability problems found in 100 Analysis samples

To summarize, Faulkner states that relying on any one set of 5 users was very risky that **nearly half of the identified problems could have been missed**. She also notes that adding more users greatly increases the number of issues uncovered (as shown in Figure 1).

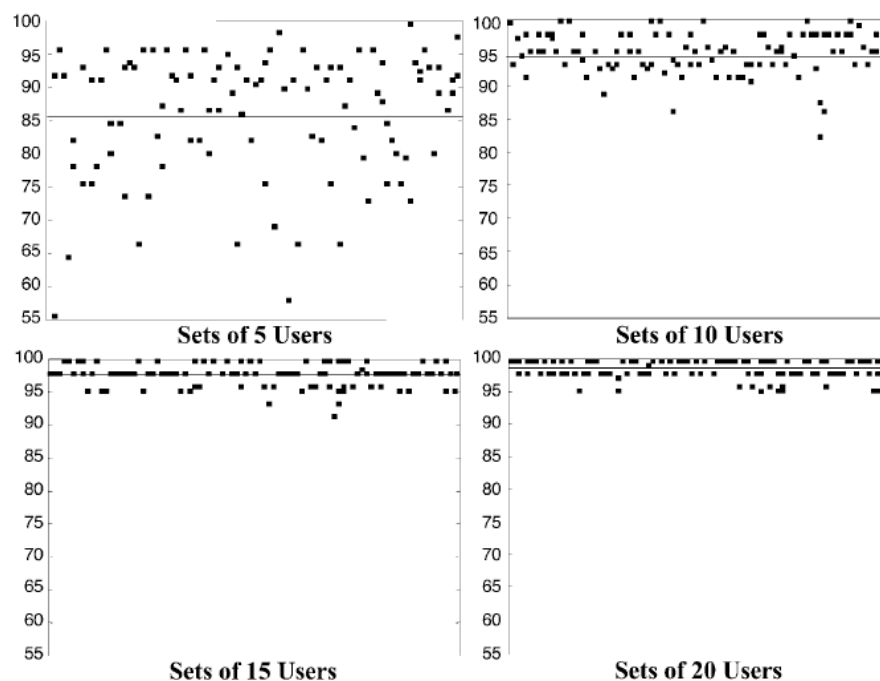


Figure 26: The effect of adding users on reducing variance in the percentage of known usability problems. Each point is a single set of randomly sampled users. The horizontal lines show the mean for each group of 100. ¹⁴

¹⁴ Charts from "Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. LAURA FAULKNER University of Texas, Austin, Texas (2003)

4.4.4 To test or not to test?

What this very interesting research shows is that there is a degree of truth in the ‘5 person user test assumption’ but only in the most narrow of senses. The 5 person user test may uncover up to 80% of issues but luck would have to play far more than a minor part for this to be the case. There is no way that you can use statistical probability with any degree of accuracy to predict usability outcomes. Probability has no memory and the discovery of some usability issues in the 5-person test method does not mean that other major usability issues will also be discovered. Many of the usability issues latent in an application could be entirely orthogonal, and the traction that the 5 user test gained within the usability community could have been actively harmful at worst, and a dangerous assumption at best.

To summarize, Faulkner does suggest that best practice would be to:

- Focus testing on users with goals and abilities representative of the expected user population.
- When fielding a product to a general population “[...] run as many users of varying experience levels and abilities as possible. Designing for a diverse user population and testing usability are complex tasks”.
- It is advisable to run the maximum number of participants that schedules, budgets, and availability allow.

It is worth noting that, in practice, most usability professionals will actually test with as many test participants that they can. There are naturally many variables that need to be considered such as the availability of testers, their suitability, time restraints and so on. Regardless of the suitability of any user testing methodology, this research aims to examine how in practice usability professionals conduct their business while practically dealing with these variables and constraints and how their user testing outputs tallies with the expected outcomes of orthodox methodologies that is in use.

4.5 Case Study #3: Evaluating the Evaluator Effect

So what about the role of the User Test facilitator? How do they impact on the testing method? Can we successfully analyse and assess the impact of the evaluator on the quality of user testing output? What about the role of bias? Can the test facilitator ever be truly dispassionate? If not, how can the impact of bias be reduced in a user test?

These issues then naturally lead to other more general questions such as, if it is ever possible to truly remove bias at all? In reality is the presence of bias even statistically significant (if such things can be truly be measured)?

Regardless of the methods undertaken to either reduce the effect of bias or remove it altogether. In the day to day reality of the usability professional - the adoption of any outputs from user testing may have more to do with the attitudes of the client and their responsiveness or willingness to change their product – than any concerns about user testing methodologies and undue test facilitator influence or evaluator bias.

4.5.1 The role of the User Test Facilitator

In this third case study the impact of the user test facilitator on the test itself in greater detail will be discussed. In usability research into the ‘Evaluator Effect’ (Jackson, John 1998) four evaluators individually analysed four videotaped usability test sessions. Their findings were that only 20% of the 93 detected problems were detected by all evaluators, and 46% were detected by only a single evaluator. From the total set of 93 problems the evaluators individually selected the ten problems they considered most severe. What is very revealing about this research is that none of the selected severe problems appeared on all four evaluators’ top - 10 lists, and 4 of the 11 problems that were considered severe by more than one evaluator were only detected by one or two evaluators. Thus, it was concluded, both detection of usability problems and selection of the most severe problems are subject to considerable individual variability.

In their research, (Jackson, John 1998) they say while Virzi et al. (1993) state that the *“...think-aloud method incorporates the users’ perspectives directly, without being filtered by an intermediary...”* (p. 312) this, while true, is rather over simplistic. Particularly in light of the ‘grey areas’ mentioned above.

Holleran (1991) observed that there can be substantial disagreement among evaluators because the collected data are primarily subjective in nature, but unfortunately he supplied no data to confirm this assertion. So Jackson and Johns research extends the study of Jacobsen et al. (1998) and they assert that the detection and severity rating of usability problems depend on the evaluators who observe and analyse the usability test sessions.

4.5.2 The Test Sessions

In the test, four experienced Apple Mac users spent about an hour working through a set of tasks in a multi-media authoring system called the *Builder* (Pane & Miller, 1993). None of the users had previous experience with the Builder, and they did not receive any instructions in the use of the system. The tests were video taped.

Four experienced HCI evaluators watch the video footage test afterwards and analysed the results. The Table 1. Below outlines their previous experience, time spent with the *Builder* application as well as time spend viewing the test footage.

Eval- uator	Occupation	Number of users previously analyzed	Initial experience with the Builder	Average analysis time per tape
E1	Associate professor	52 users	10 hours	3.8 hours*
E2	Doctoral student	4 users	5 hours	2.7 hours
E3	Assistant professor	6 users	2 hours	2.9 hours
E4	Usability lab manage	66 users	12 hours	4.5 hours

Figure 27: HCI practitioner experience, initial experience with Builder, and Time Analysing Tape¹⁵

4.5.3 Assessment Criteria

The evaluators were requested to report three properties for each detected problem:

- (a) A free-form problem description
- (b) Evidence consisting of the user's action sequence and/or verbal utterances
- (c) One of nine predefined criteria for identifying a problem.

¹⁵ THE EVALUATOR EFFECT IN USABILITY STUDIES: PROBLEM DETECTION AND SEVERITY JUDGMENTS (Jacobsen, Hertzum, John) Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting (Chicago, October 5-9, 1998),

The evaluators used the following set of problem detection criteria:

- (1) The user articulates a goal and cannot succeed in attaining it within three minutes.
- (2) The user explicitly gives up.
- (3) The user articulates a goal and has to try three or more actions to find a solution.
- (4) The user creates an item in his new document different from the corresponding item in the target document.
- (5) The user expresses surprise.
- (6) The user expresses some negative affect or says something is a problem.
- (7) The user makes a design suggestion.
- (8) The system crashes.
- (9) The evaluator generalizes a group of previously detected problems into a new problem.

Using the four evaluators' individual problem reports (276 raw problem reports), two of the authors created a master list of unique problem tokens (UPTs) that were used as a way to grade the overall severity of usability issues. To study the problem severity the evaluators received a version of the master list containing:

- (1) A short description of each UPT.
- (2) The number of users experiencing the UPT.
- (3) The number of evaluators detecting the issue.
- (4) The problem detection criteria it was attributed to.
- (5) The interface feature it involved.

Each evaluator was presented with a scenario in which a project manager had constrained the evaluators to point out the ten most severe UPTs, due to a tight deadline forcing the developer team to fix only those few UPTs in the next release of the Builder.

In this scenario, the evaluators were told that their selection of UPTs should be based on the information on the master list and on other factors, such as considerations concerning experienced versus novice users, and the Builder's use in real life settings. The UPTs on the top-10 lists were not prioritized, but each UPT was annotated with

the evaluator's reasons for including that particular UPT. This would then naturally help to create a hierarchy of issues, based on their overall severity based on the judgment of the evaluator.

4.5.4 Test Results

The percentages of the total of 93 UPTs reported by E1, E2, E3, and E4 were 63%, 39%, 52%, and 54% respectively. Thus, a single evaluator detected on average 52% of all known UPTs in the Builder interface. What is significant is that the net effect of adding more evaluators to a usability test **resembles the effect of adding more users**; both additions increase the overall number of UPTs found.

Would this be an excellent way of improving the quality of user test evaluation by simply getting more UT pros to review video footage, and other data? This may be easier to implement that adding an order of magnitude more users.

The following figure depicts the number of the 93 UPTs detected as a function of the number of both evaluators and users.

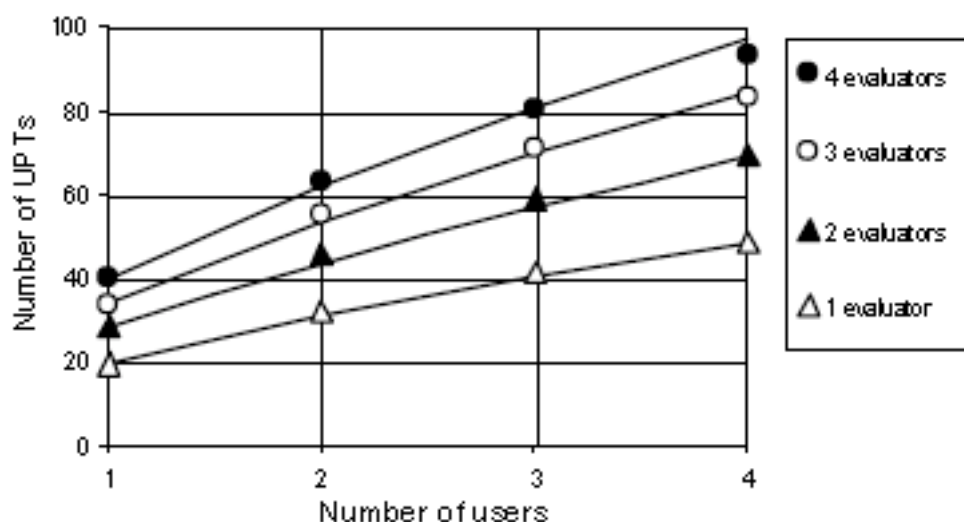


Figure 28: The number of detected UPTs depends on the number of users and the number of evaluators¹⁶

¹⁶ THE EVALUATOR EFFECT IN USABILITY STUDIES: PROBLEM DETECTION AND SEVERITY JUDGMENTS (Jacobsen, Hertzum, John) Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting (Chicago, October 5-9, 1998),

Calculating the effect of running more users, (Jackson, John 1998) found:

- 1) An increase of 55% going from one to two users
- 2) 26% going from two to three users,
- 3) And 23% going from three to four users

This was when all evaluators were included in the calculation (the topmost curve). The declining number of new UPTs detected as more users are added confirms the results from similar studies (Lewis, 1994; Nielsen & Landauer, 1993; Virzi, 1992).

The can be described with Equation 1. The fit between the equation (the curves in Figure 1 above) and the data (the data points in Figure 1) is highly significant (squared correlation coefficient (R^2) = 0.997; standard error of estimate = 2.6%; $p < 0.001$).

The following equation describes the relationship between the number of detected UPTs, number of users, and number of evaluators of the study.

$$\text{No. of UPTs} = 19.35 * (\text{no. of evaluators})^{0.505} * (\text{no. of users})^{0.661}$$

The evaluator effect for all UPTs is substantial; as much as 46% of the UPTs were found by only a single evaluator, while 20% were found by all four evaluators.

Problem criteria 9 (a problem identified as a generalisation of previously detected problems) might be more likely to differ across evaluators, since the generalization process is quite subjective.

However, only 5% of all problem reports were attributed to criteria 9. Hence the evaluator effect cannot be caused by these criteria alone.

4.5.5 Level of Agreement

In order to investigate whether the level of agreement among the evaluators differs when detecting more severe problems, three methods were used to extract severe problems.

- 1) The UPTs attributed, by any evaluator to some of the more severe usability criteria were extracted. This amounted to 37.
- 2) They then looked at the 25 UPTs that appeared on at least one evaluator's top-10 list.
- 3) Finally the 11 UPTs that were included on more than one top-10 list were extracted.

The following Table 2 shows that the evaluator effect in detecting problems was progressively less extreme for the sets of more severe problems, but for the smallest set of severe problems it was still substantial.

UPTs	No. of UPTs	Detected by			
		only 1	any 2	any 3	all 4
All UPTs	93	46%	20%	13%	20%
Violating criteria 1, 2, or 8	37	22%	19%	19%	41%
Any UPTs on top-10 lists	25	20%	20%	80%	52%
More than one top-10 list	11	9%	27%	0%	64%

*Figure 29: Percentages of the UPTs detected by only 1, any 2, any 3 and all 4 evaluators*¹⁷

4.5.6 Results

What is very interesting and significant about this research is that:

- 1) Severity judgment differed substantially between the four evaluators.
- 2) When comparing the top -10 lists, there were large differences; 56% of the 25 UPTs that appeared on the four top - 10 lists were selected by only a single evaluator, 28% were selected by two evaluators, and 16% were selected by three evaluators.
- 3) Therefore not a single UPT appeared on all four top-10 lists!

¹⁷ THE EVALUATOR EFFECT IN USABILITY STUDIES: PROBLEM DETECTION AND SEVERITY JUDGMENTS (Jacobsen, Hertzum, John) Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting (Chicago, October 5-9, 1998),

4.5.7 When is a usability problem not a problem?

This research is very interesting as it highlights the subjective issue of when usability issues are critical or not and how this can be judged. The research seems to indicate that this is also a highly subjective issue. Are the evaluators biased toward the problems they originally detected themselves?

Jeffries (1994) found that problem reports are often unclear and ambiguous, and may therefore be unreliable. So is there a risk of biased severity judgments against that of misinterpreted problem reports? The substantial differences among the evaluators in terms of their selection of problems for their top-10 lists reveal that judgments of severity are highly personal. This is asserted due to the fact that *no* UPT appear on *all* evaluators' top-10 lists.

4.5.8 Retrospective reporting

The evaluators were then asked to write a report outlining their findings. These were assessed and the researchers concluded that:

- 1) The evaluators' methods for extracting top-10 UPTs varied greatly.
- 2) The selection methods were based on multiple aspects such as:
- 3) The evaluators' favour for certain user groups.
- 4) The number of evaluators and users encountering a problem.
- 5) The violated problem criteria (as set out initially).
- 6) Expectations about real-world usage of the application.

All these aspects may catch important dimensions of problem severity but they also point out that severity is an ill-defined concept.

4.5.9 Analyzing and Communicating Usability Data

So one of the burning questions is, “How can usability practitioners effectively communicate the results of usability analysis such as user testing”? Later on in our knowledge audit we will look at how usability professionals currently tackle this issue.

The audience for these results are often non-technical people and some of the methods, due to constraints such as time and resources are often non-scientific. These more limited usability tests can be referred to as 'Discount techniques' (Nandini P. Nayak, Debbie Mrazek & David R. Smith, 1994) and are easy to dismiss as lacking sufficient rigour or validity.

Discount methods have been developed in response to time pressures of modern business and the desire to get as effective results as possible with a limited budget.

Some advantages of discount techniques are low-cost materials (typically pen and paper, post-its etc.) used for cognitive walkthroughs, prototyping etc, the rapid speed of data collection (often less than 1 day), and the small number of users required.

So it is a common dilemma as to how to effectively communicate the outputs of usability data such as user testing -when done in this way- across. As (Nandini P. Nayak, Debbie Mrazek & David R. Smith, 1994) indicate, *"without a clear analysis and communication strategy, the data from discount techniques are much easier to dismiss as unreliable or inadequate to inform design decisions"*.

That is not to say that 'Discount techniques' are not valuable. Discount techniques are a response to lack of time and resources and while they are a quick and dirty usability fix the down side is that it can be difficult to generalize into design recommendations (Nandini P. Nayak, Debbie Mrazek & David R. Smith, 1994). This can be the same for more involved methods such as ethnographic techniques or other experimental laboratory techniques.

4.6 Case Study #4: Effectively communicating the results of Usability Data

Nandini et al., 1994 asked how best to approach this issue: They examined several key areas and asked the following questions about the best approach to communicating the result of Usability studies in general.

They sought to find out if the best way was:

- Through the conversion of usability evaluation findings into "defects"?
- Through some form of protocol analysis?
- Through videotape summaries of prioritized problems?
- Through special report writing or presentation techniques?
- Through having control of the external reference specification?
- Through impact analysis or importance analysis?

Overall, 15 participants (from industry or academia) gave feedback in the form of position papers or interviews. The results are very interesting. To the question, “Why is Usability Data difficult to analyze or communicate?” the consensus was that:

1) Time constraints and limited resources play a major part.

2) Usability data is based on observation; and there is a lack of sufficient metrics as observation is subjective.

3) Use of small samples in usability data make it difficult to produce statistically significant results to back up usability professional’s assertions about the quality of the interface design.

4) Large usability reports are often never read. So how can test data be communicated in a way that is effective and yet understandable and accessible to non-technical people? The audience for usability data is often mixed (technical, and non-technical) so there is difficulty in fitting the content to the audience in a way that suits all their needs, levels of detail, overview and granularity.

5) How can observer bias be reduced when collecting data?

6) Are we testing in time to respond to the outputs of the usability data and incorporate the results into the project?

Some of the findings and suggested ways to reduce some of the above problems were:

- 1) Educating design teams about the usability engineering process.
- 2) Engaging the teams involved at the beginning of the product development process by introducing similar case studies as examples of the process and what they can expect.
- 3) Start usability data collection early on, so improvements can be made on time.
- 4) Use understood formats for the target audience to communicate effectively
- 5) Provide outputs from tests in logical way, so the audience can understand how a conclusion was reached.
- 6) Use a positive tone in presenting findings, so usability feedback is not seen as overly negative.

