Investigating the Relationship between Usability, Preferences and Usage Intentions when Banking Online

Catherine Sarah Weir

A thesis submitted for the

Degree of Doctor of Philosophy

The University of Edinburgh

2008

Declaration of originality

22nd April 2008

This thesis is submitted for the Degree of Doctor of Philosophy. I declare that it has been composed by myself and that the work described was my own research.

Catherine S. Weir

Acknowledgements

Many of my colleagues, past and present, at the Centre for Communication Interface Research (CCIR) have provided valuable contributions to the work presented in this thesis. In particular, Dr. John Foster and my supervisor Professor Mervyn Jack have offered their valuable experience throughout. Dr. Fergus McInnes has also been a tremendous source of wisdom and experience regarding experiment design, instrument development and statistical analysis.

I would like to thank Lloyds TSB bank for supporting this research and for permitting the use of user-interface designs based on their real services. The user-interface design manipulations and extensions used in the experiments described here are my own work. I was very fortunate to receive excellent technical assistance developing prototype interfaces and in this regard I would particularly like to acknowledge Ian Taylor, Dr. Iain McKay, Dr. Gary Douglas and Dr. James Anderson.

It was the enthusiastic lectures of Dr. Patrick O'Brien Holt at Herriot Watt University that inspired me to pursue research in Usability and Human Factors Engineering. The publications of Jared Spool and his associates at User Interface Engineering were particularly influential to my early work. My research was also greatly motivated by the writings of Andrew Dillon, Kasper Hornbæk and Brian Shackel.

I would like to thank my parents for their support and enthusiasm for my pursuit of academic research. Finally to my partner Donald, thank you for your encouragement, support and assistance with proofreading.

Abstract

It is widely assumed that the success of a computer system is related to its usability, yet there is little data to support this supposition. The concept of usability as it may apply to system success is reviewed. A modern, broad definition of interface usability is proposed comprising utility, attitude and performance factors in relation to specific users, tasks and environments. Appropriate usability metrics are devised to compare interface designs in controlled experiments. The experiments are conducted in the context of Banking and the Internet.

A clear and consistent relationship between attitudes toward usability and preference quality ratings for interfaces is demonstrated and extends to usage intentions in the eBanking context. Further, the relationship between preferences, attitudes and performance measures are explored and findings support previous research: that preference does not always follow performance, and that attitudes and performance (subjective and objective measures of usability) are not always directly related. Levels of utility in the Banking interface are also explored. Finally, interface characteristics highly associated with preferences and usage intentions for Internet Banking services are highlighted.

List of Publications

Weir, C. S., Anderson, J. A. and Jack, M.A., 2006, On the role of metaphor and language in design of third party payments in eBanking: usability and quality, International Journal of Human-Computer Studies, Volume 64, Issue 8, pp. 770-784.

Weir, C. S., McKay, I.G. and Jack, M.A., 2007, Functionality and usability in design for eStatements in eBanking services, Interacting with Computers, Volume 19, Issue 2, pp. 241-256.

Weir, C. S., The Relationship between Usability Metrics and Usage Intention in Online Banking Services, manuscript prepared for submission to the International Journal of Human-Computer Studies (September 2008).

List of Abbreviations

ATM	–	Automated Teller Machine
ATS	–	Automated Telephone System
BI	–	Behavioural Intention
CI	–	Confidence Interval
df	–	Degrees of Freedom
HCI	–	Human Computer Interaction
HF	–	Human Factors
IA	–	Information Architecture
IT	–	Information Technology
IS	–	Information Systems
MIS	–	Management of Information Systems
NS	–	Not Significant
P3	–	P3 model (p.19) of usability and usage (Dillon & Morris, 1998)
P3P	–	Payments to Third Parties (or Third Party Payments)
PEOU	–	Perceived Ease of Use
PL	–	Plain Language
PU	–	Perceived Usefulness
TAM	–	Technology Acceptance Model
TRA	–	Theory of Reasoned Action
TTF	–	Task Technology Fit
UE	–	Usability Engineering
UEM	–	Usability Engineering Methods

Glossary

Attitude – A person's perception of their feeling or opinion towards something; their subjective evaluation of an experience.

Clicksteam – Clicksteam is defined as a list or path of mouse clicks taken through the site contents in the performance of a task or tasks.

IA – As applied to the Internet, Information Architecture (IA) is defined as the organisation of the information within the Website: simply put it concerns the navigation, structure, classification, labelling and provision of hyperlinks to help people find, manage information and provide intuitive access to content (adapted from Rosenfeld & Morville, 2002).

Performance – The observable or measurable behaviour of a person in a particular situation; the ability to carry out a specific function or functions.

Reliability – A statistical term concerned with the consistency within a measurement set, such as the precision of the instrument in a test-retest situation, or the consistency of the elements within the measure.

UE – Usability Engineering (UE) is a process which informs and evaluates the design and engineering of usable and useful systems, to ensure quality in use for the intended user and therefore success of the system in question.

Utility – The capability of a computer system; its fitness for some desirable purpose or valuable end; its functionality or task-technology fit.

Web Portal – A Web Portal is defined as a Website which provides an entry point (or gateway) to a range of information. In this case a Website which allows access to digital information on the range of products and services provided by the Case Bank and any subsidiary companies.

List of Figures

List of Tables

Table of Contents

Chapter 1. Introduction

The thesis expounded here is that attitude measurement of usability which includes components of interaction, fulfilment, visual design and integrity relate to preference formation and usage intention in performing online information retrieval and eBanking tasks.

When designing technology it has long been recognised that human factors need to be taken into consideration (Shackel, 1991). Usable designs tend to consider the user perspective but modern industry is still dominated by technology-driven rather than user-centred products and services. The consequences of this focus are present in everyday things (Norman, 1988), for example in user frustration with mobile phones, confusing remote controls and complicated ticket machines. Systematic analysis of the interaction between users and

technology, known as usability engineering (UE or just 'usability') can foster better design by balancing what is technologically possible against human needs, expectations and capabilities. When developing a new system, resources invested in usability come from limited development budgets (Chapanis, 1991), and these expenditures need to be justified. Usability experts argue that good design results in satisfied users. In the context of service provision this means loyal service users. However, empirical evidence supporting the relationship between usability and system success has not been widely published. This thesis will investigate whether usability in design really leads to success in the marketplace, whether subjective or objective measures of usability are more appropriate in relation to success and which characteristics of an interface design are most highly associated with potential success.

It is difficult to predict how different people will respond in their interaction with different interface designs. This is where usability engineering methods (UEM) are typically employed. Usability work can predict an appropriate design for maximum benefit and make sure this design is optimised for the potential users and tasks (Tohidi et al, 2006). Usability engineering involves collections of methods applied throughout the design process, typically detailed questions and observation posed after direct use of prototype interfaces are used to make iterative enhancements to the designs (Nielsen, 1993a).

In addition to the growing collection of usability research findings, there are also a variety of adoption models which aim to predict the real world usage of a specific technology. Adoption models, such as the widely cited Technology Acceptance Model (TAM), use attitudes to constructs such as the ease of use of the technology or interface and its potential usefulness (often in the context of a job) to forecast probable real-world use (Davis et al, 1989). Typically, adoption models do not ask the types of detailed questions common to a usability engineering approach and are not concerned with iterative design enhancements (Weir et al, 2007).

There have been relatively few attempts to provide both the feedback on interface and interaction usability and quantitative statements of potential usage or success. While usability metrics focus on many different aspects of designs and consider user performance and attitude, they typically refrain from demonstrating their relationship to real world usage or adoption of one service or another. Often, this is due to the early prototype and pre-prototype stages at which many usability evaluations take place. Human Computer Interaction (HCI) work has typically associated better usability in designs to more

acceptability and adoption in the marketplace, yet these relationships have not been resolved (Dillon & Morris, 1996).