4.6.1 ‘Think aloud’ Studies

The majority of the above user test studies are ‘think aloud’ studies – a practice that has long been incorporated into current user testing methods. This is where the user is encouraged to give feedback on what they are doing, how they feel, what improvements they can suggest etc. This is very effective form of ‘real time’ feedback that is invaluable to the test facilitator and the project in general. While care certainly has to be taken on the part of the test facilitator as to how this feedback is interpreted, or these results incorporated into the test outputs, it is nonetheless invaluable. (J. Nielsen, T.Clemmensen, C.Yssing, 2002)

4.7 Conclusions

The above case studies have looked at issues like, the strength and weaknesses of various approaches to user testing, ‘The Evaluator Effect’, the issue of insufficient sample sizes, and how to effectively communicate the results of user testing.

Some interesting findings were that:

- 1) A usability report that is ignored by the developers is useless, in fact the issue of exactly how to effectively communicate user testing output is still open to question.
- 2) There is often little overlap between issues covered by several evaluators looking at a single application.
- 3) Relying on any one set of 5 users was very risky that nearly half of the identified problems could have been missed.
- 4) Nielsen used the z distribution, which is only appropriate for large sample sizes, rather than the t distribution to produce his eye-catching mantra that gained huge popularity in the early days of usability testing for the web. However it is completely misleading.
- 5) Even with the best expert evaluation, with many evaluators is not guarantee of successfully covering all usability issues.

The next section is the ‘Knowledge Audit’ and where we aim to get a glimpse of current practice among usability professionals to see if we can get closer to the current state of the art.

5 SECTION 2: KNOWLEDGE AUDIT BACKGROUND. EXPERIMENTATION & EVALUATION

5.1 Evaluating Usability Testing

Traditional usability testing involves testing with a random sample of the public or a sample of representative users, who will in practice be using a web application or website, in an attempt to assess the quality of the user experience. The outcomes of the test such as whether a user could successfully complete a certain task or set of tasks, the ease of use with which they could complete tasks - and other user feedback and observations made during the test - are all noted by the test facilitator.

This recorded information is therefore very valuable as it allows an experienced usability analyst to gain a detailed picture of what is working for the end user, or not, in a particular user interface design.

While traditional usability testing is very useful, it usually only captures a small sample of issues. It is not exhaustive but any difficulties will become immediately apparent during the test. An experienced usability professional will understand exactly how the design or implementation contributes to these problems and what can be done to fix them.

5.1.1 User Testing with People with Disabilities

Involving people with disabilities in user testing is often the best way to gain a detailed picture of how usable an interface is for this user group. By closely studying the experience of a user with a disability it is possible to gain insight into how design choices and decisions impact on the user experience for other users with similar disabilities.

While user testing with people who do not have disabilities can yield many positive results that can improve the user experience, these are generally users who have more 'standard' user requirements and may not need or use assistive technology.

By successfully involving people with disabilities in the development and design cycle of a project, through the feedback obtained from user testing with them we can gain a more rounded picture of the user experience for a very broad range of people and build applications that suit a very broad range of user needs.

5.1.2 Formal vs. Informal User testing

Formal usability analysis and user testing conjures images of a stern scientist with a white coat taking notes behind a one way glass observation screen while the test participant is relayed commands and tasks via a talkback system or feedback relay. These instructions of course must be given in a voice drained of any hint of emotion or semblance of humanity in order to avoid the sin of bias within the test.

This image, while an obvious caricature, is what many people may think when they conjure up images of observation, testing, and analysis. It is however a rather outdated view that may be at odds with the current trends and habits in user testing - which we will explore later in our research into current user testing practices.

Formal user testing is very much associated with the ‘scientific method’ and while it is certainly valid and useful – in certain domains – it is not what we are concerned with here. The formal method is associated with statistical analysis, and control experiments. What we are concerned with is the more ‘real world’ approach of informal user testing. This is where testing has often to be done quickly, as a part of an iterative development cycle (in the best of cases) and as an add-on at the end of a project as some attempt at validation, at worst. (Hill. L, Carver. L et al. 2000)

More informal user testing is where there is a test script and a series of tasks that have been outlined beforehand. The test facilitator may also have a relationship that has been built up over the years where the test facilitator and participant have done many, many tests together.

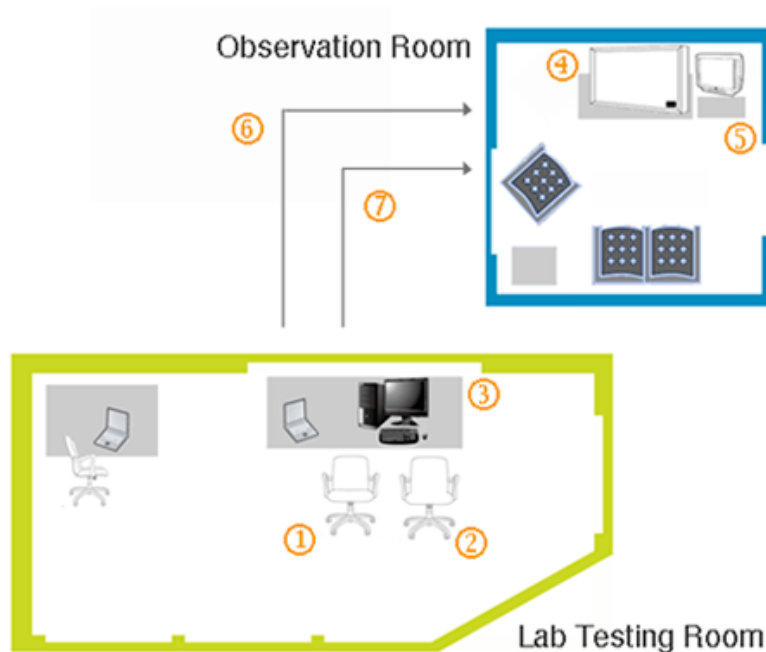
5.1.3 Measuring User Testing outputs

User Tests have certain outputs. These are varied and can be the accumulated notes of the test facilitator, the video footage collected during the test for later analysis, the collective impression of uninvolved observers of the test and so on. Some outputs are more tangible - like video that can be archived and viewed later. Some are less tangible but are still very valuable such as the lasting impressions a user test can leave on the observers when they have watched someone use their website.

These less tangible impressions and subjective experiences can result in very real outputs. A product can be dropped, a software iteration cycle abandoned and so on, if a project manager sees a 'live', 'real-time' negative user response to one of their interfaces. Conversely, a designer can be vindicated as the results of their design efforts and attention to detail come to fruition when a user says in all sincerity "Yes, that website is great, I can find the information I need really easily. I like the way the page is designed". This experience can be more profound when the person being observed has a disability.

It is not the average user experience where one may get the interesting information as a usability analyst; it is the extremes, the edge cases, both positive and negative where the really interesting aspects of user testing analysis take place. This is where both positive and negative experiences are amplified and made quite explicit - so there is often no ambiguity. The language is often less than neutral so there can be little doubt of the users impression and feelings about a particular user interface or application.

5.1.4 How does User Testing work?



The above is an example of user testing facility showing layout of rooms and equipment and the following description of user testing comes from the CFIT website. (CFIT website, 2010)

5.1.5 The User Environment

The user test participant (1) sits in a typical office environment within the testing room that is controlled for sound. The test facilitator (2) sits with the user, explaining the tasks, taking notes and critically observing the user's interactions. The test is conducted using a standard PC (3) with assistive hardware and software. Dedicated user test recording software such as Morae, together with discreet cameras and microphones capture and record every aspect of the user testing session for later analysis.

5.1.6 The Observation environment

Observers can watch the test in real time from the comfort of our observation room couches. The video from the user's monitor (6) is displayed on a flat screen TV (4) while a second signal from the camera and microphone (7) shows the user's gestures, facial expressions, body language and vocalisations on a television monitor (5).

Through these links, observers can see everything that the user does and says, as well as the interaction between the user and the facilitator.

5.1.7 Test details

A typical user test consists of 8 separate user sessions of 1 to 1½ hours each. The types of user cover a broad range of disabilities and assistive technologies. It also allows us to include younger and older users and people with different levels of experience. This results in a more representative sample of attitudes and approaches.

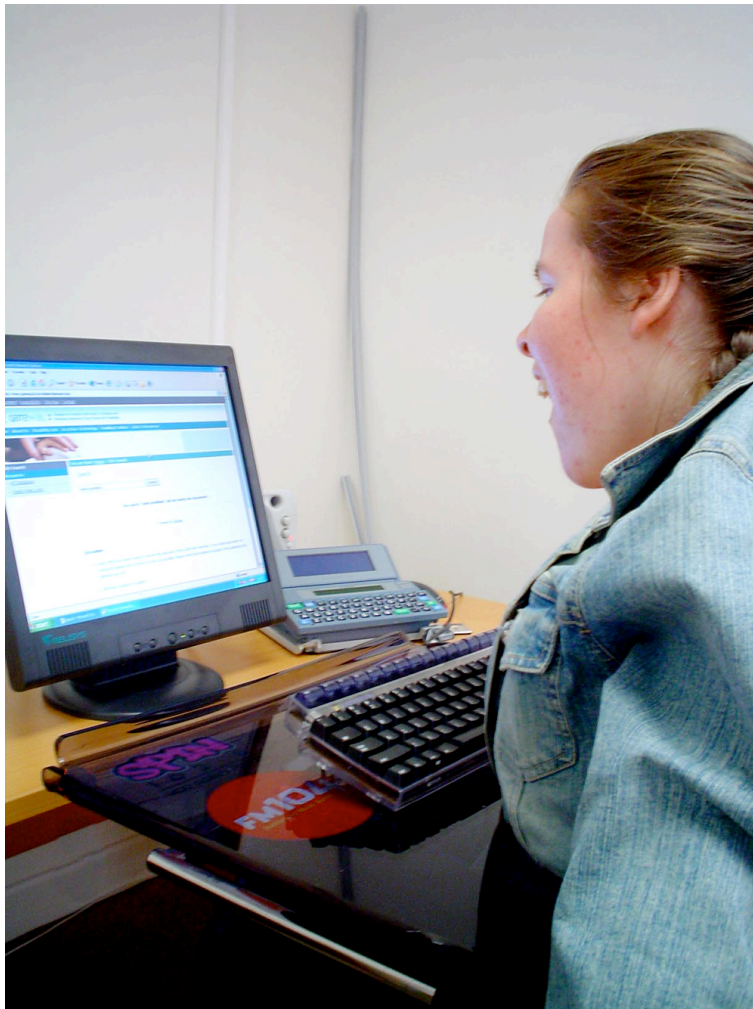


Figure 30: A User Test participant in the NCBI Centre for Inclusive Technology usability lab

Each user carries out a set of realistic tasks that have been agreed beforehand with the client.

These will usually include the most common tasks for which the product is used, as well as the most critical tasks and any tasks that test facilitator may anticipate causing problems for users. Tests are carefully designed and run so as to yield the most realistic user behaviour and therefore the most valid results.

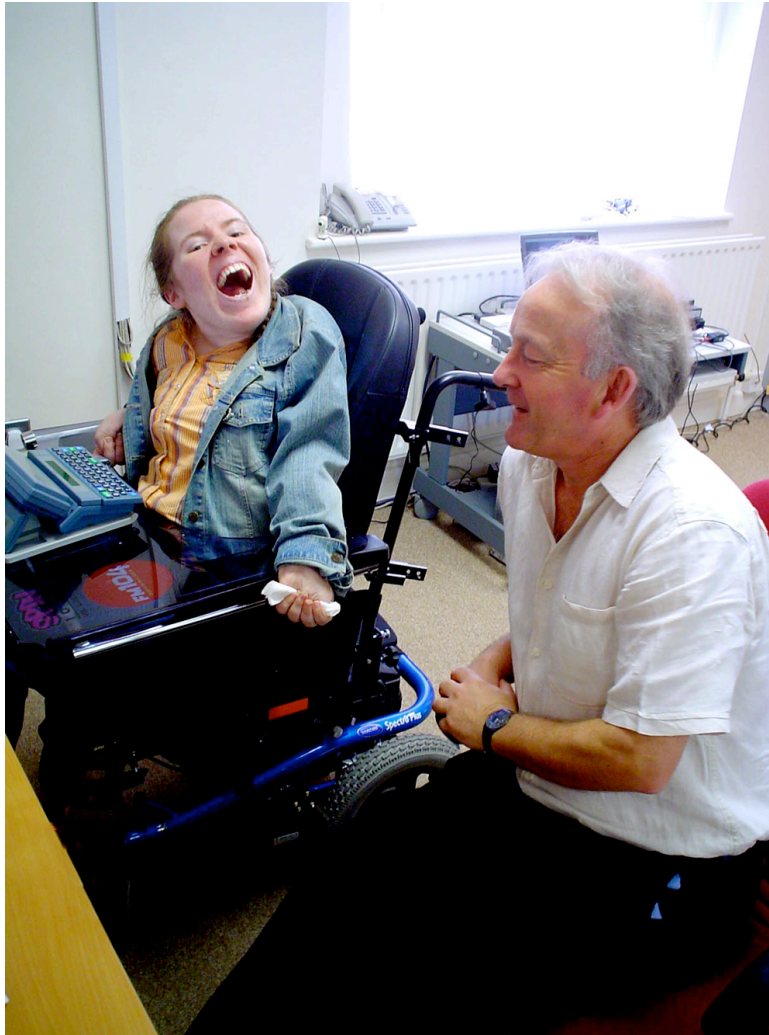


Figure 31: User Test facilitator in the NCBI Centre for Inclusive Technology lab

5.1.8 Observing a user test

Observing user tests is one of the best ways to gain a first hand understanding of what accessibility and usability really mean. Designers and developers in particular can get huge benefits from the insights they gain from observing users.

Some facilities have a dedicated observation room. Using a wide screen TV and a small video monitor, clients can watch and listen to the user tests via a remote link without disturbing the users in their tasks.



Figure 32: Observing a user test in the NCBI Centre for Inclusive Technology Observation room

Digital video recordings of user sessions are then often used to illustrate key issues. The emphasis is on building an understanding of how the design of the site contributes to users' difficulties and what practical steps can be taken to alleviate these problems.

While the main thrust of this research is to look at user testing, we need to first of all put it into context of the main issues that face a designer when trying to build applications and websites that can be used by the widest possible audience – and the corresponding evaluation methods.

5.1.9 Goals of User Testing

Rubin states that the overall goal of usability testing is to “identify and rectify usability deficiencies [...] prior to release”. This is in order to ensure that products and services are:

- 1) Easy to learn and to use.
- 2) Satisfying to use.
- 3) Provide both a utility and functionality that are highly valued.

The more specific goals and benefits of Usability testing are:

- 1) To provide a historical record of usability benchmarks for future releases.
- 2) Minimize the cost of service and support.
- 3) Increase sales of a product or service.
- 4) Acquire a competitive edge.
- 5) Minimize risk of releasing a sub-standard product.

It is interesting that there has been a large increase in demand for the services of usability professionals over the last few years, as the Internet has become more pervasive and the consumer has far more choice. Usability comes into its own when the quality of the user experience is a determinant of what they buy and the services that they use.

5.1.10 Limitations of Testing

Some of the main drawbacks of testing, and some of the reasons that unfortunately mean that Usability testing is not a silver bullet is:

- 1) **Testing is always an artificial solution.** Testing in the lab, or in the field is a depiction of usage and not the real situation itself.
- 2) **Test results do not prove that a product works.** Test results that are statistically significant are actually not indicators that a product will work, but rather that the results of the test were not arrived at by chance.
- 3) **Participants are rarely fully representative of the target population.**
- 4) **Testing may not always be the best technique to use.** Expert evaluation for example, may be a better technique in some cases.

Rubens last point is particularly interesting and is very pertinent where domain expertise is needed to successfully use a software application etc in the first place.

Having stated some of the limitations of user testing Rubin then goes on to say that user testing is still the most “*infallible indicator of potential problems and the means to resolve them*”.

5.2 Experimentation

In order to make sense of the contents of the Knowledge audit, it is important to first understand some of the mechanics of current User Testing methodologies.

5.2.1 Basics of a Testing Methodology

The origins of the basic methodology for formal user testing are in the classic approach to undertaking a scientific controlled experiment. This involves the setting up of a test question or hypothesis and then, under controlled circumstances changing some of the variables and observing the outcome.

Cause and effect relationships are thus examined and by using a relevant statistical technique the hypothesis is either confirmed or rejected. Formal testing therefore requires that:

- 1) **A hypothesis is formulated:** This is ideally as specific as possible and must be clear outlining what the scope of the test aims to prove (and also inferring what it does not aim to prove).
- 2) **Randomly chosen participants are assigned to work under experimental conditions:** A representative sample if chosen from the target population.
- 3) **Tight controls are employed:** This is vital for the integrity of the test data and corresponding conclusions.
- 4) **Control groups must be employed:** To validate the results, a group that must form the basis of comparison must be used. The treatment of the group must vary only on the single variable being tested at any one time.
- 5) **The sample of users must be sufficiently large to measure statistically significant differences between groups:** Too small sample sizes can very easily lead to poor or misleading data. (Rubin, 1994)

Rubin then goes on to note that this method may not be suitable to use as a methodology to conduct usability tests in today's fast paced, pressurized development environment. Also usability testing is not in practice used to formulate and test certain hypothesis but to improve the usability of products and services.

5.2.2 Basic Elements of User Testing

Rubin then outlines a methodology for more informal testing:

- 1) Development of problem statements or test objectives rather than hypotheses.
- 2) Use of a representative sample of end users, which may or may not be randomly chosen.
- 3) Use a representation of the work environment.
- 4) Observation of end users (with a representative product). Controlled interrogation and probing of the participants by test monitor (facilitator).
- 5) Collection of both quantitative and qualitative performance and preference measures.
- 6) Recommendations of improvements to the design of the product.

Rubin then outlines the four main types of informal testing. He maps each of them to the different stages of the product development cycle in an attempt to more clearly outline what kind of testing would be beneficial at various stages in the design process. He also notes how confusing the varying terminology can be in describing very similar testing techniques – so to simplify matters he uses the product development cycle as a reference point to describe the various tests.

He states that these tests are also most effective when used as a part of an iterative design process – which is a cycle of design, test, measure and redesign. (Gould and Lewis, 1995) and that this process must be iterative as it allows steady progress, built upon empirical evidence to shape the design. They are:

- 1) Exploratory Test
- 2) Assessment Test

3) Validation Test

4) Comparison Test

The first three are used at equivalent stages, only the final method – the Comparison test can be used at any stage as a part of the other test and is not explicitly associated with a life cycle phase. This is illustrated in the following diagram.