Meanwhile studies using the TAM often fall short of collecting empirical usage data. In fact, in many cases TAM researchers tend to regard behavioural intention to use as a substitute for real-world usage. There are several criticisms to this approach (Legris et al, 2003; Horton et al, 2001; Szajna, 1996) although generally, data supports the application of the TAM and it's extensions in many fields of usage (Premkumar & Bhattacherjee, 2008; King & He, 2006; Taylor & Todd, 1995; Mathieson, 1991).

The research presented here builds upon two previous attempts to unify the usability and adoption concepts. The first is a model of usability and technology acceptance called the P3 model which was proposed to bridge the gap between Information Systems (IS) and HCI approaches to understanding usage and success (Dillon and Morris, 1998). To date, empirical validation of the P3 model has not been published. More recently, Microsoft's usability guidelines were used to obtain usability feedback in the context of intention to use (Agarwal and Venkatesh, 2002) indicating practical benefits from a usability evaluation approach. However, the study did not examine the usability-behaviour relationship.

Recent challenges in the usability field also question the plethora of usability evaluation techniques and the variety of metrics collected. In order to answer these challenges, associations between the various subjective and objective usability measures must be studied and their relative importance determined (Hornbæk, 2006). In this work, the relationships between different usability metrics are studied in conjunction with the usability-preference and usability-adoption relationships. Therefore salient usability and interface design characteristics which appear to be important in relation to preferences and intentions can be illuminated. This work sets out to outline a methodology which can iteratively improve designs whilst predicting acceptability and explaining the important usability and utility characteristics which drive potential adoption. Finally, it will discuss the use of subjective and objective measures of usability and their associations with preference and adoption intentions.

This work considers the discretionary use of the Internet for banking needs. The development of the Internet has brought challenges and opportunities to many industries including banks. The UK has a large financial services industry and most UK banks now offer Websites and Internet Banking (eBanking) services. Public access Websites allow for the marketing and delivery of product information to a range of customers and non-

customers. eBanking services offer applications for personal account and transaction management. Typically of the Web medium, many sites and applications have been developed using guidelines, principles and opinions. Often these guidelines are inferred from studies in quite different disciplines, applications and environments (LeCompte, 1999). When it comes to understanding what makes a good website, some industries still have "a long way to go" (Spool et al, 1997). There is very little published data specific to the design of a successful eBanking service.

Banks are also pledged to adhere to the Banking Code (British Bankers' Association, 2005) which requires them to provide "secure and reliable banking and payment systems" and act "fairly and reasonably" in dealings with their customers. The code applies across banking channels. The eBanking channel is a self-service transaction environment. People with a wide range of abilities perform the tasks traditionally done by trained employees (with considerable financial benefits to the companies involved). In self-service, users rely on the interface to provide any guidance that they need to perform banking tasks. Intuitive, error tolerant, usable designs are important in this context to ensure the reliability of transactions performed.

Therefore Banking Websites and in particular eBanking service Websites are a good candidate for usability engineering work. In this context the systematic, iterative design and development offered by a user-centred process can ensure the provision of the right design for the application, and to refine that design into a high-quality service (Tohidi, 2006). Therefore public-accessed informational Websites and personal account-management applications (Internet Banking or eBanking) were selected as the platform for a series of usability experiments. The experiments aimed at producing usable and useful eBanking interfaces and to explore the relationships between subjective and objective usability metrics, preferences and usage intentions. The experiments and evaluations were performed in the context of strategic planning for the Website and eBanking channel of the Case Bank[1]. The work aimed to enhance usability in the user-interfaces of these Web offerings. By incorporating usability practices, the service provider hoped to encourage migration from staffed to the lower-cost Internet channel. From an academic point of view, the evaluations were designed and conducted in order to provide data to explore relationships between usability aspects, preferences and intentions to use banking Websites. Finally, to validate a

[1] Lloyds TSB Bank plc.

usability questionnaire suitable for the subjective evaluation of Websites and eBanking services.

1.1. Thesis

The thesis investigated here is that usability metrics will relate to usage intentions, that both subjective and objective measures of usability as evaluated through usability engineering techniques will relate to preferences and usage intentions. This work attempts to empirically validate the relationship and thus confirm the potential for usability measures to inform the potential acceptability, usage and therefore success of a Web service. The confirmation of this relationship will establish the benefit of including usability engineering work in the development budget.

The model of usability and intention is adapted from the P3 model (Dillon and Morris, 1998) that usage intentions are positively associated with the degree of functionality (utility), the ability to operate (performance) and the subjective perceptions of the interaction (attitude).

The work further examines the structure of a Web usability attitude questionnaire for eBanking interfaces and suggests that subjective evaluations of usability which include interaction, fulfilment, visual design and integrity attributes are highly correlated with preferences and usage intentions. The relationship between objective measures and the subjective usability questionnaire will be considered for each specific interface and associated tasks.

A usability-preference (quality) relationship is established in the context of public-accessible Banking information Websites and authenticated eBanking transactions. Further, a usability-usage intention relationship is established for adopting electronic instead of paper statements in an eBanking interface. Actual adoption rates are compared to estimates based on early prototypes and show the predictive value of the intention metric.

1.2. Terms

The thesis is work in the multidisciplinary field of Human Computer Interaction (HCI). Within this field, Website usability, usability engineering methods and metrics and customer preferences are of specific interest. It also considers related work from outside this field, specifically in terms of adoption of technology. In this respect it draws on research from

Information Systems (IS) which is concerned with managing Information Technology (IT) in organisations, IT development, implementation and use. IS offers a different perspective on technology adoption than the HCI field.

The experiments are all conducted within a context of designing and evaluating Websites and Web applications for the Case Bank.

Participants in the studies were all 'computer savvy' – that is people who were able to use a computer when asked during recruitment; or 'Internet-savvy' – that is people who, at recruitment, had used the Internet within the last month.

A range of other terms in the fields of usability engineering and adoption of technology will be defined and reviewed in Chapter 2 and a glossary is included on page

1.3. Contribution

The main contribution to knowledge arising from these studies is the quantification of the relationship between subjective usability metrics and preferences or quality ratings for interface designs and an extension of that relationship to usage intentions. These relationships are presented in the context of Banking Websites (see Chapters 4, 5 and 6)

The thesis also contributes a usability engineering approach and a validated attitude questionnaire (see Chapters 3 and 5) adapted from well-researched methods to study the design and refinement of Web interfaces. The method and metrics are then extended to consider acceptability and use of an end-product (see Chapter 6).

A novel preference metric is proposed offering both information on overall quality levels perceived and preference data for one design or another (see Chapter 3). The relationship between usability and preference (or quality) is demonstrated for a range of banking tasks on the Internet. Further, the most important aspects of interface usability which predict and explain preferences are identified (see Chapters 4 and 5).

The preference metric is also adapted into a measure aimed at describing usage intention, a concept associated with many adoption studies. The relationships between usability and usage intention in the context of eBanking services are established. The most important aspects of usability which predict and explain usage intentions are also highlighted (see Chapter 6).

Actual usage data from subsequent live service offerings is then used to determine the extent to which the gathered intentions predict real-world usage. This provides evidence that the

intention metric offers robust quantitative guides which are valuable in predicting success from fairly early prototypes. The conservative intention estimates predicted 13% adoption from the results of the experiment, when the first real-world adoption numbers were calculated at 23%. The intention tendencies measured in the experiment at some 54% of the cohort were higher than these initial real-world usage measurements (see Section 6.6.8, p.234).

Finally, the relationship between objective (performance measures) and subjective (attitude) usability metrics is discussed. The relationship was more dependent on the specific interfaces and tasks being evaluated, and no consistent results were found (see Chapters 4, 5, 6 and 7).