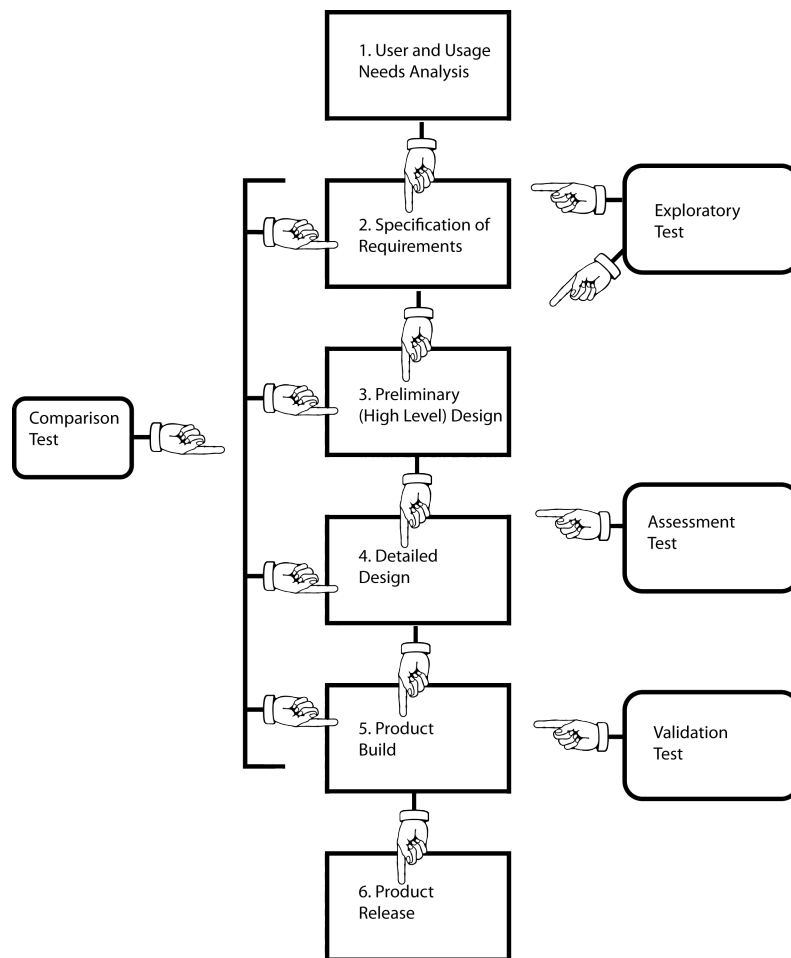


Figure 33: The Product Development Life Cycle (based on Robins Iterative Model, 1994)

5.3 *The Exploratory Test*

5.3.1 When:

The Exploratory test takes place early in the development cycle. This is when the product is being defined and designed. The user profile and usage models should already have been defined at this stage also.

5.3.2 Objective:

This is when the effectiveness of the design concept is explored. At this stage a rather high-level approach where the users mental model of the product is assessed – is taken.

For example, if the product is a user guide or interface for a software application the kind of high-level issues being assessed may be:

- Overall organisation of subject matter
- Whether to use a graphical or verbal approach
- Accessibility of the proposed format
- Anticipated points of help access
- How to address reference information

At this stage the designers will wish to evaluate any assumptions they may have made about the users themselves. So some user orientated questions that would be asked at this stage may include:

- What does the user think about using the product?
- Does the basis functionality appeal or have value to the user?
- Is the User Interface intuitive?
- Are former CLI (Command Line Interface) users able to now use the GUI?
- Is there any a priori knowledge needed to use the product?

- What functions of the product need the manual, and what are “*walk up and use*”?
- How should the table of contents be organised?

Rubin then emphasises the importance of this kind of early analysis and research as at this stage important design decisions are set in motion and the project, in order to be successful much not be based on incorrect assumptions on the part of the designers. So this stage very much determines the underlying structure and that determines the quality of all that follows. (Rubin, 1994)

5.3.3 Overview of the Methodology

An exploratory test usually means that there is extensive interaction between the test facilitator and the test participant in order to work out how effective the basic design concepts are. This can involve various prototypes being assessed by the participant. These can include static on screen mock-ups of an interface or design, more advanced functional prototypes, widgets or site sections based on certain functionality that the designers are interested in assessing or even simple paper prototypes that model the suggested design.

Working prototypes can be both ‘horizontal’ and/or ‘vertical’. This is to say that the former represents a shallower working model that offers the user limited functionality and the latter allows the user to delve more deeply into the application to test specific aspects of its design.

During the test of the prototype, the user would often attempt to perform a series of representative tasks or if this is not possible at the early stages of development, a ‘walkthrough’ under the guidance of the test facilitator may be used. This is where the product is reviewed and questions asked by the test facilitator to try and understand the users preferences. Both methods may be used depending on the status of the product/interface to be tested.

This stage in the testing process is quite informal and is really collaboration in order to explore possibilities for the interface.

The main point here is to try to explore the users' thought processes by encouraging the user to 'think aloud' as much as they can. In later tests there may be much less interaction between test facilitator and participant.

In later tests, quantitative data is often collected when assessing how well the user performs when using an interface whereas at this stage, the user is encouraged to air their views and ideas about how to improve aspects of the design, or clear up confusion about aspects of the design the developers are unsure about so the emphasis is on qualitative exploration with an emphasis on discussion, an examination of high-level concepts and thought processes. (Rubin, 1994)

5.4 *Assessment Test*

5.4.1 When:

This is one of the most typical types of usability test. It is one of the simplest to design and conduct and is usually undertaken early or midway in to the development cycle.

5.4.2 Objective:

The purpose of the Assessment test is to expand on the findings of the exploratory test. It aims to examine how effectively the initial concept has been executed. Rather than assessing how intuitive the product is likely to be this method aims to look at the detail of how well a user can perform tasks and attempts to uncover any inherent usability issues.

5.4.3 Overview of the Methodology

This is an information-gathering test and the methodology is a cross between the more informal exploration test and the more tightly controlled validation test. In the Assessment test however the user will:

- 1) Perform tasks rather than simply walking through and/or commenting on screens and pages.
- 2) The test monitor will lessen their interaction with the participant, as there is more emphasis on the users behaviour rather than their thought processes.

- 3) Quantitative measures will be collected. (Rubin, 1994)

5.5 *Validation Test*

5.5.1 When:

This is also referred to as the verification test and is conducted late in the product development cycle (often close to release) as it is intended to certify the products usability.

5.5.2 Objective:

The objective of the Validation test is to measure how the product performs against a recognised standard of evaluation. This could be usability standard or a performance related standard set up by the company themselves or even against a competitors product.

These tests aim to see if the product meets these standards prior to release and if not, aims to find out why. The Usability objectives are often stated in terms of performance criteria, such as speed and accuracy and how well a user can perform tasks etc. Objectives can also be stated in terms of user preferences and ratings.

Another major aspect of the validation test is to see how well the various components of a product work together. In larger companies this can be particularly interesting as it often proves how well various departments are or are not communicating during a project. Certain components could be created in relative isolation from each other so the integrated validation testing stage is an important acid test.

5.5.3 Overview of the Methodology

The validation test has more emphasis on experimental rigour and consistency is conducted in a similar way to the Assessment test with three exceptions:

- 1) Before the test, standards are identified for the tasks to be measured against.

- 2) Participants are given tasks to perform with very little interaction from the test facilitator.
- 3) The collection of quantitative data is the main focus.

Since this kind of testing invariable means that a standard will be used as a benchmark, it is also vital to ensure that there is agreement on how adherence to the standard will be measured and also that there will be suitable steps taken when the product does not meet the standard. (Rubin, 1994)

5.6 Comparison Test

5.6.1 When:

This testing method is different from the others in that it is not linked solely with any phase of the development cycle. In early stages it can be used to compare different designs and interfaces using the exploratory test, in the middle stages it can be used measure the effectiveness of different elements in the design and at the end of the cycle, it can be used to measure against similar finished products that are in the wild.

5.6.2 Objective:

It can be used in conjunction with any of the other tests to compare two or more designs, interface styles or indeed any other component that is relevant in order to see what is good or bad about its usability and whether it should make the final iteration of the products design.

5.6.3 Overview of the Methodology

This methodology involves comparison of two or more designs, or components of a design. Performance data etc can also be collected and compared in order to provide empirical evidence to back up a design decision. Comparison tests can be conducted informally, as a part of the exploratory test, as a part of a more rigorous controlled experiment etc. It is very flexible depending on the context of the need. (Rubin, 1994) Rubin does suggest however that, the best results are from comparing designs that vary greatly and are not similar. He states that this may be because:

- 1) The design team is forced to stretch its conceptions of what will work rather than just continuing along with a predictable pattern.
- 2) During the test, the participant is forced to really consider why one design is better than another and focus on the aspects that make it so.

The classic approach to undertaking a scientific controlled experiment may not be suitable for the more informal world of commercial user testing, but are there aspects of it that we can gain from it? This research aims to find out how practitioners undertake user testing, identify any flaws, find out something about what their level of knowledge regarding existing methodologies is, and finally if they are put into practice. The objective of the experiment in this research is exploratory. By capturing a snapshot of the current state of practice, via a social anthropology study, of experts we can assess what they know and identify issues and trends in the field.

The 'Knowledge Audit' form of this research has learnt itself well to giving a rich qualitative overview. Some more qualitative statistical methods of research would not have been entirely suitable, as they may not effectively reflect what is a very nuanced area of work. (Yeung, 1995)

5.7 Evaluation

The Knowledge Audit survey aims to provide some rich qualitative data that will give some insight into what is a rather nuanced practice. From the case studies, it is clear that there have been some assumptions that have turned out to be incorrect. The four case studies are particularly interesting as they give a good overview of some of the major issues such as:

- What is the correct amount of people to test with?
- Can expert evaluation be relied on considering the great of variability in their results?
- What are the best ways to communicate the results of user tests to clients?

Later the results of the Knowledge audit will be framed against these questions.

5.8 Conclusion

In this chapter we see the many excellent methodologies developed by Rubin that are the basis for many of the current approaches to User Testing. This chapter also aimed to critically evaluate usability testing, and outlined research into comparing, and evaluating, user evaluation methods. Thus, as Lund has suggested the lack of standard metrics means that it is impossible to develop a definitive User evaluation methodology. The weaknesses in the '5 User Assumption' were also investigated, which may have had an adverse effect on user testing as it was taken as canonical.

Further research by Faulkner did discover that adding more users increased the number of errors discovered, and relying on only 5 users could result in only half of the important UI problems being identified. The role of the UI facilitator was also investigated and the author suggests that the client's responsiveness and the relation built may be most important. Finally the evaluator effect was looked at and discovered that opinions can vary wildly between experts, particularly when critical subjective assessment by an evaluator is intended to improve a user interface.

6 SECTION 3: KNOWLEDGE AUDIT RESULTS

6.1 Methodology

The following research was undertaken in the form of a Knowledge Audit. This is a qualitative method where respondents answer survey questions, mostly with descriptive prose. The survey was distributed via email in MS Word format to the recipients who agreed to take part in advance. There were 14 surveys sent out and 10 came back fully completed.

6.2 TYPE OF WORK

6.2.1 Please describe your role?

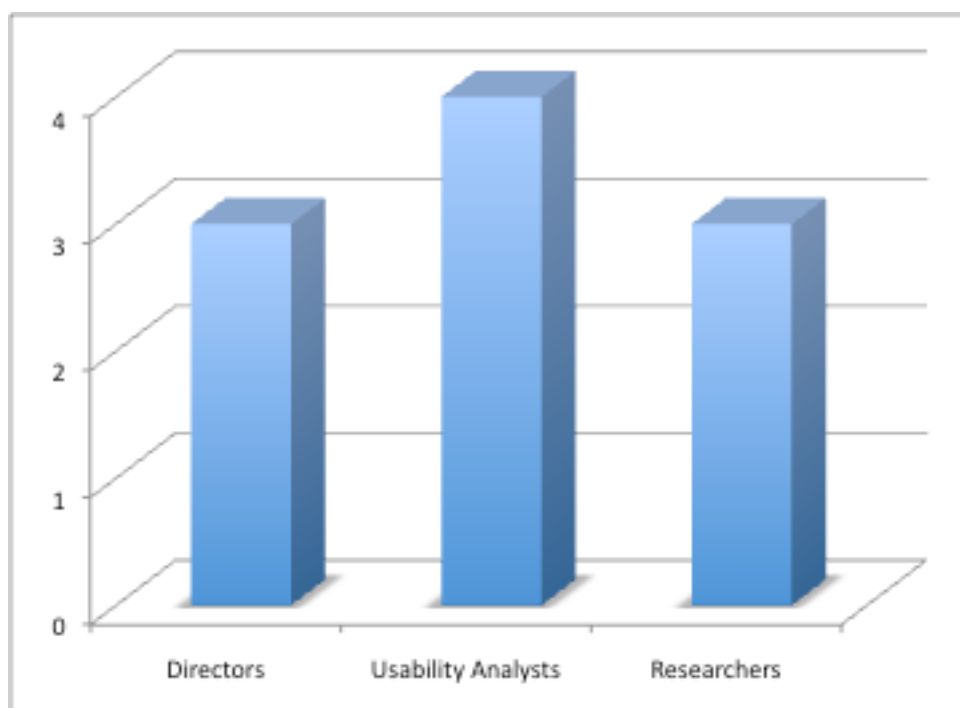


Figure 34 Role description

6.2.2 How would you describe the work that you do?

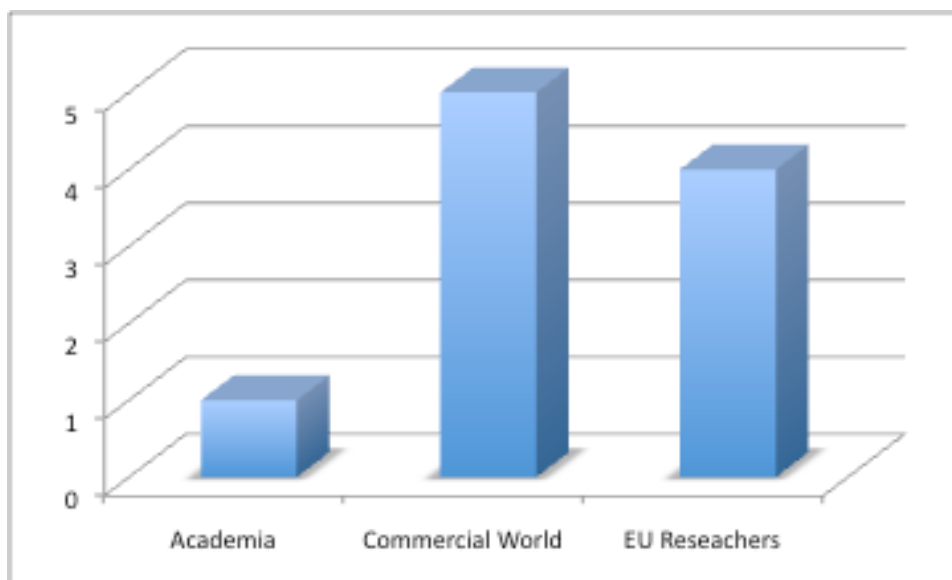


Figure 35 Groups of role types #1

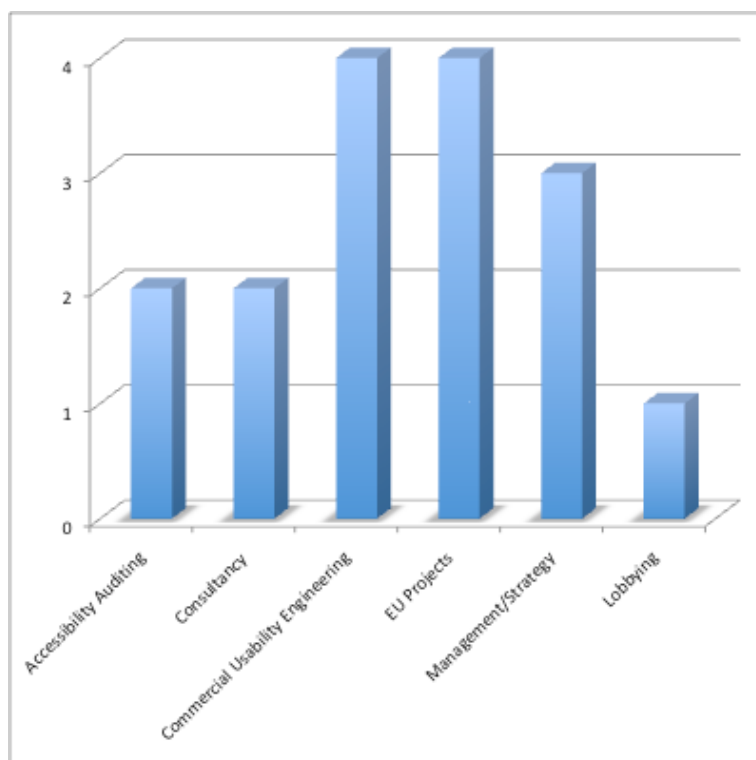


Figure 36 Groups of role types #2

6.2.3 EDUCATIONAL/BACKGROUND

The Knowledge Audit participants had the following educational backgrounds:

3 x HCI,

3 x Psychology (1 x Msc Cognitive Science)

1 x Msc Computer Science

1 x Business Systems

- 1 x Graduate Diploma in IT
- 1 x Industrial Design
- 1 x Special education (Disability)

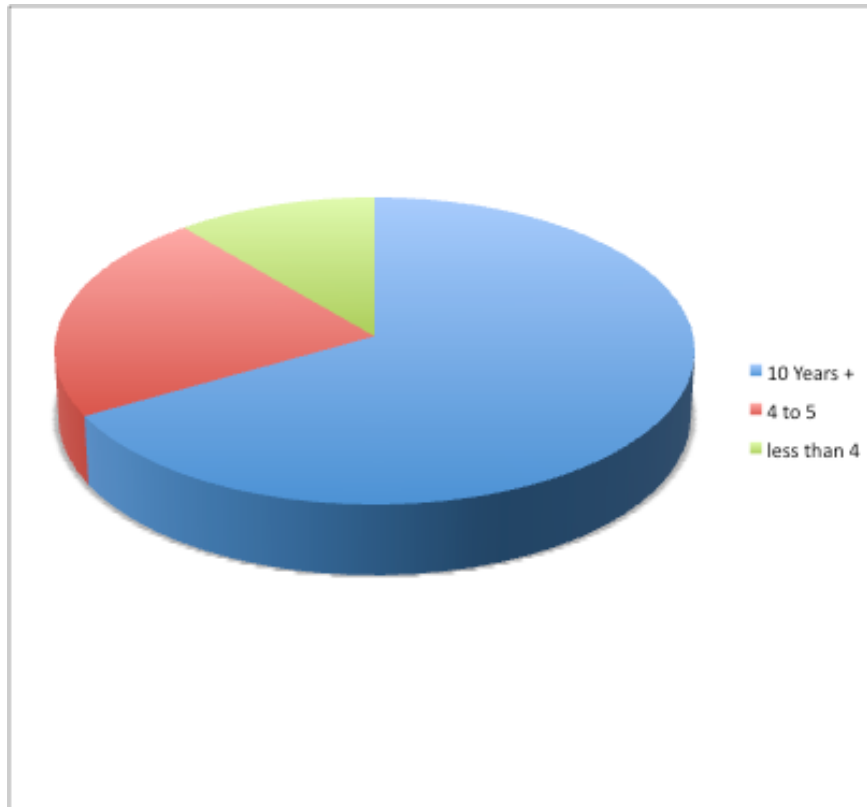


Figure 36 Level of professional experience

In terms of professional experience, the respondents had the following levels of experience in their field:

6 x have 10 years + experience.