1.4. Outline

Chapter 1 outlines the scope and context of the work and proposes the thesis to be tested and the contribution to knowledge arising from the work.

Chapter 2 contains a review of usability engineering literature and scope, experiment design and analysis. It introduces several models and theories related to the adoption of technology. Finally, eBanking, the banking industry and the role of Internet services in this industry are discussed.

Chapter 3 presents the development of a Web usability questionnaire for use in general financial information retrieval and to be adapted for account-specific eBanking tasks.

Chapter 4 presents the pilot studies of the Web questionnaire and usability metrics. The pilots consider general Portal design and the reduction of interface clutter using three designs for a Banking Portal of product and service information. The chapter will discuss the design, methodology and results. The usability questionnaire will be examined in detail, along with the comparative preference metric, objective measures and the use of the verbal protocol. The relationship between usability and preference will be examined and discussed. The structure and scope of the usability questionnaire will be presented for use in the subsequent evaluations.

Chapter 5 presents experiments on interface metaphor and dialogue design. These relate to usability and preference evaluations for making transactions in self-service eBanking applications. The chapter will discuss the design, methodology and results of the two experiments. The relationship between usability and overall quality (preference) ratings will

be examined and discussed with a view to using usability metrics and user-interface design to maximise actual usage.

Chapter 6 presents the eStatements experiment investigating levels of utility in the design of new functionality. This evaluation focuses on maximising usability and therefore encouraging migration of simple statement tasks to the eBanking channel. The design and methodology will be presented. In this chapter, the usability metrics will be compared to measures of usage intention. The proposed model of usability and intentions to use the service in real life will then be validated. The relationship between usability metrics and usage intentions will be discussed. Usage data from a subsequent live service will be introduced to allow comparisons with predicted intentions to actual usage data to be made.

Chapter 7 discusses and compares the relationships between usability attitudes, performance, utility, preference and intentions to use across the various Online Banking experiments presented. The conclusions of the work will then be presented and opportunities for further work will be highlighted.

References: The thesis concludes with a full reference list.

Appendices consist of statistical equations; experiment materials, tasks and interview questions for each experiment; summaries of eBanking design guidelines resulting from each individual experiment; additional data tables and charts.

Chapter 2. Background and Methodology

This chapter will introduce readers to the field of user interface design and usability research. Definitions of usability and methods of incorporating usability into the design process will be reviewed. The chapter will also introduce the concepts of system success, such as adoption and acceptance of information technology and describe models to predict real world use. From this foundation, an experimental evaluation method for Web services will be described. The exposition will include relevant statistical techniques, their limitations and how conclusions can be drawn from their results. Finally, the specific context of Banking services and the Internet as the platform for the usability investigations will be described.

2.1. Introduction to Usability

"Engineers make things that are useful to people. In collaboration with designers, ergonomists make things that are usable by people. The concept of usability means making artefacts easy, efficient and comfortable to use (anything from a corkscrew to a control room in a nuclear power station). Most people have experience of poorly designed objects. At best they cause frustration and annoyance (for example when a video recorder fails to record your favourite programme). At worst they can lead to injury or even death (as in the release of radioactive material from a nuclear reactor)."

(Stanton & Young, 1999).

In order to measure a quality, demonstrable criterion must be specified; usability needs to be defined. Usability is defined by the ISO as:

"the efficiency, effectiveness and satisfaction with which specified users can achieve specified goals in particular environments."

(ISO 9241-11, 1998).

Many alternative definitions exist but all serve to reinforce the idea that there are several aspects of system design that will influence interaction (Gould, 1988). Early definitions and measures of usability focused on effectiveness, learnability, flexibility and attitude (Shackel, 1986). In this definition, effectiveness related to speed and performance, and attitudes related to human costs, such as tiredness, discomfort, frustration and effort. Most definitions stress that systems should be intuitive, memorable, assist in the prevention of errors and in error recovery (Nielsen, 1993a). Another important consideration is how people feel about using the system (Quesenbery, 2002). Usability definitions are also concerned with what makes a product successful (Kuniavsky, 2003), from the end-users point of view this is suggested to be:

- Functionality (performing useful functions to those using it)
- Efficiency (how quickly and error tolerantly the functions can be operated)
- Desirability (the emotional response and satisfaction of using well-designed functions).

This leads to the more modern concept of usability, that of the whole user experience – the idea that task fit and ability to perform tasks is insufficient in explaining the choices people

make and therefore determining what will make a successful product or service. Crucially, definitions imply both a subjective and an objective component is involved in usability. The focus of usability is often on the user interface as this is where the user directly manipulates and experiences the process (Shneiderman & Plaisant, 2005; Dix et al, 2004). Functions can be offered by way of many contrasting interface designs, but equivalent functionality will not necessarily correspond to equal usability. Designing and engineering the user experience to ensure successful products has become central to modern development work, and is commonly known as Usability Engineering.

2.1.1. Usability Engineering

Usability engineering methods have been developing for over 20 years as an applied engineering discipline taking contributions from Human Factors and Ergonomics. Experimental methods of usability engineering borrow extensively from the formal techniques of experimental psychologists (Hartson, 1998; Preece et al, 1994; Karat, 1988). Human factors engineers were first involved with new technologies such as factory automation, military applications and industrial control systems. From early Time and Motion studies of factory workers, through to the design of dials, cockpits and telephone keypads, the human factor and our propensity to make errors have always been important considerations (O'Brien Holt, 2006). In recent years more and more people came into contact with computers and automated systems as a part of their daily lives: for example, around 13.9 million households in Great Britain had Internet access at home according to the Office of National Statistics in August 2006 (National Statistics, 2006). Modern computer systems serve a larger, more diverse user base than ever before. Thus human factors and usability engineering have developed techniques to assist the design and evaluation of a multitude of different applications (Shackel, 2000). Televisions, remote controls, telephones, ticket machines and the Internet are being used by a wide range of people from a variety of backgrounds – in both voluntary (e.g. for leisure activities, in self-service applications etc.) and compulsory (e.g. workplace) settings. The Internet, for example, offers a huge variety of applications and Internet users carry out a wide range of tasks online, with knowledge and experience levels that may differ radically from person to person. Attention has to be paid to this diversity when designing new Websites, and inevitably this means that large groups of users need to be studied in order to evaluate designs sufficiently (Hartson, 1998).

There are two main uses of the term Usability Engineering. The name was originally used to describe a process for specifying (in advance) the usability of a finished system, then attempting to demonstrate that these specifications had been achieved through iterative user testing and system re-engineering. The aim was to ensure that a finished product was suitable for its users and fitted the purpose for which it was created. Specifications were quantitative, such as the time to learn certain functions or the ability to perform certain tasks within a specified time limit. Later, a broader use of the term emerged, viewing usability engineering as a "discipline aimed at enhancing the usability of products" (Nielsen, 1993a). In this definition, usability engineering is used to assess usability by emphasising a cyclic process of evaluation and redesign based on user observations.

This thesis considers Usability Engineering as a process which informs and evaluates the design and engineering of usable and useful systems, to ensure quality in use for the intended user and therefore success. It is a process rooted in traditional engineering disciplines (Faulkner, 2000) providing techniques to support resource management in system design and development (Whiteside et al, 1988). The aim is to design and engineer the best solution for an individual system by centring the process on the user and their task (Nielsen, 1993a). Direct experience is key to the process (Karat, 1988). Questions about interface artefacts, from navigation design to visual characteristics, are posed. Weaknesses are identified to provide feedback for the development of the production interface (Hartson, 1998). The evaluation process aims to predict and explain consumer attitudes and behaviours (Howell, 1985). The usability engineering process sets out to find the optimal balance between various competing metrics such as efficiency and subjective satisfaction (Landauer, 1988). As such trade-offs are made dependent upon the type of system being built since "a repetitive data-entry system, an air-traffic control system or a game" all require a different balance (Shneiderman, 2002). The implication of usability by definition is that it requires different considerations for the design of different artefacts (Stanton & Young, 1999).