5 x said that their experience of people with disabilities came mostly from User Testing.

2 x have 4/5 years experience.

1 x have less than 4 years experience.

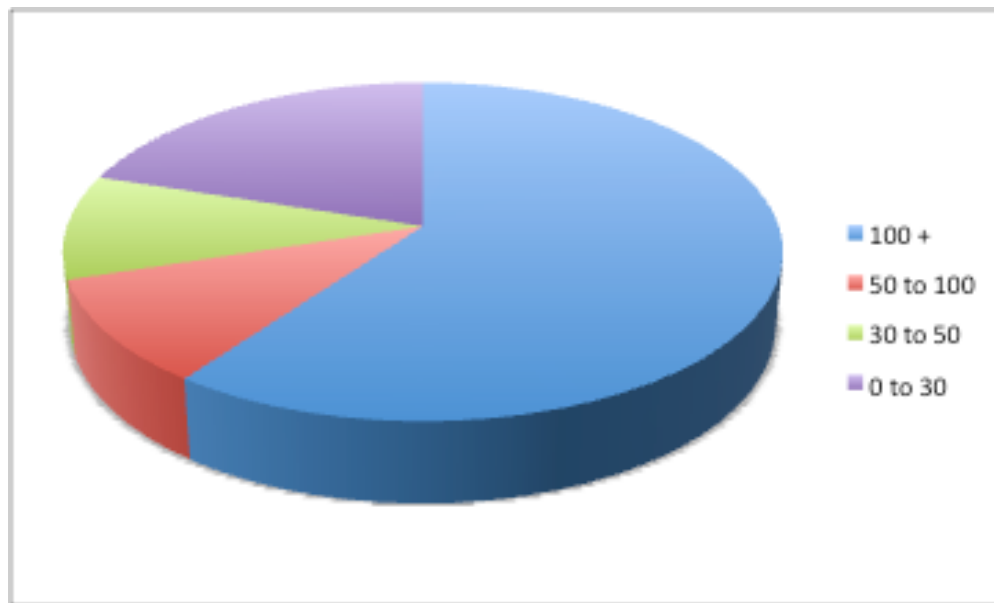


Figure 37 The number of user tests performed

The number of actual user tests that the Knowledge Audit participants have performed were:

6 x 100 +

1 x 50 – 100

1 x 30 – 50

2 x 0 - 30

An interesting observation is that 4 felt that they didn't have extra qualifications that helped them in their role. 6 felt that they did, or that their main qualifications were useful to them in this field.

6.3 METHODOLOGIES

6.3.1 Are you aware of any existing user testing methodologies? If so please outline.

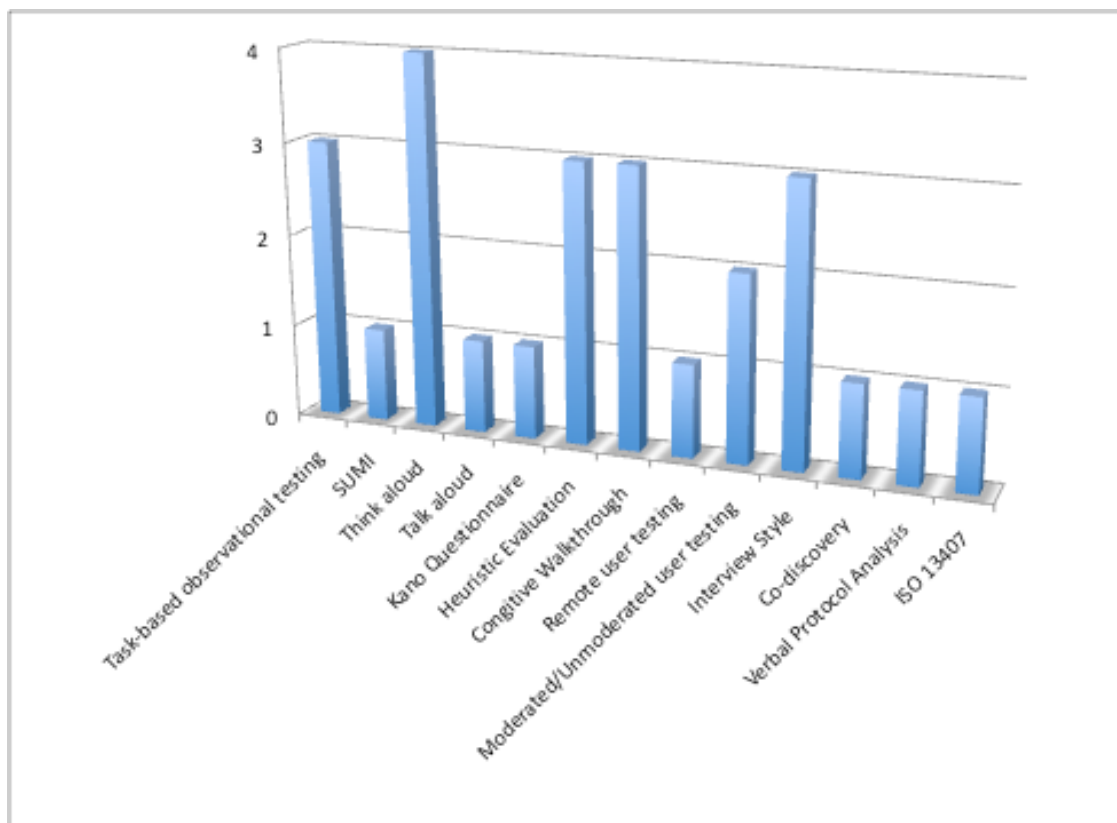


Figure 38 Awareness of Methodologies

An interesting observation is that many of the Knowledge Audit participants use aspects of a methodology but many tests are actually bespoke]. Also the kind of tests that would be undertaken would be very much dependent on context.

Please also note that SUMI, Kano Questionnaire, Heuristic Evaluation, Cognitive Walkthrough and ISO 13407 are not explicitly user testing methodologies, I am including them here as they were referred to in this section by the respondents. They would be considered to be supportive methods or parts of the user centered design process. Does this reflect a sense of confusion among practitioners about what constitutes a methodology in the first place? Some would consider them to be secondary supportive methods in the user centered design toolkit.

6.3.2 Do you use any other usability methods in your projects e.g . Case studies. Focus groups? If so please outline.

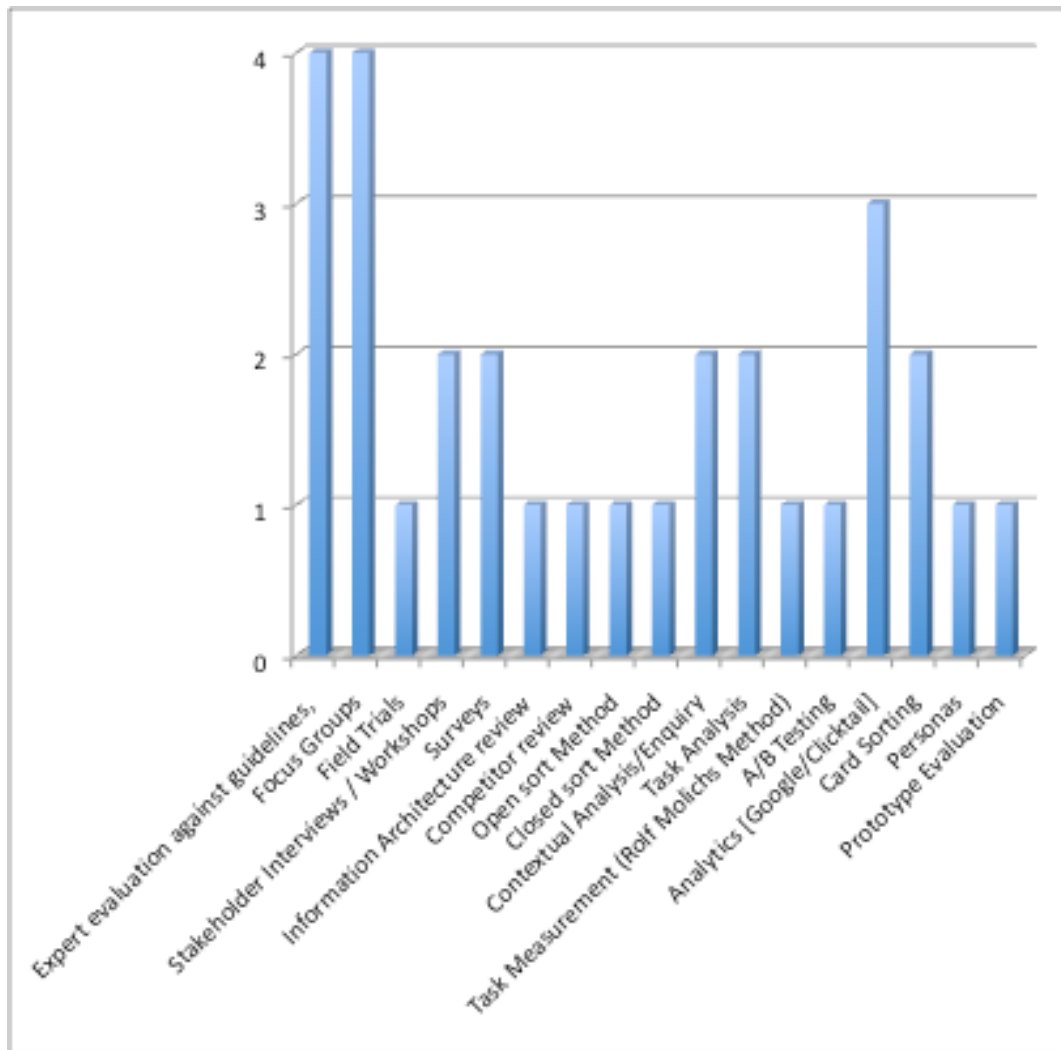


Figure 39 Other usability methods

Listed above are some of the secondary techniques that were used by the Knowledge Audit participants. ‘Evaluation against Guidelines’, the use of ‘Focus Groups’ and thirdly the use of ‘Analytics’ were the most popular.

6.4 SECONDARY SUPPORTIVE METHODS

6.4.1 Are you doing user testing/usability analysis? If so please describe.

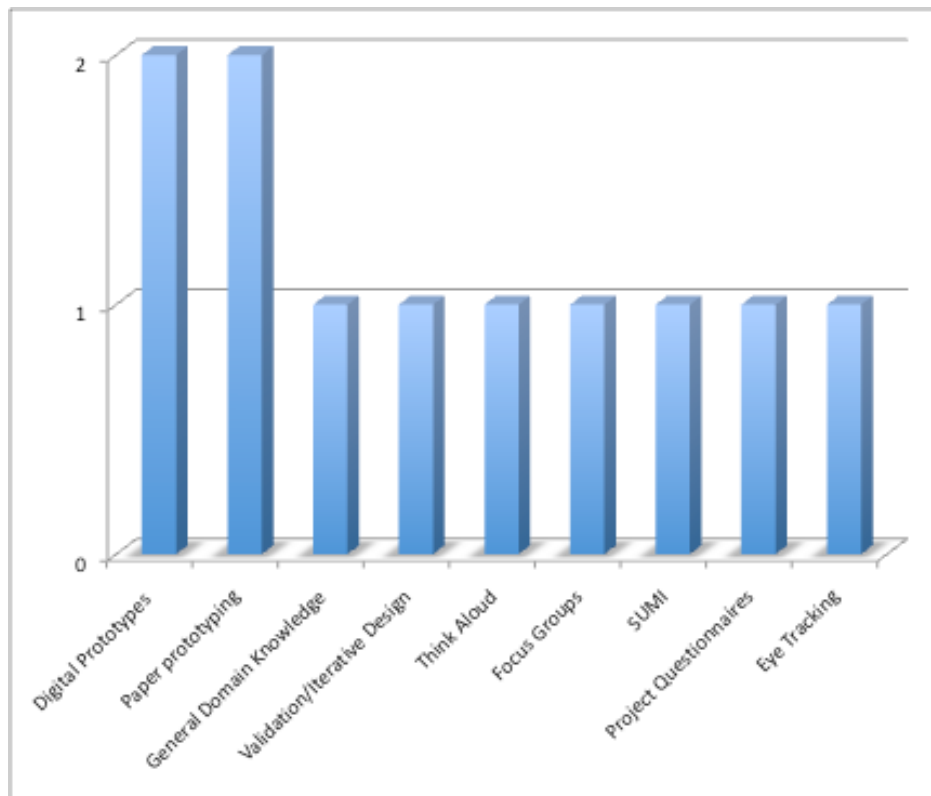


Figure 38 Description of main methods used

Some of the main methods used are:

2 x Prototypes (e.g. digital wireframes), where you can discover a variety of issues, from the interface, the terminology used, structure, or mismatches in expectations

2 x paper prototyping

1 x (By domain, I mean the general knowledge domain, e.g. clothes for Next's website.)

Also mentioned were:

- 1) Web sites, with the comment: *"where you find just about any issues, depending on the tasks you use, the people you ask, and the domain of knowledge."*
- 2) 1 x validate and iteratively improve the design, and in analysis projects to uncover problems.

- 3) Comment: *“I have often used the method of identifying critical incidents, i.e. ask users to perform typical tasks, thinking aloud, while we observe and note any incident of errors or expressed irritation/anger/disorientation etc.”*
- 4) Comment: *“I have sometimes applied focus groups to discuss special issues in more detail. I have applied questionnaires to measure user satisfaction (SUMI).”*
- 5) Comment: *“I have applied project-specific questionnaires to ask for user experiences after a trial.”*
- 6) 1 x employ eye tracking equipment.

6.5 STANDARDS

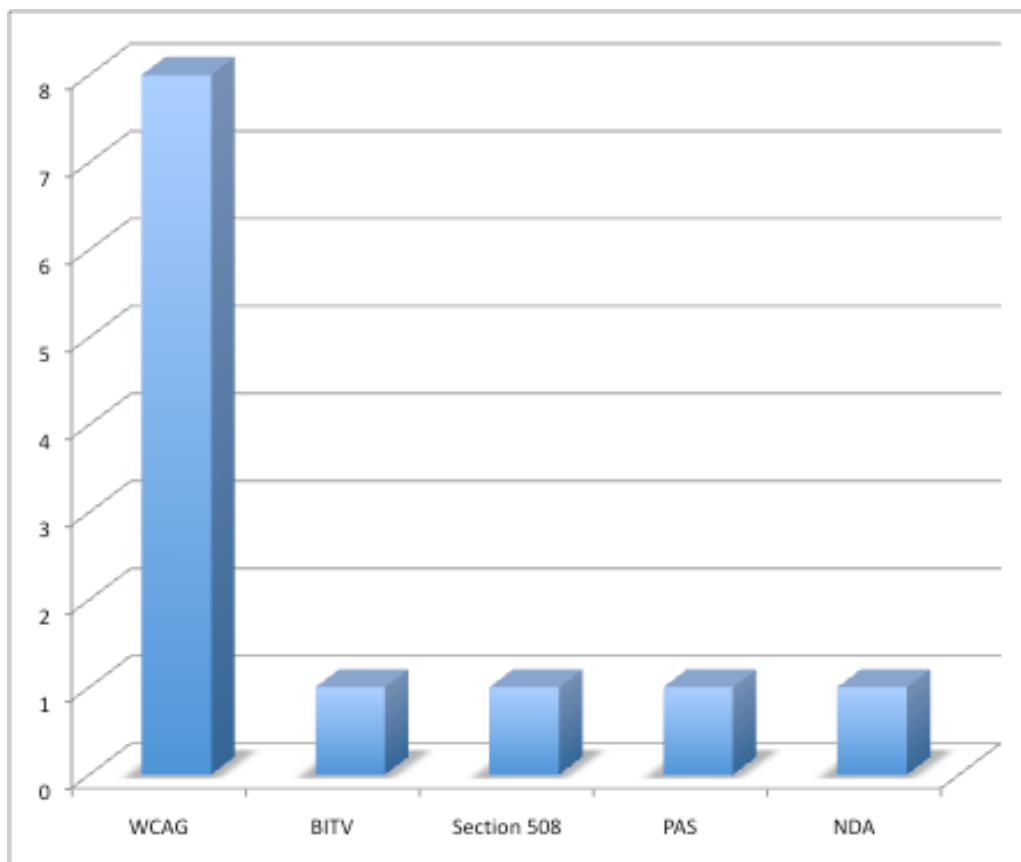


Figure 39 Awareness of Standards

The above graph shows the main standards that the Knowledge Audit participants are aware of, they are:

8 x WCAG

1 x BITV

1 x Section 508

1 x PAS

1 x NDA IT Accessibility Guidelines.