Figure 2.1 describes a typical usability engineering lifecycle with early focus on users and tasks and iterative steps in design and development as well as in offering upgrades to the finished product.

There has been much usability engineering research, method, theory and practical application published in recent years in the field of HCI, and many leading texts in the field review this work (for example: Shneiderman & Plaisant, 2005; Dix et al, 2004; Faulkner, 2000; Baeker et al, 1995; Preece et al, 1994; Nielsen, 1993a; Helander et al, 1988). Usability

engineering is now an established practice in software development and has also been used effectively to inform the design of Web sites (i.e. Nielsen, 2000; Spool et al., 1997) and eBanking applications (Weir et al, 2006; Weir et al, 2007).

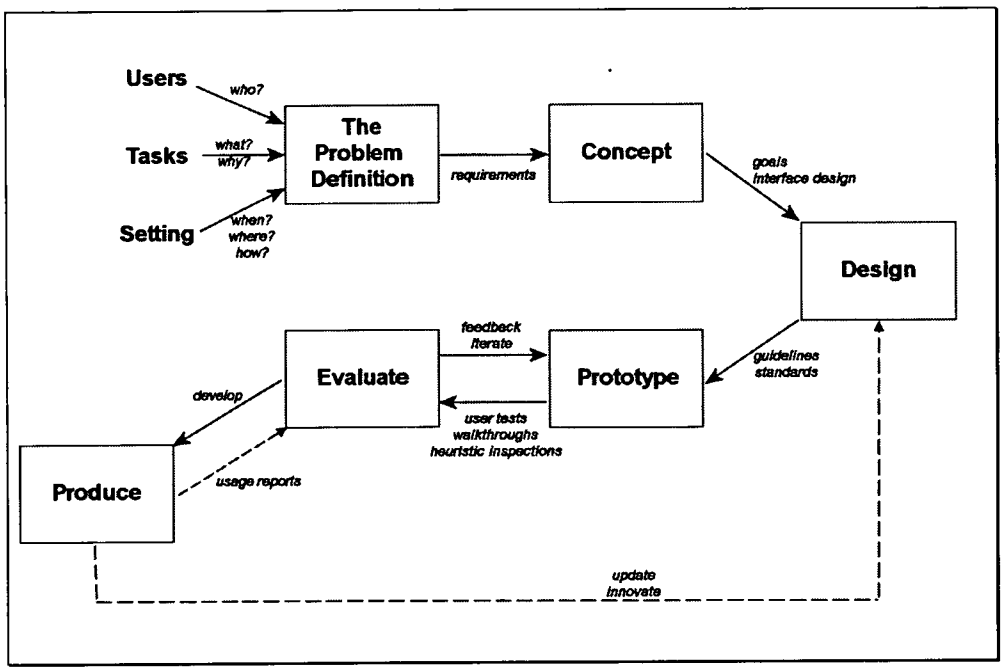


Figure 2.1. Usability Engineering Lifecycle

2.1.2. Web Usability

Web usability aims to ensure efficient, effective and satisfying experiences, but is specifically interested in use of the Internet. In particular, it has been noted that satisfied users may spend longer at a Web site, may revisit the Web site later, and may recommend it to others, issues particularly relevant to eCommerce and eBanking (Zhang & von Dran, 2000). For self-service banking, efficiency is a key driver of users to the online service (Dandapani, 2004; Wright, 2002). eBanking interfaces should ideally be quick and easy to use to meet the needs of eBankers (customers). Additionally, an eBanking interface must be effective, carrying out the desired transactions in a way that avoids errors, which may then be costly to resolve (Gopalakrisnan et al, 2003), or worse, go unnoticed by the user (Weir et al, 2006). Finally, subjective satisfaction and desire to use the channel over other contact channels will be crucial to the success of any Web application such as eBanking. Eventual adoption of the lower cost channel being the focus of the service provider.

A further consideration of Web usability is the diverse population of users which can access content and services over the Internet. Web usability attempts to make the Internet accessible and usable to a wide range of these potential users, although studies of particular groups are also useful in some situations (i.e. visually impaired users, the aging population for accessibility issues and the different Web browsing technologies they use such as screen-readers or magnifiers). Generally, the aim is to optimise usability in design to suit the main body of potential users. However, including demographic factors allows the identification of specific groups and individuals who react differently to the majority, and who may benefit from further study or a tailored design.

2.1.3. Adoption of Technology

From the HCI viewpoint, usability concerns making an interface design suitable for the majority of potential users. Usability experts consider that the main barriers to technology adoption are difficulty, complexity and frustration in use of the technology (Ceaparu et al, 2004). By evaluating and iterating towards a usable interface, providing a good user experience in a broad sense, and providing useful functionality, the usability barrier can be overcome. This enables use of the interface in real life. A taxonomy of the contributions to system acceptability (or success) is suggested in Nielsen (1993a). Usability and utility combine to explain system usefulness, a practical aspect of acceptability, cost is also considered and a social component to acceptability is suggested, as shown in Figure 2.2.

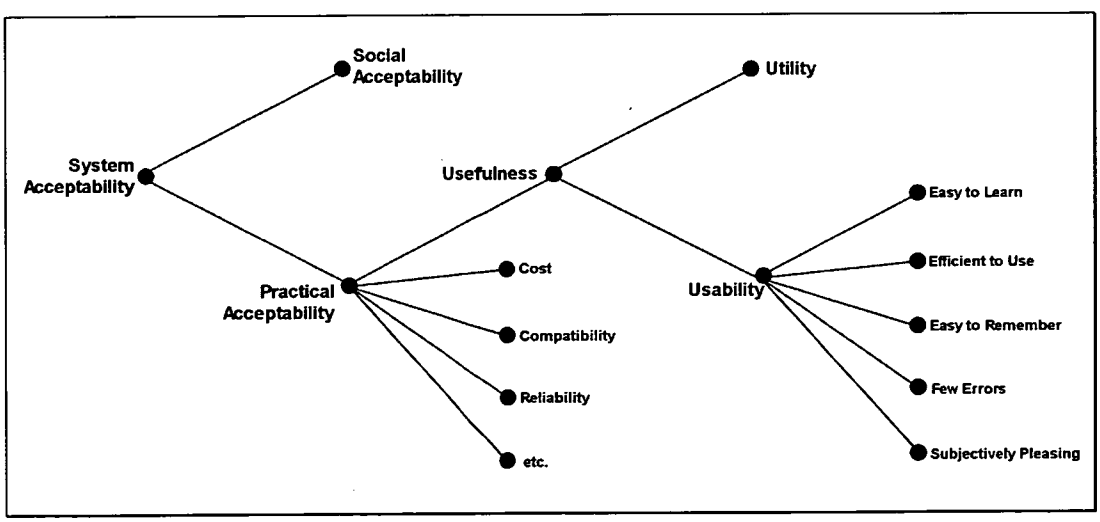


Figure 2.2. Nielsen's Taxonomy of System Acceptability (from Nielsen, 1993a)

This taxonomy leads to a usability approach to acceptance or adoption, and one of the main aspects of this definition is that usefulness of a system is linked to both its utility (system capability to perform functions) and its usability (the users experience of trying to do it). This is slightly different the most widely used adoption model: the Technology Acceptance Model (TAM) (Davis, 1989). With roots in IS research, TAM studies have generally focused on the organisational implementation of IT in job contexts.