6.6 INFLUENCES

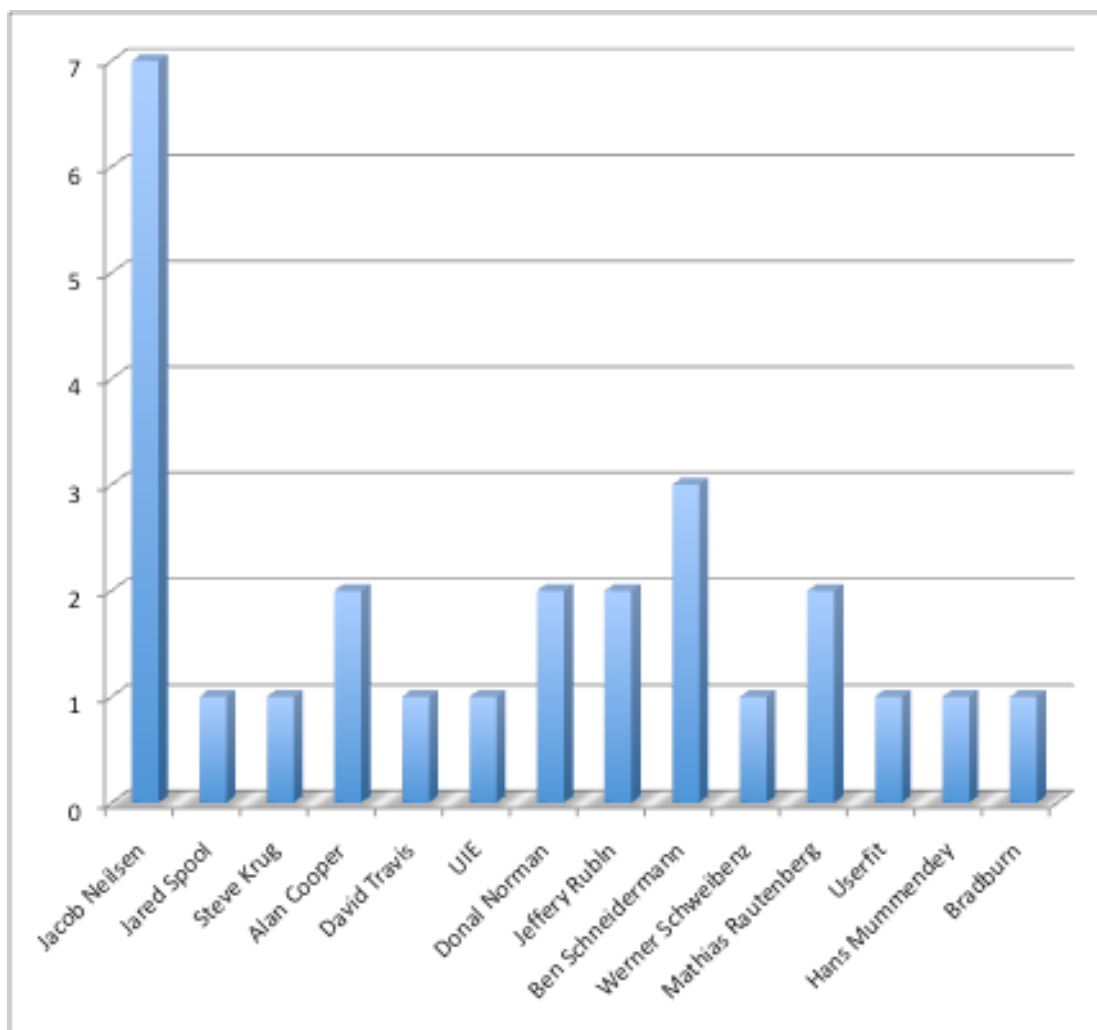


Figure 42 Influences

The following are a list of the main influences, on the professional careers of the Knowledge Audit participants, they are:

- 7 x Jacob Nielsen
- 3 x Ben Schneiderman
- 2 x Alan Cooper
- 2 x Donald Norman
- 2 x Jeffery Rubin
- 2 x Mathias Rautenberg
- 1 x Jared Spool
- 1 x Steve Krug

- 1 x David Travis
- 1 x UIE
- 1 x Werner Schweibenz,
- 1 x Userfit,
- 1 x Hans D. Mummendey (Die Fragebogen-Methode)
- 1 x Bradburn

6.7 Assistive Technology (AT)

While having in depth knowledge of Assistive Technology is not necessarily a requirement to effectively perform user testing. The author does feel that it certainly helps the validity of suggested changes to interface design when the test facilitator understands how the AT works. In general, there is a good level of awareness of AT amongst those questioned.

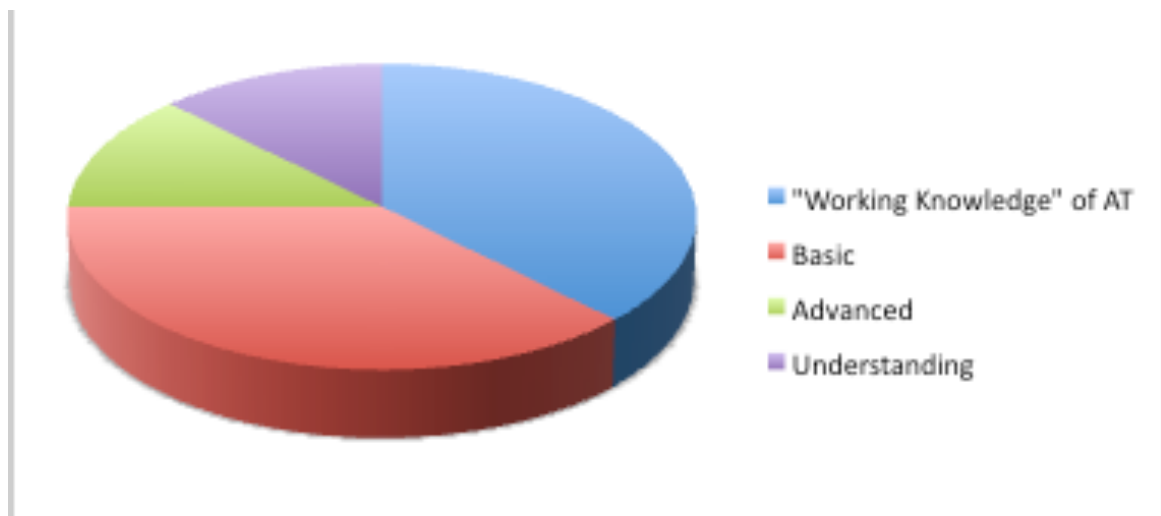


Figure 43 Level of AT Knowledge

4 x of the Knowledge Audit participants described themselves of having a “Working knowledge” of AT

3 x “Basic”

1 x “Advanced”

1 x has some “overall general understanding”.

Note that one said that his knowledge was "Advanced however, he would not base ‘usability’ decisions on my use". This is an interesting comment, as even though he perceives himself to have an advanced knowledge he recognizes the importance of testing with real users to give validity to the results. This is not to say that expert testing has no value but it is important that there is an awareness of the varying levels of experience, literacy and competence among AT users.

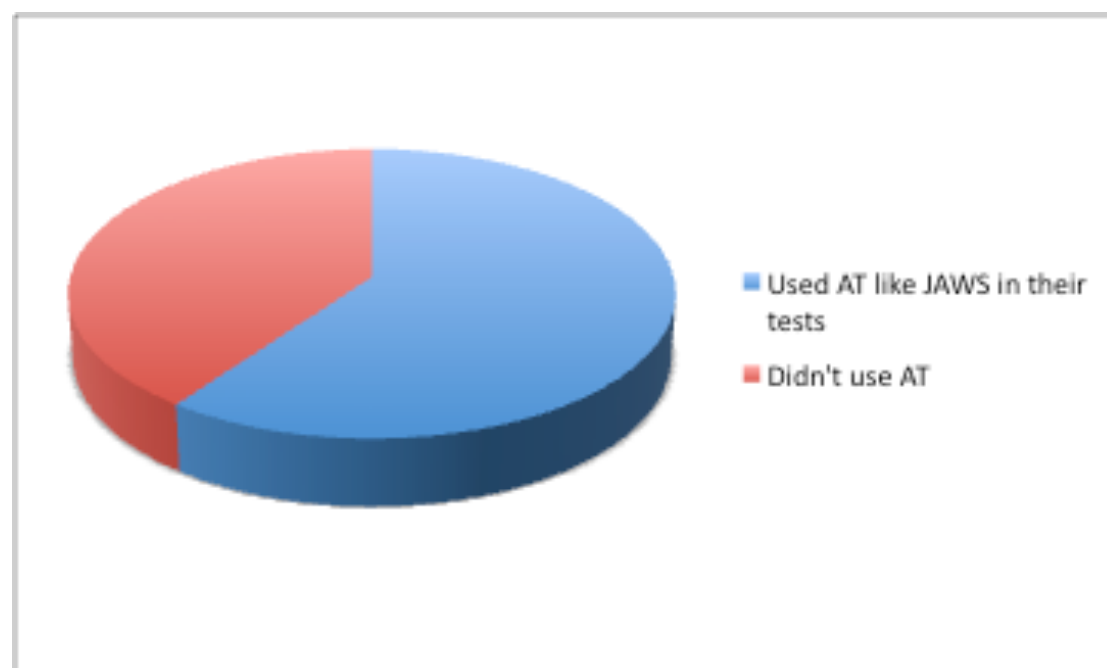


Figure 40 Overview of AT testing/experience of facilitator

6 x would use AT like JAWS in their tests to some degree. This may be just to get a ‘feel’ for the application:

3 x said they were confident in the results that would get using AT themselves (as non native users as such) but there was a certain degree of hesitancy by the majority to have any absolute confidence in this kind of testing.

1 x stated that *"We don't rely on that type of testing, so not really."*

1 x stated that *"the technical aspects of testing are easy to test. But there are issues such cognitive aspects and the users familiarity with their AT that effect the outcomes.]"*

There is an awareness that regular users will just do things differently, and while testing with AT (if you have some degree of fluency) is useful to double check issues, unless the user is advanced there may not be much confidence in the results.

6.8 USER TESTING PRACTICES

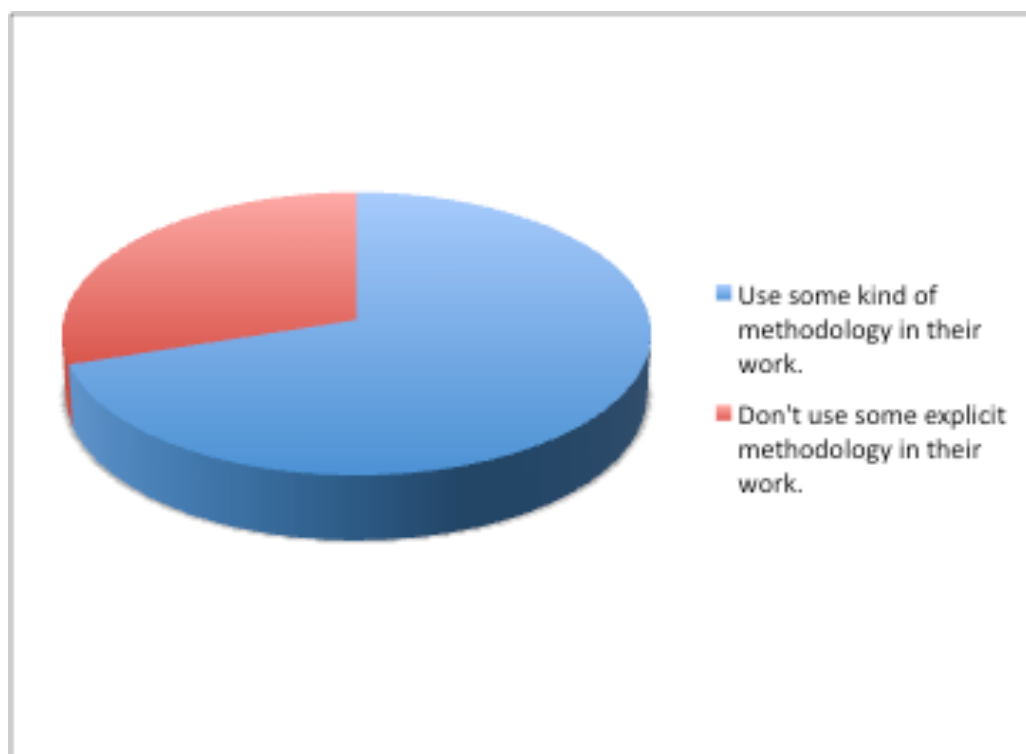


Figure 41 Is a methodology used in your work? Overview of Methodology use

7 x say they use some kind of methodology in their work.

The question of awareness of user testing methodologies is a core issue in this research (as raised also in Chapter 6 “Current Methodologies and the Research Question”) most of the Knowledge Audit participants stated that they use some kind of explicit methodology in their work. The following are some interesting comments from the respondents are that should give some context to this statement:

1) *“We use these methods mainly for requirement engineering and implementing our user-centred design approach when developing software. “*

2) *“Most of the time, I have done usability engineering while the system was still under development, therefore formative evaluation was needed.”*

3) *“Moderated user testing allows us to interact with the participant and get the most out of it. Remote or un-moderated user testing doesn’t allow this. “*

4) *“Perhaps it would help to outline usability testing techniques/methods that I haven’t found useful? “*

5) *“I haven’t used remote, un-moderated testing, where you setup an application that people login to, and are given tasks automatically, and put feedback into a box. To me this combines the worst aspects of several other methods:
- It does not allow you the observation of the participant. Even if you only have audio and can see their screen, you can observe what they haven’t seen or understood. If you only go from their feedback, you only know what they think they understood, which often misses the most valuable information.*

*- It doesn’t have the breadth of results you get from analytics.
- It doesn’t have the realism you get from a survey. The setup isn’t that dissimilar from an online survey that is linked to from within a site, except that people already on a site you know have a reason to be there. People logging in to make some money from testing may not have any reason to be there. I also don’t tend to use eye-tracking. There are a few niches where it can be useful (e.g. checking for advertising visibility), but for most situations it doesn’t add anything. It does provide lovely looking deliverables for clients, however, not that many are willing to pay the extra cost these days!”*

(See <http://www.uie.com/brainsparks/2006/06/13/eyetracking-worth-the-expense/>) “

6) *“Task-based observational testing with think-aloud or, more usually, talk-aloud. Not protocol analysis”*

6.8.1 If you do not adhere to a particular methodology, please outline why? Alternatively, have you created your own methodology that works within the context of your role? If so please outline how you work.

The following are some interesting comments that outline how the Knowledge Audit participants work.

More interesting Knowledge Audit participant quotes/comments:

- 1) *“Task-based testing is relevant because it is the effect of accessibility and usability barriers on task performance that is of most consequence. Observation is relevant both as a way of discovering problems and the reasons they arise and as a way of giving other people (especially the developers of the product or service being tested) insight into users’ needs and design flaws. Think-aloud will generally give better results than talk-aloud if the test user can do it, but most can’t so it’s not that useful. “*
- 2) *“Hopefully the above answer this, usability testing is a method within UCD.”*
- 3) *“We are moving away from think aloud style tests based on the results of eye tracking data. The results suggest that users behave very differently when they are explaining their actions and often post rationalise their explanations. Eye tracking allows us to play back and review what really happened in much more detail giving us a more accurate prediction of how products will perform in real life. “*
- 4) *“They are critical in helping us understand why a design does or doesn’t work.”*
- 5) *“We run to a pretty standard format – but are constantly changing and tweaking our method with a mind to getting the most useful and natural results. “*

- 6) *“There’s not much in terms accessibility testing which is what we mostly do. I think expertise we have within the organisation beats all textbooks. Although I wouldn’t mind formal training.”*
- 7) *“I have preferences (iteratively doing formative evaluation, using observation and interviews) but I can choose from a variety of methods to suit a project situation and purpose.”*
- 8) *“Mostly we do evaluations in the field of accessibility”*

6.8.2 If there is any other user testing methods information you feel is relevant, please feel free to add it here, thanks.

The following are some interesting comments that outline how the Knowledge Audit participants work.

1) *“Something I’ve noticed recently is a trend towards usability testing with people who have disabilities, rather than working to standards. Although you’ll almost always get useful information from usability testing, I worry that people are essentially doing this because they can’t meet standards (e.g. Sharepoint <http://alastairc.ac/2009/11/sharepoint-2010-accessibility-event/>). But testing with a few people using specific technologies does not ensure accessibility.*

“Usability in general (and usually usability testing specifically) is about optimising for the majority. Accessibility is about making usability more widely applicable, making sure the edge cases work.”

2) *“The use of specific methodologies are useful where the results of the tests needs to be statistically significant but may not be necessary where the end goal is to ensure that the product design is improved - therefore the tests in the commercial world often need to be designed to that end.”*

The question was asked in the section ‘Usability as a quality objective’ about whether there was an ‘ideal methodology’ or if ‘one size fits all’? It seems from the above comments that practitioners often use different methods as needed depending on

circumstance, and there is an awareness of a degree of flexibility in the user centered design toolkit.

6.8.3 LAB DETAILS

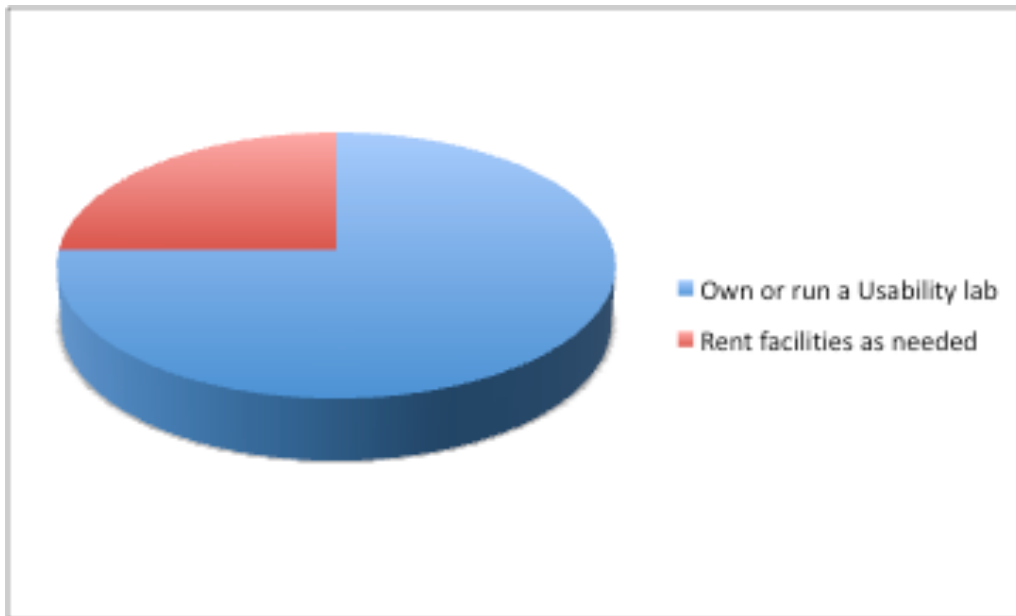


Figure 42 Do you own or run a usability lab?

6 x

Knowledge Audit participants either own or run a usability lab
2 x Rent facilities as needed

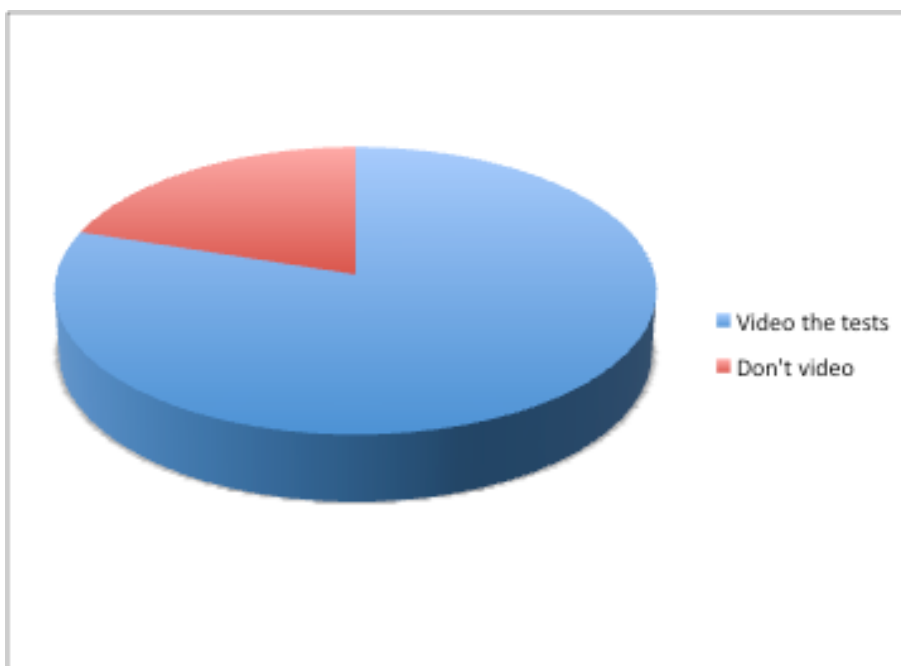


Figure 43 Use of video in the lab

8 x Knowledge Audit participants video the user tests

8 x use screen capture software (Morae being the most popular, then Silverback and VLC)

An interesting observation is that so few use the data analysis features in applications like Morae, of those interviewed only:

2 x use the data analysis features in Morae.

1 x used to but stopped.

1 x commented that there is often someone in the test who is taking notes.

1 x uses exit surveys, task time to completion data.

1 x the video clip markers are all they really need to use.

It is very interesting that many of the advanced features are not used at all! This then introduces the questions, are they useful at all? Or are practitioners missing some vital metrics in their analysis that could help to improve their outputs?

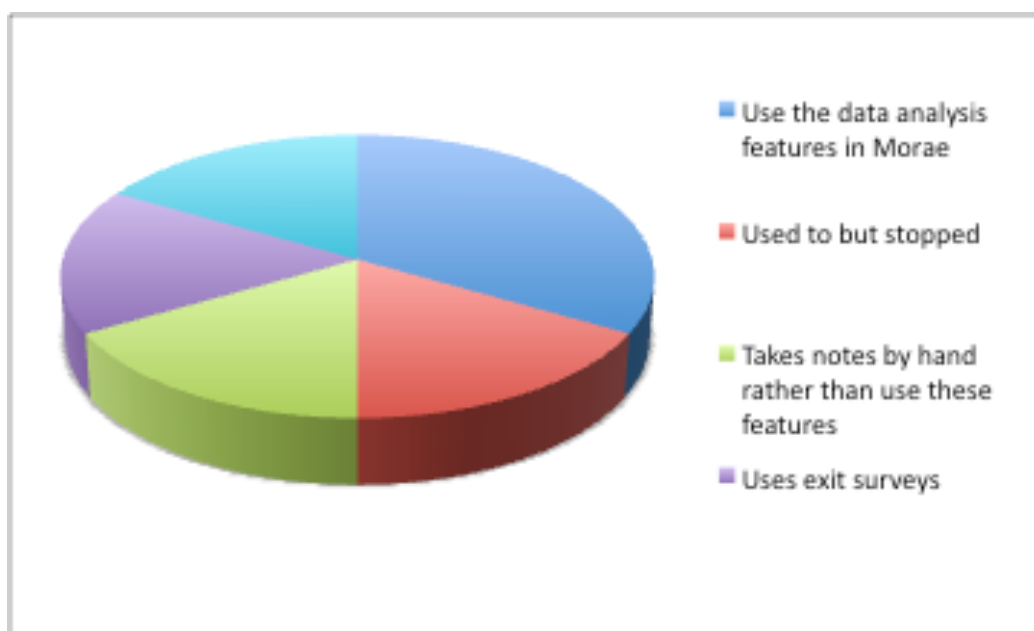


Figure 44 Use of data analysis tools

8 x give video footage to their client.

7 x prefer to give edited footage that is illustrative of issues.

1 x charges extra for this and prefers to give raw footage.

Of the Knowledge Audit participants:

5 x highlight issues and recommendations.

2 x Give high-level data analysis sometimes (e.g. number of people completing or failing a task). More qualitative than quantitative.

1 x User comments are used as examples to re-enforce points.

1 x Give an overview as well as detailed reports.

1 x Highlights issues with user comments and quotes.

This is interesting as the desired outputs are dependent on the project.

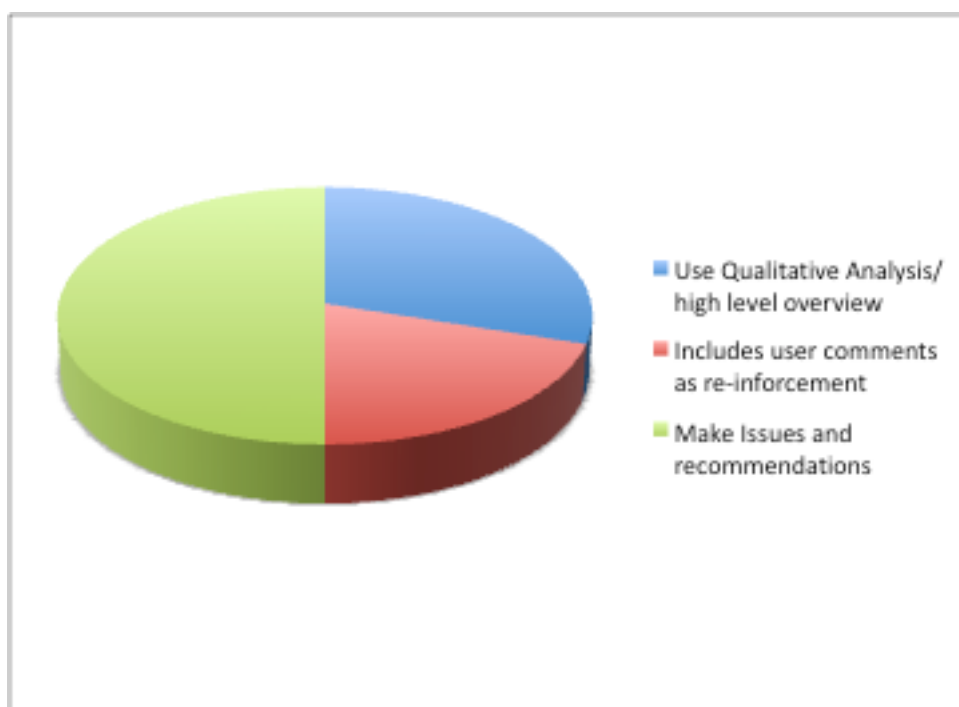


Figure 45 How are the outputs from tests are used?

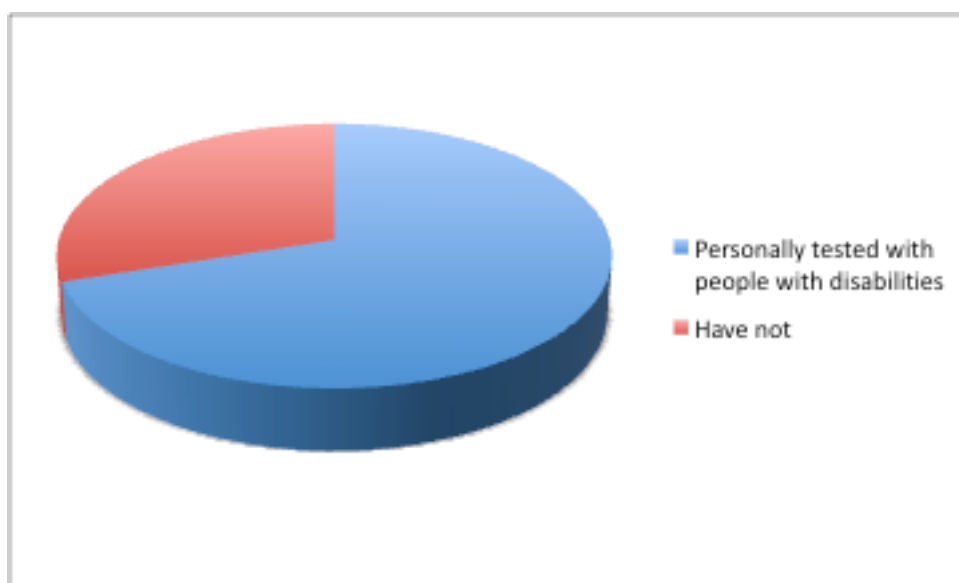


Figure 46 Do you test with people with disabilities/older people?

Out of all of the Knowledge Audit participants:

9 x test with older people

7 x personally tested with people with disabilities

1 x used eye tracking software regularly

1 x use eye tracking rarely.

6.8.4 Disability Types

In terms of disability types, find listed below details of the groups that the Knowledge Audit participants have worked with.

9 x Blind users

9 x VIP

5 x Cognitive/Intellectual

7 x Physical

5 x Combinations such as physical/VIP

2 x Dyslexia

3 x Deaf

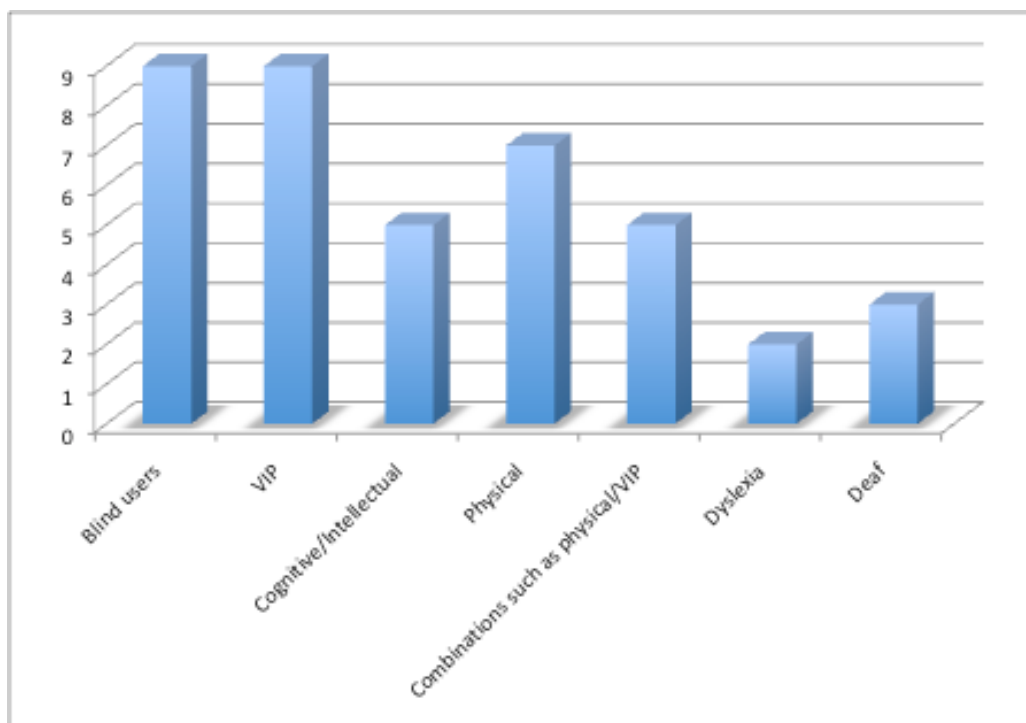


Figure 47 Disability types tested with

6.8.5 Numbers in tests

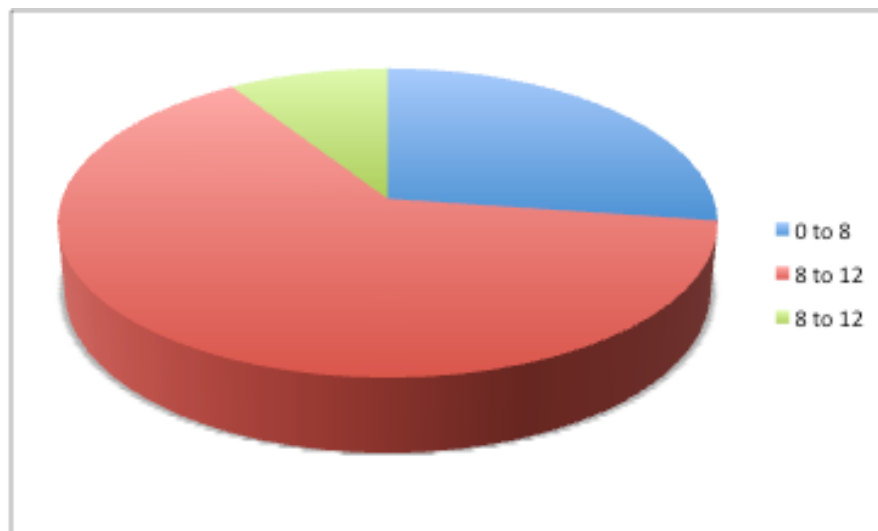


Figure 48 User Test sample sizes

For the Knowledge Audit participants the majority of test sizes use between 8 – 12 users.

7 x 8 -12

3 x 0 - 8

1 x 12 for a summative evaluation, or when user group is not homogeneous.

Of the Knowledge Audit participants 8 say that they prefer iterative design process, 2 had no preference.

As was asked in section “Iterative Models: Agile” about consensus amongst practitioners on whether iterative models are best, the above data seems to indicate that from the point of view of the practitioner that the most practical method of incorporating user feedback and analysis into development process are flexible iterative methods such as the Agile family.

6.9 Outcomes of User Testing

6.9.1 What do you feel the main benefits of user testing are?

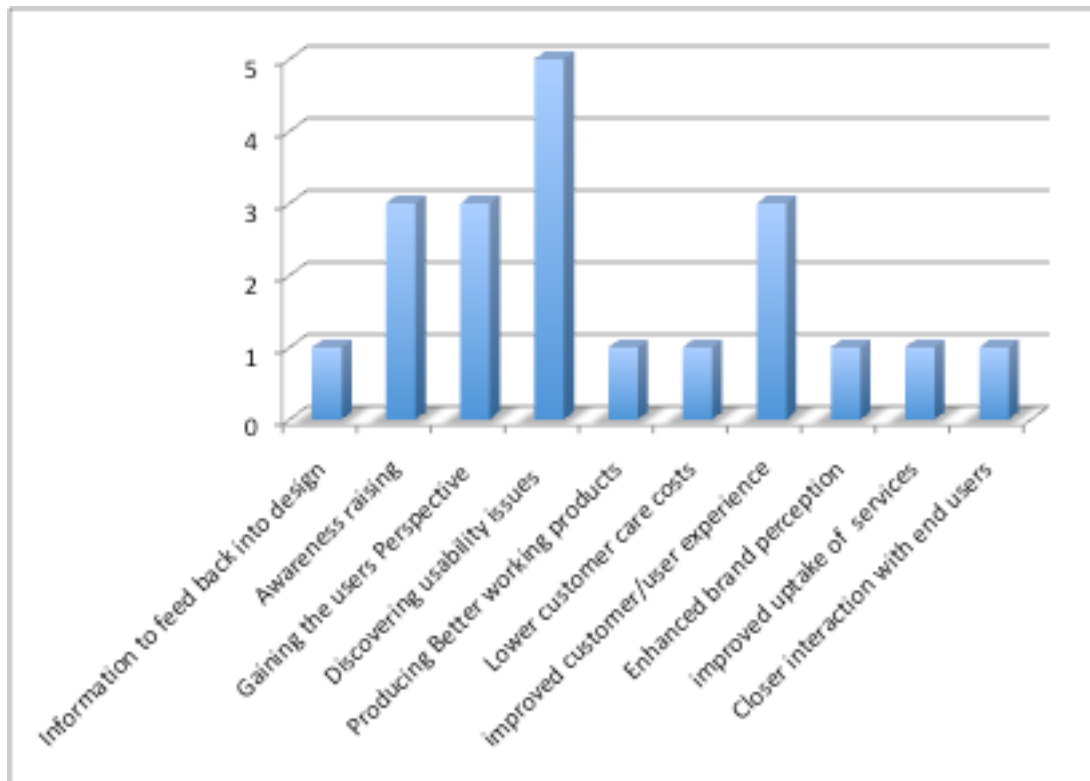


Figure 49 Practitioners view of the benefits of user testing

From the perspective of the practitioner, the main benefits of user testing are:

5 x Discovering usability issues interface, navigation, structure, functionality and objectives levels

3 x Awareness raising

2 x Gaining the users Perspective

3 x Improved customer/user experience

1 x Providing Information to feed back into design

1 x Producing Better working products

1 x Lower customer care costs

1 x Enhanced brand perception

1 x Improved uptake of services

1 x Clients see the benefit of closer interaction with their customers.