The TAM is based on a well-researched psychology model, the Theory of Reasoned Action (TRA) (Fishbein & Ajzen, 1985) which suggests that attitudes and subjective norms influence intentions which in turn influence behaviour. The TAM extends this relationship by considering external variables influencing two key concepts: perceived usefulness (PU) and perceived ease of use (PEOU), which in turn influence attitudes, behavioural intentions (BI) and actual use. See Figure 2.3 for an illustration of the relationships proposed in TAM and confirmed by empirical study (the arrows indicate that relationships are causal).

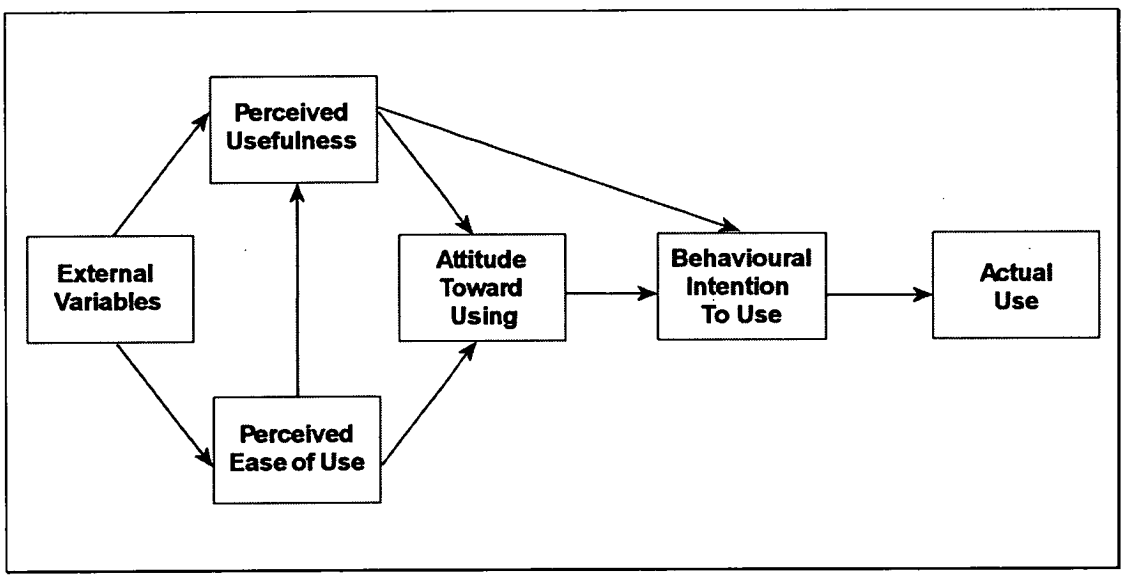


Figure 2.3. Technology Acceptance Model (TAM) adapted from Dillon & Morris, 1996 and Davis, Bagozzi and Warshaw, 1989.

Two versions of the questions used in TAM studies are available in Appendix B1, p.286.

The TAM has been widely used and many studies have been published establishing the relationships within the model. Fairly consistent findings suggest that PU is more important in determining BI than PEOU. In fact, PEOU influences BI through PU – often reported that a system is only useful if it is easy to use (Davis, 1989). This is similar to the Usability

perspective of Nielsen's taxonomy. However, these findings have also been used to imply that usability is less important than usefulness in the adoption model. This assumes that PEOU is akin to usability. However, the concepts of PU and PEOU may not directly map onto usability and utility (Dillon & Morris, 1999; Nielsen, 1993a). Furthermore, the concept of PU has been criticised in the literature. PU defined by Davis in his original model focused on advantage, efficiency and productivity (in a job context). As such, PU is a very broadly defined term that some have suggested will also be time dependent. Thus the concepts of current and future usefulness have been introduced (Chau, 1996) as an extension. They explain why ease of use influences usefulness: that is does so in the short term rather than the long term due to short term learning effects. The suggestion is that in the long-term usefulness will dominate in motivating adoption. This approach would have to be studied longitudinally. For the purposes of this study of self-service banking Websites, long-term effects are much less of a consideration as application usage is discretionary, unlike IT infrastructure in an organisational setting where learning and training may be relevant to system use. Various discussions about the definitions of PU and PEOU illustrate the difficulty of unifying the TAM approach with usability methods (Dillon & Morris, 1996 and 2002; Nielsen, 1993a).

A second aspect of the TAM which has been criticised is that it does not consider task issues or user goals. If the user's task or the system capability is not a strong consideration in the model, this constitutes a major difference between TAM and Usability research. The Task-Technology Fit (TTF) model proposed in parallel to the TAM focuses on the role of task in the technology adoption process. TTF considers the ability of the system to support tasks and individual performance (Goodhue & Thompson, 1995). The TTF has been used to extend the TAM and the PEOU → PU → BI relationship was again apparent. However, the results indicated no strong correlation between TTF and PU, Pearson's $r = .21$, not significant (Dishaw & Strong, 1999). Although the relationship is suggested in Nielsen's taxonomy, it has not been established empirically that utility is related to usefulness, or indeed PU.

Thirdly, criticisms of the TAM focus on the work context in which many studies have been conducted (Legris et al, 2003). The concern is that the TAM is not generalisable outside a job-performance context. However, in response to this concern, there have been several recent studies of more general populations, including one study of Web users (Chuan-Chuan Lin and Lu, 2000). So there is now further evidence that the model can be applied to wider contexts successfully. Other studies have focused on student populations, which similarly

could pose a problem as to whether relationships hold in a larger, more diverse user group (Legris et al, 2003). This however, is a common limitation to large scale research – universities and organisations have the ability to easily draw on large samples of users, but these samples many not be a good representation of the wider public, e.g. of potential Web users.

One of the major limitations of the TAM approach from a usability-centric viewpoint is that the method lacks the diagnostic qualities key to early prototype usability studies (Weir et al, 2007; Agarwal & Venkatesh, 2002). The TAM does not collect data on behaviour in performance of tasks, or on attitudes to various interface characteristics. Such data, when collected during or immediately after direct use, provide valuable feedback for an iterative design cycle (Shneiderman, 1987; Root and Draper, 1983). In conducting usability evaluations with a view to measuring usage intent, the TAM must be applied in addition to a UE method. This may be duplicating some efforts, as the TAM questions represent a small subset of typical questions posed in usability evaluation (see Appendix B for a comparison of questions). This suggests that it may be more appropriate to incorporate Usability techniques into the attitude → behaviour relationship of the TRA, which assumes that attitude influences intent and so actual usage. This aspect of the TAM has been incorporated into a recent study of Web site success (Agarwal & Venkatesh, 2002). Their study compared two different usability metrics, one based on Microsoft Usability Guidelines reformatted into an evaluation instrument and a short 3-question usability summary metric. They reported Web site usability a critical measure of quality and their metric an important predictor of success in eCommerce environments. Their research implied a link between usability and actual behaviour, but did not include assessing this empirically.

One final criticism of TAM studies is the use of self-report usage data as a substitute for actual use. Although many researchers suggest that the connection is valid (e.g. Igbaria et al, 1994; Davis, 1993), there are actually very few studies which include both subjective and objective measures of system use. In one such study, self-report use was found to be different from real-world behaviour (Szajna, 1996). Therefore many of the TAM studies which rely on self-report usage may be limited in their predictions of real world adoption behaviour. Unfortunately, the lack of real world adoption reports combined with the limitation that some studies also infer actual usage from behavioural intent reports, results in some confusion over the application of the TAM and its validity in explaining adoption (Legris et al, 2003). Much of the HCI literature stresses that observations of what people

actually do may not correspond with what they say they do. Self reports are often modified by what people think the interviewer wants to hear, or what is socially acceptable (McGrath, 1995).

Unfortunately, controlled usability experiments are subject to this social phenomenon. When combined with the unreality of the laboratory setting, this makes measuring likely real-world behaviour very difficult. However, measures of ease of use and usefulness in a TAM study have been shown to successfully predict choice between software applications (Szajna, 1994). This indicates that preferences may be influenced by the user experience, but no empirical evidence could be found to determine whether user preference, or choice between different options, can model potential real world adoption.