6.9.2 Are the results of user testing incorporated into projects? If so, how?

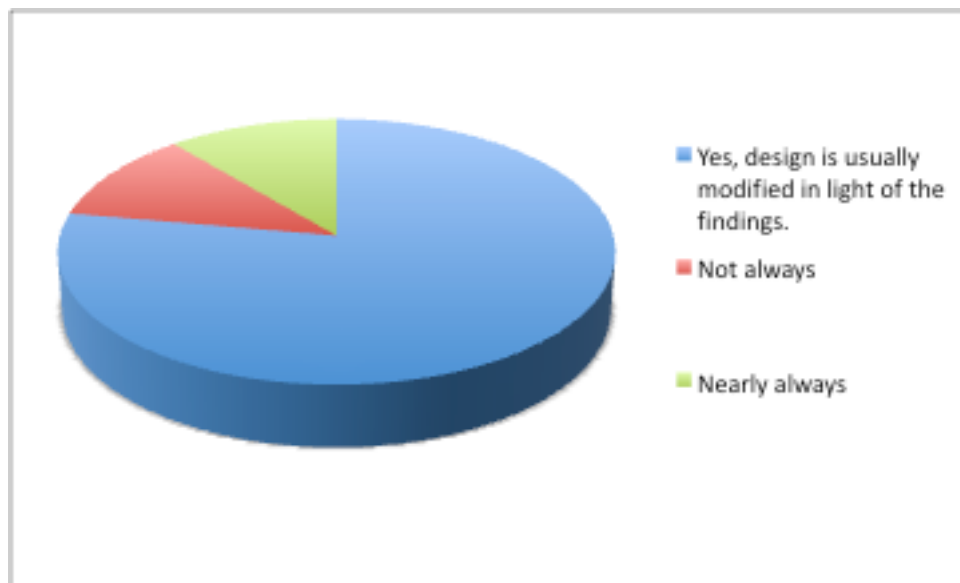


Figure 50 Incorporation of user testing results

7 x Yes, the design is usually modified in light of the findings

1 x Directly fixing an issue, by adjusting the interface/code or adjusting content

1 x Working around an issue (e.g. adding help).

1 x Yes, when tests are part of a design process that is structured to incorporate user test results

1 x In analysis projects the output [of the tests] is combined with some other data source e.g. analytics to generate recommendations and design improvements.

1 x If the findings leads to actionable fixes,

1 x Not always

1 x Nearly always

1 x Depends on the project, its management, stage and goals. List of necessary improvements and updated requirements handed over to the developers.

1 x Findings were addressed in the functional specifications [of the project].

6.9.3 Have you ever undertaken user testing more than once in the same project?

All of the audit participants said that they have undertaken user testing more than once on the same project. One said that they *"do this regularly"* another said that *"they do it at least three times in a project where the project is a replacement for an existing product or service we advise a benchmark test on the original, a prototype test and then a late stage pre-release test."*

1 x will test the existing site and then the new site.

1 x as a requirement as a part of the human-centered design process according to ISO 13407. [for EU research projects as they have an iterative work plan. However, for paying clients, the user test may be a singular event]

6.9.4 For multiple user testing sessions was it beneficial, if so how?

1) *"Yes, it's the only way to do that kind of thing. You get to improve the interface based on one set of tests and test to see whether the improvements have made \neg the intended \neg difference. Or you get to add new features and see what impact they have and whether they are well integrated."*

2) *"A lot less issues were built into the site!"*

"It's always difficult to say, because the site without the usability testing input doesn't get made. However, aspects such as navigation, labeling, and even content tend to be further optimised with each iteration of testing."

"Sometimes you can build this into one day of testing, where you change the site after each (few) sessions. However, you do lose the ability to compare across participants if you do this."

3) *"Yes – we were able to see how our redesigns had improved the user experience (or made worse)"*

4) *"Very, we uncovered issues at each stage and resolved them as we went."*

5) *"I would rather involve testing early on in the development process. Issues can get lost in reports and often we find designers not correcting the issues as they've not really read or understood the report."*

6) *“Yes. A research project aims to develop some innovative technology. This is only feasible with several iterations. BTW, I count scenario evaluation as an early opportunity of user evaluation, testing for validity of concept and of our understanding of context of use and user needs. “*

7) *”Yes, we adapted the user interface, functionalities etc. of the developed system according to the findings.”*

6.9.5 When testing often in the same project did it reinforce you initial findings or contradict them in any way, or did it shed fresh light?

1) *“Often it would reinforce. Sometimes it would contradict but generally in those cases, the problem was due to the tests having been poorly designed and/or poorly controlled or not having enough users.*

2) *“Assuming you can make changes, it will generally find new things. Often you will extend the testing to cover areas that haven’t been explored yet, because participants complete them much more quickly. [In general you find new things]”*

3) *“Depends, most of the time results show a gradual improvement. However on the odd occasion a change made to solve an issue in a previous test throws up a new issue in subsequent test rounds.”*

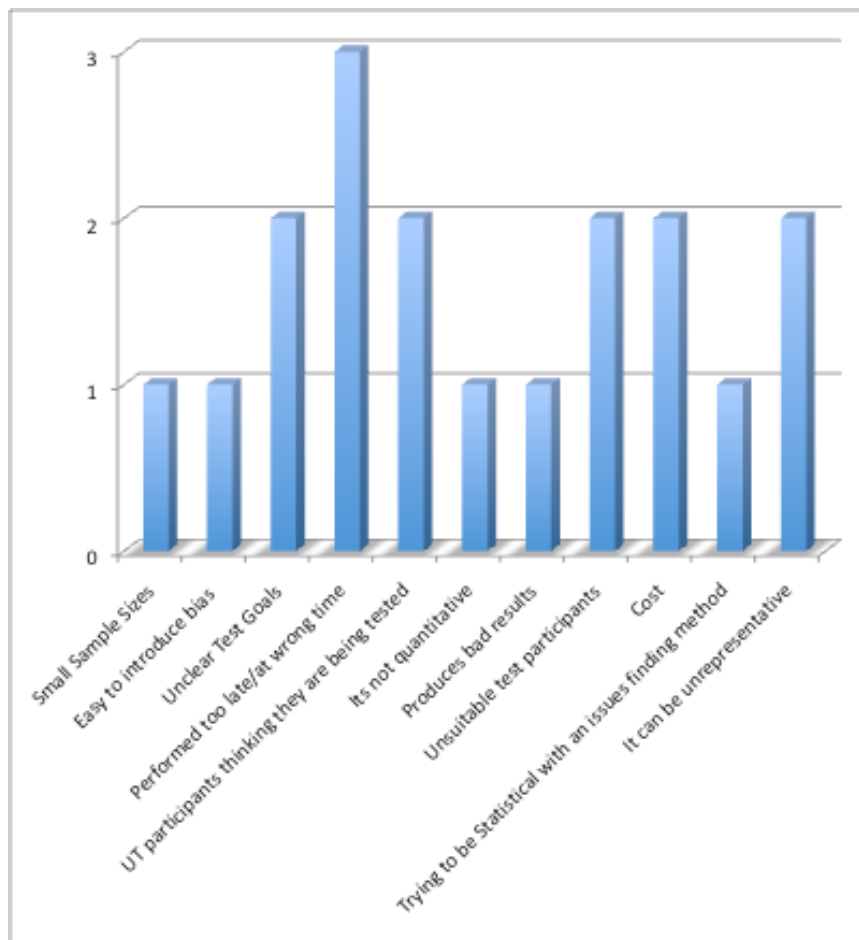


Figure 51 Main deficiencies in User Testing

- 4) *“The client has no obligation to make the changes we suggest. Sometimes I find sites which do accessibility testing just to tick the box. They do not amend the site in any way unfortunately.”*
- 5) *“The next prototype may be quite different with different, fresh usability issues. Also, we could identify some more requirements, we eventually corrected our rating of user requirements. “*

6.9.6 What, in your opinion, are the main deficiencies with user testing?

- 3 x Performed too late/ or at the wrong time
- 2 x Unsuitable test participants or testing with the wrong audience
- 2 x Easy to introduce bias through poor design of tasks and scripts and poor facilitation
- 2 x Unclear goal of the tests
- 2 x Cost
- 2 x It can be unrepresentative, non-exhaustive
- 2 x It's easy to jump to the wrong conclusions
- 2 x People [user test participants]
- 1 x Small test samples
- 1 x Validity and reliability, thinking you're testing their skills [The white coat effect.]
- 1 x It's not quantitative
- 1 x it can produce bad results
- 1 x Trying to be statistical with an issues-finding method.
- 1 x Confusing optimisation for the majority (usability) with ensuring access (accessibility).

6.9.7 Are there aspects of how you undertake user testing that you would like to improve?

- 2 x Increase the number of test participants
- 2 x Experiment with new kinds of testing (such as un-moderated testing or focus groups)
- 1 x More time for preparation
- 1 x Improve my moderation of tests
- 1 x More iterative testing of the product 1 x run more tests quickly
- 1 x More consistent test design to allow us to compare the results of test on different projects

6.9.8 If there is any other practical user testing information you feel is relevant, please feel free to add it here, thanks.

One very interesting comment from a test participant was:

“Apart from using video to provide illustrative information to the client, it is very important, though very time consuming, to review the video when doing the analysis and drawing conclusions. This is partly because it is very difficult to write down the pertinent information in real time while observing a test. It is also because it can take many times of rerunning a sequence before you really figure out exactly what has happened and why, because you need to piece the ‘story’ together from what is usually very little and very subtle evidence from screen activity, user actions and user utterances. It is like being a detective and very easy to jump to the wrong conclusions if you don’t take time to review the evidence carefully. It often requires the analyser to compare the footage from different users on the same task.”

7 CONCLUSION

7.1 Introduction

The Knowledge audit has produced some interesting and varied results. We have gotten an insight into how current user testing is undertaken in a range of commercial and academic domains. This provides a rich ‘snapshot’, so what can be extracted from this data and what conclusions can be come to?

7.2 Research Definition & Research Overview

The objective of the experiment in this research is an exploratory one, to capture a snapshot of the current state of practice in UCD, so it can be seen as a social anthropology study.

By exploring what people are doing at the moment, it is possible to assess what they know and to identify issues and trends in the field. The ‘Knowledge Audit’ form of this research has leant itself well to giving a rich qualitative overview of this domain and a certain ‘ontological depth’. Some more qualitative statistical methods of research would not have been entirely suitable, as they would not reflect the intricacies and depth of what is a very nuanced area of work. (Yeung, 1995)

7.3 Contributions to the Body of Knowledge

Some of the striking issues that arise from this research are:

In practice approaches to the informal user testing process seems to be rather random. It is possible that the adoption of some aspects of more formal testing may benefit the discipline of user testing as a whole. Certainly more ‘metrics’ need to be incorporated into the outputs of user testing (such as within reports and recommendations etc) in order to be able to back up assertions that are made about the quality or lack thereof within a certain User Interface design.

For most user testing in the ‘real world’ (outside of a research situation) there seems to be little room for scientific rigour and a general ‘make do’ approach.

This is probably a natural consequence of the context within which much usability analysis is undertaken. It is not that informal user testing should be abandoned however, we have seen in the cited Case Studies in Chapter 3 how adversely effected user testing can be by because of this lack of rigour.

The feedback from professionals in this research indicates that isn’t any real consensus on how to actually do user testing. There seems to be a great deal of variability in how it is undertaken, what is done with the outputs etc.

The general issues outlined in this research are applicable to the discipline as a whole, whether the testing is with or without people with disabilities. Interacting, working with people with disabilities can certainly be seen as an effective way of gaining a much deeper understanding of the needs of a diverse user group, for effective UI design. However, effectively doing the tests, and successfully incorporating the outputs from testing - with any user group - seems to be the overarching issue.

So a burning question is how could some of the quantitative advantages of the more formal method be used to support and improve the quality of the outputs from more informal user testing? Informal user testing is often dismissed as being at best (by its nature) unscientific, and at worst misleading. This research does indicate the need for aspects of the more formal method to be introduced into informal usability testing methods to improve its integrity. Also any change in the current model must be sympathetic with the context of the testing such as commercial pressures, limited budgets, deadlines and so on.

A key finding of this research is that there appears to be very little definitive consensus regarding the effectiveness of usability methodologies amongst practitioners and this may lead to confusion in the field. The only real strong evidence is that there is a general preference towards the iterative testing family of methodologies and an appreciation of a responsive design process.

However, there is little agreement of exactly how best to achieve this beyond the need to constantly design, evaluate and redesign.

Chapter 2 examined User Evaluation Methods in ‘Iterative Design Process’ H. Rex Hartson, Terence S. et al and it was suggested that there is a “lack of knowledge amongst usability professionals as to the strengths and weakness of UEMS” this research suggests that this is not the case. The research undertaken in the Knowledge Audit indicates that there is actually a great deal of awareness of the faults and failings of user testing and evaluation methods in general. There is also a consensus that it is a good way of improving the quality of the user experience - even if some of the practices are less than ideal.

However, Lund’s hypothesis mentioned in Chapter 2 ‘Iterative Design Process’ about the need for greater metrics in usability analysis certainly seems to be true. What is interesting is that hardly any of the respondents of the test would really use metrics of any kind at all in their final report writing. Also hardly any of the respondents used some of the analysis features in video capture software like Morae, that can allow the presentation of these metrics.

Another important issue is looking at ways of improving exactly how usability practitioners can effectively communicate the results of their analysis and testing. In my experience as a practitioner, one of the most effective ways is to actually have the tests observed in real time by the designers and developers of the system, this has an immediate and kinetic effect on the observers when they see the raw, real-time interaction. There is a tangible immediacy in this kind of observation and also the repetition of observing test after test can certainly be effective in hammering home any of the issues that can surface during the test. Unfortunately, as observed earlier in Chapter 3 ‘Are 5 Users Enough?’ you cannot be guaranteed that even major usability issue will be repeated or even discovered in small tests, but when it does happen the impact on the developer can be profound.

Finally, there is no doubt of the important of the role experience plays in effective user testing. User Test Facilitator experience is incredibly valuable in making qualitative judgment calls in what is effectively a ‘subjective user experience’.

What is apparent is that among practitioners there is just no ‘one right way’. Therefore, it is worthwhile contextualising user testing as being only a part of the overall usability toolkit that collectively helps to reduce barriers in the digital environment. The goal is after all to improve the quality of life for all of us and to help us to better realise a more inclusive society.

7.4 Experimentation, Evaluation and Limitation

The experiment, which is a social anthropology study, did uncover some care questions:

- ‘What is the correct amount of people to test with?’
- ‘Can expert evaluation be relied on considering the great of variability in their results?’
- ‘What are the best ways to communicate the results of user tests to clients?’

The answers to these questions based the evaluation of the data indicate that, one of the limitations of user testing is having the correct number of test participants who are suitable. Obviously having more people to test with will increase the likelihood of uncovering a greater number of issues, and this is indicated in the research. Also however, they must be suitable, in the sense that they represent the qualities and characteristics of the target user population, that they have a suitable skillset and general level of digital literacy that is suitable for the target group. Exactly how many, is hard to define but the more the merrier.

While expert evaluation may not always be entirely suitable (solely) for analysis, it does have some benefits. AS the experts themselves may even be a good fit for the target population. There are certainly some issues that experts will easily spot within a UI. So while expert evaluation should not be considered a silver bullet, it is useful and should sometimes be incorporated in tandem with traditional User Testing, maybe even in some cases replacing it where for example, there is an expected fluency and high level of expertise or power use, in a sample population.

In order to effectively communicate the results of user testing, certainly dead static reports are not the way forward.

It seems that engaging the client in the process of inclusive design and capturing their imagination, and helping to see the many ancillary benefits, such as increased sales, easier maintenance etc are important. The use of observation suites where tests can be viewed in real time is effective in engaging the designers' attention. Using video clips that illustrate outstanding issues is also useful. Finally, it seems that metrics are to play a key part in backing up the assertions made by UI analysis, as the reporting of a test facilitator or usability analyst is subjective and therefore greater metrics will help to backup assertions about the need to change certain aspects of a UI, and to convince the client to allow this change to take place.

Finally, some limitations of this research are finding a 'perfect' methodology. This just may not exist.

7.5 Future Work

It can be inferred from the findings that there is no consensus on just exactly 'how' to best do user testing and user evaluation. It is hard to tell if this issue is exacerbated when testing with people with disabilities, as this does require more specialised knowledge and experience, so more research work is needed in this area.

An interesting factor in user testing that is often encountered is "But the user said [...] so it must be [...]". It is implicit in this kind of observation that the user is always right and that their feedback should always be acted on. However, what if a user with a disability has an overall poor level of digital literacy? What if they do not know how to use their Assistive Technology properly? Can we then with impunity blame the UI for a poor user experience?

It also is worth exploring the issue of what a user says, vs. what they mean, and being able to tell what they do, from what they actually meant to do. Actually, it can be very difficult to make a proper assessment of subjective user experience. There are times when the user comments and 'think aloud' feedback can be very useful (such as "This is terrible, I would just give up now" or " I love this! It's so simple") but these examples reinforce what is already there and are easier aspects of the user experience

to understand - while they are both at the extreme ends of the user experience – they are more binary responses to certain system qualities or states.

The grey areas of analysing the user experience are where it is harder to respond as a professional in any way that is meaningful. As mentioned, there are often other background variables that need to be considered such as – the users general digital literacy, their familiarity with AT (if they are using it), the complexity and suitability of the task in order for the user to achieve their goal and so on.

Other interesting questions that arises out of this research:

- Are there actually methodologies for effectively measuring the outputs of the user test?
- How do we create suitable metrics for real world testing such as commercial environments etc that can be used to aid analysis or user testing data?
- Considering the time and financial pressure that the usability professional may be under, how can improvements be effectively adapted to existing workflows?
- What is the best way to effectively train professionals to gain the skills necessary to progress this domain?

7.6 CONCLUSION

A wise man once said that the measure of a man is his highest ideal, so it is good that so many are aspiring to at least try to create a world that has an equitable foundation, even if it may never appear to be fair.

The field of inclusive design is an exciting and challenging domain with many aspects that reflect the diversity of humanity itself. As life is in a state of constant flux so is the environment and this presents obstacles, and opportunities, to us all.

BIBLIOGRAPHY

Abrahamsson, P. Warsta, J. Siponen, M.T. Ronkainen, J.;
Tech. Res. Centre of Finland, V77 Electron., Oulu, Finland (2003) New directions on agile methods: a comparative analysis. Software Engineering, 2003. Proceedings. 25th International Conference, p 244 – 254. ISBN:0-7695-1877-X

Agile Manifesto (2001) Agile Software Development. Available from:
<http://agilemanifesto.org/> [Accessed January 2010]

Agile Principles (2001) Agile Software Development. Available from:
<http://agilemanifesto.org/principles.html> [Accessed January 2010]

AT Animation (2011) Inclusive Technologies. Available from:
http://www.inclusive.com/AT_boogie/at30.swf [Accessed February 2010]

AT definition (2011) National MS Society. Available from:
<http://www.nationalmssociety.org/chapters/vax/programs--services/programs/community-programs/assistive-technology/index.aspx>
[Accessed February 2010]

Azure (2011) Axure – Wireframes, Prototypes, Specifications. Available from:
<http://www.axure.com/> [Accessed February 2010]

Bevan, N. (1995) Usability is quality of use. In: Anzai & Ogawa (eds) Proceedings of the 6th International Conference on Human Computer Interaction, Yokohama, July 1995. Elsevier.

Beyer, H. Holtzblatt, K. (1998) Contextual design : A customer-centered approach to systems designs. Available from:
<http://portal.acm.org/beta/citation.cfm?id=291229&coll=ACM&dl=ACM&ret=1>
[Accessed June 2010]

Butler.S, Kindlund. E, Miller. D, Kirakowski. J. (1999) Comparative Evaluation of Usability Tests in Proceedings of UPA98 (Usability Professionals Association 1998 Conference) (Washington DC, June 1998), UPA, 189-200.

Card, S., Moran, T., & Newell, A. (1983) The psychology of human-computer interaction. Hillsdale, NJ: Lawrence Erlbaum Associates. ISBN: 0-89859-243-7

CEUD UD Principles (2011) NDA Centre For Excellence in Universal Design.

Available from: <http://www.universaldesign.ie/exploreampdiscover/the7principles>

[Accessed February 2010]

CFIT Website (2011) NCBI Centre For Inclusive Technology. Available from: <http://www.cfit.ie/user-testing> [Accessed February 2010]

Chapman, C.N, and Milham, R. P (2006) 'The persona's new clothes: methodological and practical arguments against a popular method' Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting, pp. 634 –636. Available from: <http://cnchapman.files.wordpress.com/2007/03/chapman-milham-personas-hfes2006-0139-0330.pdf>). [Accessed February 2010]

Clicker (2011) Cricksoft. Available from: <http://www.cricksoft.com/us/products/clicker/> [Accessed February 2010]

Cooper, Reimann (2003) About Face 2.0: The Essentials of Interaction Design. Wiley; 2nd edition (March 17, 2003). ISBN-10: 0764526413

Donald Howard, (2003). Swimming around the Waterfall: Introducing and Using Agile Development in a Data Centric, Traditional Software Engineering Company. Lecture Notes in Computer Science, 2003, Volume 2675/2003.

Dumas, J. (2002). Human Factors And Ergonomics archive. The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications. ISBN:0-8058-3838-4

EZKeys. (2011) Words-plus. Available from: <http://www.words-plus.com/website/products/soft/ezkeysexp.htm> [Accessed February 2010]

Fraternali. P. Rossi. G, Sánchez-Figueroa. F (2010) Rich Internet Applications - IEEE Internet Computing, 2010. Available from: <http://www.computer.org/portal/web/csdl/doi/10.1109/MIC.2010.76> [Accessed July 2010]

Gould, Lewis, (1985) Designing for Usability: Key Principles and What Designers Think. Communications of the ACM Volume 28 Issue 3, March 1985 pp. 300-311

Grid (2011) ZYGO Industries, Inc. Available from: <http://www.zygo-usa.com/> [Accessed February 2010]

Grosvenor, L. (1999). Software usability: Challenging the myths and Grosvenor, L. (1999). Software usability: Challenging the myths and assumptions in an emerging field. Unpublished master's thesis, University of Texas, Austin.

H. Rex Hartson, Terence S. Andre, and Robert C. Williges (2001) Criteria for Evaluating Usability Evaluation Methods. International Journal of Human-Computer Interaction, Volume 13, Issue 4 December 2001 , pages 373 - 410

Hartson, H. R., Castillo, J. C., Kelso, J., and Neale, W.C. (1996) Remote evaluation: The network as an extension of the laboratory. *Proceedings of CHI'96 Conference*, pp.228-235 ISBN:0-89791-777-4

Hartson, H.R, Andre, T.S. and Willages, R.C. (2003) Criteria for evaluating usability evaluation methods. International Journal of HCI. Available from: http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/h/Hartson:H=_Rex.html [Accessed March 2010]

Haungs (2001) Pair programming on the C3 project. Computer, vol. 34, no. 2, pp. 118-119. Available from: <http://www.computer.org/portal/web/csdl/doi/10.1109/2.901173> [Accessed February 2010]

HCI Glossary 1 (2001). Usabilitysa.co.za. Available from:
<http://www.usabilitysa.co.za/> [Accessed February 2010]

HCI Glossary 2 (2000) SQAtester.com. Available from:
<http://www.sqatester.com/glossary/index.htm#tz> [Accessed February 2010]

HCI Glossary 3 (2011) Don Norman's jnd.org. Available from:
<http://www.jnd.org/> [Accessed February 2010]

HCI Origins (1996) SIGCHI. Available from:
http://old.sigchi.org/cdg/cdg2.html#2_2_1 [Accessed Feb, 2010]

Henry, Shawn L. (2007) Just Ask: Integrating Accessibility Throughout Design.
Madison, WI: ET\Lawton. ISBN 978-1430319528

Hill, L, Carver, L et al. (2000) Alexandria digital library: user evaluation studies and
system design. Available from:
[http://onlinelibrary.wiley.com/doi/10.1002/\(SICI\)1097-4571\(2000\)51:3%3C246::AID-
ASI4%3E3.0.CO;2-6/abstract](http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(2000)51:3%3C246::AID-ASI4%3E3.0.CO;2-6/abstract)
[Accessed June, 2010]

ISO 9241 (2011) Ergonomics of human-system interaction -- Part 210: Human-centred
design for interactive systems. International Organisation for Standardization.
Available from: http://www.iso.org/iso/catalogue_detail.htm?csnumber=52075
[Accessed Feb, 2010]

Software Methodologies. (2010) Noop.nl. Available from:
[http://www.noop.nl/2008/07/the-definitive-list-of-software-development-
methodologies.html](http://www.noop.nl/2008/07/the-definitive-list-of-software-development-methodologies.html) [Accessed Feb, 2010]

ISO 13407 (1999) Human-centred design processes for interactive systems.
International Organisation for Standardization. Available from:
http://www.iso.org/iso/catalogue_detail.htm?csnumber=21197 [Accessed July 2010]

ISO TC 16071 (2008) Ergonomics of human-system interaction -- Guidance on accessibility for human-computer interfaces. International Organisation for Standardization. Available from:

http://www.iso.org/iso/catalogue_detail.htm?csnumber=30858 [Accessed July 2010]

Nielsen, J. Clemmensen, T. Yssing, C. (2002) Getting access to what goes on in peoples heads? – Reflection on the think-aloud technique. NordiCHI 2002 Available from: <http://portal.acm.org/citation.cfm?id=572033> [Accessed July 2010]

Chattratchart, J. Brodie, J. (2004) Applying Use Testing Data to UEM Performance Metrics. Available from: <http://portal.acm.org/citation.cfm?id=985921.986003> [Accessed June 2010]

Jacobsen, N. Hertzum, M. John, B. (1998) THE EVALUATOR EFFECT IN USABILITY STUDIES:PROBLEM DETECTION AND SEVERITY JUDGMENTS Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting (Chicago, October 5-9, 1998), pp. 1336-1340. HFES, Santa Monica, CA.

JAWS (2011) Freedom Scientific Website. Available from: http://www.freedomscientific.com/fs_products/JAWS_HQ.asp [Accessed February 2010]

O Connor, J. (2007). Joomla Accessibility, Packt Publishing ISBN: 1847194087

Krug, S. (2005), Don't Make Me Think: A common sense approach to the Web. ISBN-10: 0789723107

Faulkner, L. (2003) Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. University of Texas, Austin, Texas. Available from: <http://brm.psychonomic-journals.org/content/35/3/379.short> [Accessed February 2010]

Long, F. (2009) 'Real or Imaginary: The effectiveness of using personas in product design' *Irish Ergonomics Review, Proceedings of the IES Conference 2009, Dublin* ISSN 1649-2102

Long, J, Whitefield, A, (1987) *Cognitive Ergonomics: Cognitive ergonomics and human-computer interaction*. Cambridge University Press, 1989.

McInerney, P, Maurer, F, (2005) *UCD in Agile Projects: Dream Team or Odd Couple?* Available from: <http://portal.acm.org/citation.cfm?id=1096554.1096556> [Accessed February 2010]

Meister, D. (1999) *THE HISTORY OF HUMAN FACTORS AND ERGONOMICS* (LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS Mahwah, New Jersey London) ISBN 0-8058-2768-4

Molich, R. Bevan, N. & Curson, I. (2004) *Comparative usability evaluation*. Available from: <http://www.informaworld.com/index/713819776.pdf> [Accessed February 2010]

Nandini P. Nayak, Debbie Mrazek & David R. Smith (1994) *Analyzing and Communicating Usability Data. Now that You Have the Data What Do You Do? (CHI Workshop) CHI '94 Conference companion on Human factors in computing systems*. ISBN:0-89791-651-4

Nielsen, J, *Heuristic Evaluation*. J. Nielsen and R.L. Mack (Eds.), (1994) *Usability Inspection Methods*. CHI '94 Conference companion on Human factors in computing systems. ISBN:0-89791-651-4

Nielsen, J. (1993). *Usability engineering*. Boston: AP Professional. ISBN 0-12-518405-0

Nielsen, J. (2000, March). *Why you only need to test with 5 users: Alertbox*. Retrieved April 15, 2003 from <http://www.useit.com/alertbox/20000319.html> Perfetti, C., & Landesman. [Accessed February 2010]

Norman, D. (1998 – republished frequently) The Design of Everyday things. ISBN-13: 978-0-262-64037-4

NVDA (2011) NVDA Project. Available from: <http://www.nvda-project.org/> [Accessed February 2010]

ORCA (2011) Orca –GNOME Live! Available from: <http://live.gnome.org/Orca> [Accessed February 2010]

PDF (2011) Webaim:PDF Accessibility. Available from: <http://www.webaim.org/techniques/acrobat/> [Accessed February 2010]

Principles of UD. (2008) Centre for Universal Design NCSU. http://www.design.ncsu.edu/cud/about_ud/udprincipleshtmlformat.html [Accessed February 2010]

Riemann, J. Franzke, M. Redmiles, D. (1995) Usability Evaluation with the Cognitive Walkthrough. <http://portal.acm.org/citation.cfm?id=223735> [Accessed February 2010]

Rubin, J. (1994) Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests (Wiley Technical Communications Library 1994, Second edition 2008)

Sanders, M. McCormick, E. (2002) Human factors in engineering and design. Serena. <http://www.serena.com/products/prototype-composer/> [Accessed February 2010]

Spool, J. & Schroeder, W. (2001). Testing web sites: Five users is nowhere near enough. In CHI 2001 Extended Abstracts (pp. 285-286). New York: ACM Press. [Accessed January 2010]

SuperNova (2011) Dolphin Website. Available from: <http://www.yourdolphin.com/productdetail.asp?id=1> [Accessed February 2010]

Tullis, T. Fleischman, S. McNulty, M. Cianchette, C. and Bergel, M. (2002) An Empirical Comparison of Lab and Remote Usability Testing of Web Sites Available from: <http://home.comcast.net/~tomtullis/publications/RemoteVsLab.pdf> [Accessed January 2010]

Usability First, definition of waterfall. (2010) Usability First Glossary. Available from: http://www.usabilityfirst.com/glossary/term_718.txt [Accessed January 2010]

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? Human Factors - Special issue: measurement in human factors archive Volume 34 Issue 4, Aug. 1992 Human Factors & Ergonomics Society, Inc. Santa Monica, CA, USA Available from: <http://portal.acm.org/citation.cfm?id=141700> [Accessed February 2010]

VoiceOver (2011) Apple website. Available from: <http://www.apple.com/accessibility/voiceover/> [Accessed February 2010]

W3C Business Case Examples (2010) W3C/WAI. Available from: <http://www.w3.org/WAI/bcase/resources.html> [Accessed July 2010]

WAI Business Case for Accessibility (2010) W3C/WAI. Available from: <http://www.w3.org/WAI/bcase/Overview> [Accessed February 2010]

WAI Introduction to Accessibility (2005) W3C/WAI. Available from: <http://www.w3.org/WAI/intro/accessibility.php> [Accessed February 2010]

WAI Introduction to WCAG (2008) W3C/WAI. Available from: <http://www.w3.org/WAI/intro/wcag.php> [Accessed February 2010]

WinEyes (2010) GW Micro – Window-Eyes. Available from: <http://www.gwmicro.com/Window-Eyes/> [Accessed February 2010]

Woolrych, A., & Cockton, G. (2001). Why and when five test users aren't enough. In J. Vanderdonckt, A. Blandford, & A. Derycke (Eds.), Proceedings of IHM-HCI 2001 Conference: Vol. 2 (pp. 105-108). Toulouse, France: Cépadèus.

Word Accessibility Techniques (2010) Webaim website. Available from: <http://www.webaim.org/techniques/word/>. [Accessed February 2010]

Yeung (1995). Qualitative Personal Interviews in International Business Research: Some lessons from a study of Hong Kong transnational corporations. <http://courses.nus.edu.sg/course/geoywc/publication/IBR.pdf> [Accessed Jan 2011]

ZoomText (2010) AISquared. Available from: <http://www.aisquared.com/index.cfm> [Accessed February 2010]

APPENDIX

User Testing Research Questions

This survey evaluates the practices of the professional usability community when undertaking usability tests of web sites and Rich Internet Applications (RIAs) by examining the current user-testing methodologies involved in user testing with older people and people with disabilities.

By interviewing usability practitioners, comparing their practices and methodologies, the research aims to evaluate the quality of their respective outputs and analyze how the results of user testing can be best incorporated into web development projects.

No participant in this survey will be named or identified in anyway, all answers will be aggregated and anonymised

Section A: Background Information

1. Please give your job title:
2. Please describe your role?
3. Are you doing user testing? If so please describe
4. Do you use other usability methods in your projects? E.g Case studies. Focus groups. If so please outline..
5. Please list any other aspects of your job that you think relevant to improving the user experience/accessibility/usability of a project (for example if you undertake accessibility auditing etc.
6. If you do accessibility auditing please list any relevant standards and/or guidelines that are most relevant to your work?

7. Do you feel that there are aspects of your previous experience that you have that help you to be a competent usability professional? Such as working with people with disabilities in some capacity volunteer work etc.

8. Continuing on from the previous question, would you identify your skill set as being most closely aligned to some previous role. For example, would you consider yourself primarily a web designer/developer/ etc?

If so, please outline

9. Do you have any qualifications directly relating to usability analysis?

If so please give details

10. Are there other qualifications that you feel help in your role?

11. Please give a short overview of your educational background (Web Design, Multimedia, Computer Science etc)

12. Whether you have a directly related qualification or not. Do you consider yourself to be qualified to conduct user testing? If so please outline why.

13. Who would be the most influential thinkers/practitioners in the field of user testing/usability analysis in your opinion and why?

14. Does your experience of interacting with people with disabilities primarily come from the user testing?

15. And one more question in the section, approximately how many user tests have you performed in your career?

[] 0 - 10

- ☐ 10 - 30
- ☐ 30 - 50
- ☐ 50 - 100
- ☐ 100 +

16. If there is an other background information you feel is relevant, please feel free to add it here, thanks.

Section B: Assistive Technology

1. I'm going to ask the same question in a few different ways, answer any version of it - What do you think AT is? What does the term AT mean to you? If you describe AT to a friend what would you say? Please tell the story of AT briefly in your words.

2. What assistive technologies would you describe yourself as being familiar with or aware of? Please list any that you can think of.

3. How would rate your level of knowledge regarding the operation of the AT that you are most familiar with (from the technical perspective) such as how screen readers work etc?

☐ Advanced

☐ Working knowledge

☐ Basic

☐ Understanding

☐ Weak

4. Do you ever use a screen reader or screen magnification app in your own testing or auditing of websites/applications? If so please list.

5. Do you have confidence in the results that you get from the testing that you do with your chosen AT?

6. If there is any other assistive technology information you feel is relevant, please feel free to add it here, thanks.

Section C: User Testing Methods

1. Are you aware of any existing user testing methodologies?

If so please outline.

2. Do you use any of these user testing methodologies when conducting your work?

3. If so, please describe how these methodologies are (or are not) relevant to the work that you do.

4. If you do not adhere to a particular methodology, please outline why? Alternatively, have you created your own methodology that works within the context of your role? If so please outline how you work.

5. If there are another user testing information you feel is relevant, please feel free to add it here, thanks.

Section D: User Testing in Practice

1. Do you run or are you involved with a usability lab?

If so, please briefly describe the lab and its equipment..

2, Do you use video to record tests?

3. Do you use screen capture such as Morae?

4. Do you use any of the advanced data analysis features in Morae (or a similar package), if so please outline.

5. Do you provide your clients with video footage after a test?

If so, do you prefer to give edited footage, footage with analysis and comments, or the raw test footage?
Please outline.

6. Do you provide a written report after a test for a client?

If so what level of granularity do you provide (overview, details of user comments/actions etc?)

7. How do you deliver the report?

8. Do you use eye tracking software, If so please outline?

9. Do you personally user test people with disabilities?

10. Do you personally user test people with older people?

11. Please state if you test with people from any or all of the following user groups:

- ☐ Blind user
- ☐ Visually Impaired Persons (VIP)
- ☐ Cognitive/Intellectual Disabilities
- ☐ Physical Disabilities
- ☐ Combinations
- ☐ Other

If *Other* or *Combination*, please outline

12. How large are your average user tests?

- ☐ 1
- ☐ 2-4
- ☐ 4-8
- ☐ 8-12
- ☐ 12 +

13. Do you feel that these sample sizes are sufficient?

Please outline your views on the importance of the number of users involved in the tests.

14. When drafting test scripts, do you involve your client? If so how?

15. Have you ever undertaken user testing more than once in the same project?

If so, please describe.

16. Was it beneficial, if so how?

17. Did it reinforce your initial findings or contradict them in any way, or did it shed fresh light?

18. If there is any other practical user testing information you feel is relevant, please feel free to add it here, thanks.

Section E: Outcomes of User Testing

1. What do you feel the main benefits of user testing are?

2. Are the results of user testing incorporated into projects?

If so, how?

3. Would you feel that an outline of user testing is an effective way of improving the quality of the overall project?

4. What, in your opinion, are the main deficiencies with user testing?

5. Are there aspects of how you undertake user testing that you would like to improve?

Is so, how would you do this?

6. Are you involved with the wider usability community?

If so, how?

7. Do you contribute to accessibility/usability mailing lists? If so, which ones?

8. Have you had the opportunity as a part of your work to undertake any research?

9. Are there other ways that you give feedback of any interesting observations to the wider community?

If so, how?

10. If there is any other practical user testing outcomes information you feel is relevant, please feel free to add it here, thanks.

O.K., that's it, thanks for your help again, and as I said above your answers will be combined with all the others for my research. Neither I, or any institute I represent, nor any other third party will record you name, email address or any other personal details, nor will it be possible to identify you in any way from the report I will publish as part of my MSc dissertation. I would also like to thank you again for taking the time to fill in this questionnaire.