Although the TAM is the most widely used and cited technology adoption model, it is far from being unique. In fact, with reference to some of the other contributions to the adoption field, the TAM has been extended to include social and organisational factors, such as the TAM2 (Venkatesh & Davis, 2000) which considers social norms, image and job relevance amongst other determinants of PU. This has similarities to Neilsen's taxonomy, where social acceptability and practical considerations are both important in determining overall acceptability. Other extensions to the TAM have considered the aspects of computer anxiety, perceived fun and perceived enjoyment (Roberts & Henderson, 2000; Igbaria et al, 1995; Davis et al, 1992).

Other competing adoption models include Rogers' Diffusion of Innovations (Rogers, 1983). Rogers categorised adopters as innovators, early adopters, early majority, late majority or laggards. Each individual's willingness to adopt depends on their awareness of the innovation, them forming attitudes toward it, deciding and engaging in activities using the innovation and finally evaluating their experiences. Factors which Rogers considered influential included relative advantage and trialability. Thus the theory can be considered more of a marketing approach than the TAM. That by ensuring utility, providing a superior product compared to competitors, ensuring knowledge through a promotional campaign and allowing trials and demonstrations, adoption of the innovation will spread from innovators to the larger marketplace. A further factor that Rogers considered important is innovation complexity which has obvious parallels with usability constructs.

The DeLone-McLean Model of Information System Success has also been widely cited and recently updated for eCommerce. It suggests that system, service and information quality will all influence user satisfaction, usage intentions and net benefits (Delone & McLean,

2003). Again, this model is loosely related to the TRA relating intentions and actions. It is also highly related to usability and HCI: system quality could be considered utility or functionality and service and information quality are similar to ISO usability considerations of effectiveness and satisfaction.

A recently proposed adoption model which is clearly related to a modern usability definition is the P3 model (Dillon and Morris, 1998). P3 combines three main concepts: Power (utility) – an objective match of tool to task; Perception (user attitude or satisfaction) – a subjective evaluation of the tool; and Performance (effectiveness) – the objective behavioural data collected during user interaction with the tool, see Figure 2.4 (where double arrows indicate inter-relationships whilst single arrows indicate causal directions). Notice the similarity between this model and the definition of usability emphasising user experience in predicting success: functionality, efficiency and desirability on p.10 (Kuniavsky, 2003).

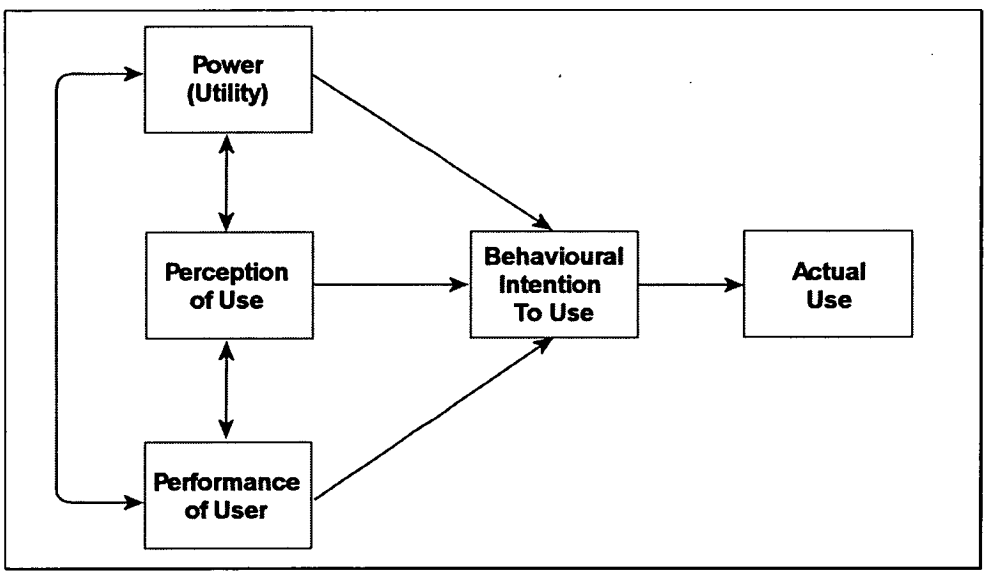


Figure 2.4. The P3 model of Adoption adapted from Dillon & Morris, 1998

The P3 model is more consistent with usability research than some of the other adoption models discussed. For example, usability work has shown that that performance does not always predict preference (Nielsen & Levy, 1994) and that user attitudes and the system capability should also be considered (Dillon and Morris, 1998, 1999; Frøkjær et al, 2000, Landauer, 1988).

Again, there are parallels between the P3 construct of Power and TTF, with P3 considering this utility component alongside subjective evaluations of use and objective performance metrics. However, there is no empirical validation of this model in the published literature.

One recent review of competing adoption models studied a variety of fields (Jeyaraj et al, 2006). They found the best predictors of perceived system use included: PU, attitudes, experience using computers and BI. Furthermore, the best predictors of BI were PU, relative advantage and subjective norms. In summary, the adoption literature generally describes themes common to usability definitions. Therefore the extension of using common usability metrics to guide potential adoption behaviour should be valid.

From the HCI point of view, there are several key issues in an adoption framework. First is to identify a task that could be supported by an IT system, then establish whether there is a potential market for the tool. If the potential market is large enough, this will provide the justification for development costs. The usability engineering method then considers what trade-offs to make in order optimise usability for the interface functionality and to suit the majority of the market. It would be of great interest to determine what association there is between subjective and objective usability measures (such as attitude and performance) and what amount of variance in usage intentions they account for. Models derived from outside HCI often focus on explaining why technology is adopted. In contrast, the focus of these studies was on alternative interface designs and user experience differences in explaining intentions to use.

2.1.4. A Proposed Relationship between Usability and Adoption

The review of various adoption and acceptance models and theories suggests the need to consider utility, usefulness, or functional capacity as well as objective measures of performance and subjective perceptions. Therefore an HCI experimental approach to measure and predict acceptability based on the P3 model (Dillon and Morris, 1998) is proposed to be explored in this thesis:

- Description of functionality and potential level of utility
- Hands-on experience by a range of potential users
- Exposure to alternative designs in randomised order of presentation with typical task scenarios
- Collection of objective performance data on task completion and errors
- Assessing perceptions of using the designs – subjective and comparative ratings
- Predicted self-report of intent to use (Shackel, 1991).

For the purposes of this thesis, the ISO definition of usability is expanded with reference to user experience and the definitions of Shackel, Nielsen and Kuniavsky. The working definition therefore includes:

- UTILITY: explores the usefulness and functionality of the artefact.
- PERFORMANCE: an indicator of performance, ability to use. Measured by rates of completion, error rate, error recovery, speed, etc.
- ATTITUDE: the subjective perceptions in use of an artefact. Measured by a collection of attitude statements – emotional appeal, cognitive effort, desire to use, etc.

Where the user and the interface interact, attitudes are formed about the interaction. The user and their task are concerned with performance, the ability to complete their task. The interface must have the utility of being able to perform the task functions. The combination of these interactions within general and specific contexts is known as the usability.

This definition is illustrated in Figure 2.5 and is used as a basis for the design of appropriate usability metrics. The figure depicts the main interactions between important elements of HCI in an evaluation of usability (not a Venn diagram).

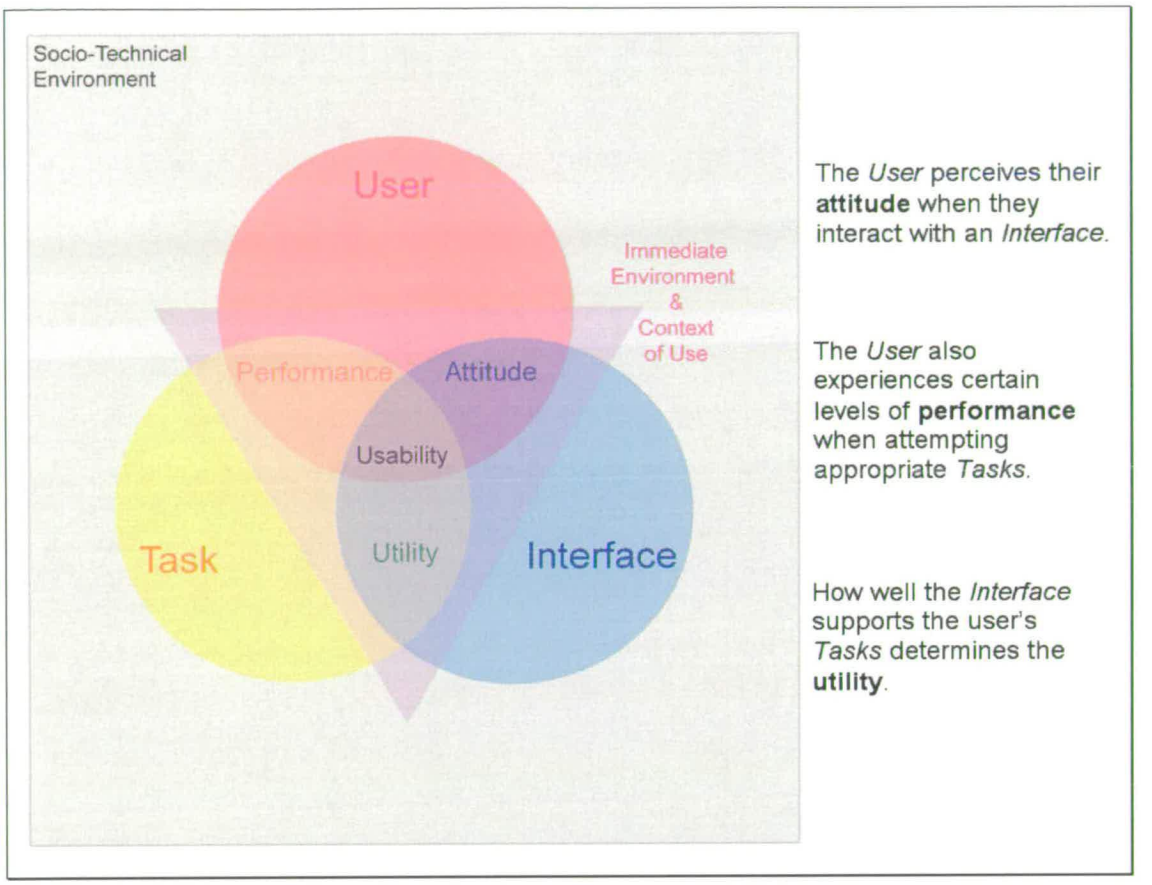


Figure 2.5. Model Based on the Extended Definition of Usability

The illustration concentrates on the interactions between the important aspects of the HCI process – the *User*, their *Task* and the *Interface* to the system. Usability is at the centre, involving a combination of factors based on the interaction of each HCI aspect within a general socio-technical environment and the specific situation in which use occurs.

Some of these measures are purely perceptions from the user point of view, whereas others can also be objectively measured.

2.2. Usability Engineering Methods

Usability engineering involves a collection of methods that take place throughout the project life-cycle (Nielsen, 1993a) with each being applied at an appropriate time. The variety and type of methods a usability engineer makes use of depends upon the time frame, resources and specific goals of a particular project (Preece et al, 2002).

Usability processes should be situated before design and development begin, as the focus on users, tasks and environments suggests. A product or service designed with the user in mind, fulfilling their requirements and providing utility has the best possibility of success. This is particularly true in a highly competitive domain such as the Internet. Profiling, field work and design principles, standards and heuristics are applied early to help concept formation. However, requirements, guidelines and heuristics can only go so far in determining the specific design of any interface. Early prototyping and testing with real users can deliver insight to refine user-interface designs or select the most appropriate design to develop further.

Some examples of established usability engineering methods and their respective applications include:

- Requirements gathering: task analysis and user profiling, investigation of environmental and physical attributes (Faulkner, 2000).
- Field studies (ethnographic inquiry): to understand usage and usefulness in the real-world and contribute to requirements gathering (Gould, 1988).
- Standards and guidelines: specification of relevant guidelines for the interface design for each project, checking adherence to imposed standards and *de-facto* conventions (Canny, 2006).
- Heuristic evaluation: expert inspection of potential designs following documented information and rules (Nielsen, 1993a).
- Cognitive walkthrough: experts and user groups performing task-based inspections of potential design flows ensuring intuitive and logical processes (Preece et al, 1994).
- Informal iterative prototyping and testing: using a small number of representative users to try the system, assessing their performance, making changes to the design as they are detected and retesting towards a more usable interface (Nielsen, 1993a).

- Formal experimental evaluation: using prototype interface designs to demonstrate they have met the required usability metrics, or to compare different implementation ideas (Shneiderman & Plaisant, 2005).

2.2.1. Formal vs. Informal Methods

Formal evaluation can compare a wide range of measures in a controlled way which allows scientific analysis of interface designs and a basis for creating design guidelines (Shneiderman & Plaisant, 2005). Experimental evaluation should ensure any specified usability requirements are met in the final product - but formal approaches have their limitations, e.g. in casual Web browsing (surfing), the artificiality imposed by experimental controls can make findings difficult to generalise and interpret. Different measures can not be effectively compared and attributed to interface components if each participating user does something totally different on a variety of Websites. Therefore this process is most appropriate to apply in environments where the interactions are task-based and specific. For such environments, formal testing is understood to provide more accurate and balanced assessments, less dependent on the tastes and preferences of any individual (Gould, 1995). eBanking transactions are task-orientated and highly specific, therefore provide a sound basis for controlled evaluation. Information retrieval on the Web is a less specific task domain, but using appropriate tasks on specific Websites should still offer enough control for formal study.

Informal, or 'discount', usability testing practices have become widely used in recent years. Using small numbers of representatives (as few as 5) to participate in a test, problems are often fixed as they are discovered. As such the interface is iteratively designed, tested, re-designed and re-tested until the appropriate levels of usability criteria are reached (Hudson, 2001; Nielsen, 1993b; Virzi, 1992). Informal approaches make usability improvements fast and cost-effectively in practical settings. However they apply to the continuous refinement of a specific interface or function and as such do not generate scientific data to test hypotheses.

The whole range of usability engineering methods have their place in the design of novel interactive products, facilitating improvements to current interfaces, updated versions or in comparing competing products or services.

However, it is of most interest to the academic community to perform formal evaluations, for example: to compare new designs to existing ones (Preece et al, 2002), or to compare

competing interfaces providing the same functionality. Doing so, it is often possible to determine whether to add a feature to a system (Landauer, 1988) or which competing design to carry forward into production. With formal data collection and analysis, these recommendations can be made reliably.

In practice, methods selected depend on the problems specific to individual projects. The usability engineering experiments which form the basis for this thesis are a specific process of evaluation used to compare design options, by examining the attitudes and behaviour of people using them under controlled conditions. This is the only type of usability engineering method considered in this work. However, the prototype interfaces used in the experiments were all subject to the application of standards, guidelines, were designed and heuristically evaluated by a team of usability and user-interface design professionals. Thorough walkthroughs were also performed to ensure the prototypes were robust and suitable to be evaluated by members of the public.

The selection of a formal method allowed specific hypotheses to be tested using statistical methods. Robust empirical data was complemented with qualitative information (comments and observations). The combination of data built up an important and useful picture of people interacting with Banking services online performing general information retrieval tasks, transactions and statement record searching.

Formal methods are based on experiment design techniques which will now be reviewed and discussed.

2.2.2. Experiment Design

In contrast to natural observational techniques, experiments attempt to control variables in order to investigate their apparent effects by prescribing the environments, the conditions and the tasks to perform. Experimental manipulations take the form of treatments (known as the independent variables) and are controlled by the experimenter. The experiment measures the outcomes of these changes on the dependent variables and uses statistical analysis to determine any differences and provide clues as to why different treatments behave as they do (Robson, 1983; Cochran & Cox, 1957).

Experiment design is a process where variables thought to be of importance to the outcome are identified and various controlled manipulations of them are planned. Other variables are identified and balanced or held constant in order to reduce their effects if they are not of

primary interest. Confounding occurs when two (or more) variables cannot be distinguished, for example if women use one interface, and men another, the effect of gender and interface can not be distinguished. Some uncontrollable variations are inevitable but it is important to ensure they are not systematically confounded. Introducing randomisation into the experiment design can control some of these effects sufficiently to validly use statistical tests (Fisher, 1971).

In Usability Engineering experiments, at least two treatments, typically alternative interface designs, are assessed in terms of usability (Shneiderman, 2002). There are several advantages to using more than one treatment, most notably it is more economical in terms of numbers of participants (this will be discussed further in Section 2.2.2.) but also there is evidence that the exposure to alternatives offers a more accurate subjective rating and the ability to make comparative criticisms and comments (Tohidi et al, 2006). Usability measurements typically include task completion, timings, behavioural observations (e.g. error counts) and attitude questionnaires, as shown in Figure 2.6.

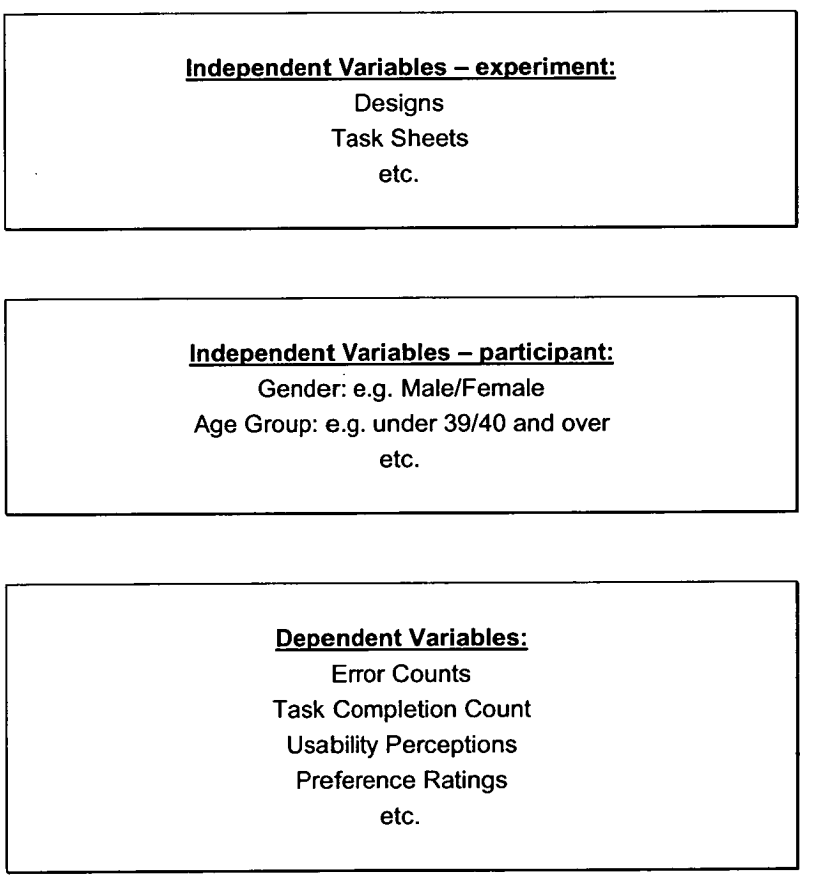


Figure 2.6. Examples of Types of Variable in this Research

The aim of an experiment (by convention) is to refute the null hypothesis, that manipulations of the independent variable will not effect the dependent variable (Moore, 2001; Cochran & Cox, 1957). Different statistical techniques for hypothesis testing will be explained in more detail in Section 2.3.

Some basic principles of experiment design are to control variables of interest, randomise other variables to reduce bias, to use enough subjects (Moore, 2001) and select the appropriate target population for the sample, so far as this is possible.

2.2.3. Advantages and Limitations to Usability Experiments

The advantage of experimentation is that factors can be studied in isolation, making use of statistical techniques to draw conclusions that, whilst not certain, are highly probable. A controlled experiment can also be replicated, for example in a second city. Such replications validate any findings and the discovered issues can be considered more reliable – although often the scope will be narrow (Shneiderman & Plaisant, 2005).

There are however, several common limitations to controlled usability experiments. Generally, experiments are likely to suffer from uncontrolled variations from many sources. Observer biases can occur, so procedure is standardised and each participant initially receives minimal instruction and is treated in a similar way. Randomisation eliminates any bias in allocation of participants to experimental conditions (Cox, 1958). Observing users may change their behaviour, making them, for example, more diligent towards completing tasks than unobserved users who might choose instead to give up quickly or more frequently (Schutle-Mecklenbeck & Huber, 2003). Other external factors such as noise, the environment, the time of day are sources of uncontrolled variation that may be controlled through randomisation techniques.

Experiment sample size can affect the chance of detecting significant results (Fisher, 1971). Too small a sample will not have the power to detect differences even when they are present. Obtaining a random sample of the population is also difficult. In practice, in eBanking research, inviting random users is not guaranteed to provide a truly random sample. The resulting sample will be made up of those who tend to be interested in volunteering in general. This is known as participation bias. In particular, certain types of personality, or very busy working people may not be represented and this is known as undercoverage (Moore, 2001).

For usability experiments in particular, the experiment situation and environment only approximate reality. However, it is common for experimental studies to try to simulate natural settings to offset a laboratory's artificiality (Baeker et al, 1995). Participants undertake specific tasks with various interfaces in the context of a convincing scenario. When engaged with the designs in a semi-realistic situation they are invited to suspend their disbelief and react to the designs as they would in the natural setting.

Control over experimental variables practically requires the use of limited prototypes and sets of prescribed tasks. Long-term, natural learning of the interface can not be modelled in this type of experiment (however longitudinal approaches are evolving (e.g. Vaughan & Courage, 2007)). Similarly, actual usage data can not realistically be collected experimentally, real world behaviour is inferred from the behaviour measured and observed in the laboratory. Finally, experiment-based research may also suffer from the Hawthorne effect, that the research process itself can alter behaviour and influence the variables under study (Burgess, 1993).

Nevertheless, the advantages of being able to control exposure to interfaces, and collect a wide range of data from real users is usually sufficient reason to continue to practise experimental research in usability. Controlled experiments and evaluations are well-suited to investigate banking applications where tasks are well established and can be presented in specific terms within well-defined scenarios that participants can relate to.

2.2.4. Basic Experiment Designs

"...Adding to natural knowledge by experimentation requires statistical procedure and experimental design."

(Fisher, R. A., 1971).

Experiments gather data from random samples of a population and use statistical techniques to infer the results for a population as a whole. In a between-subjects (classic) experiment, a group is subjected to a treatment and the results are compared to the results obtained from a control group with no treatment, in order to determine the effect of the treatment condition (Figure 2.7). This is a between-subjects experiment as each group is made up of different participants (Burgess, 1993).

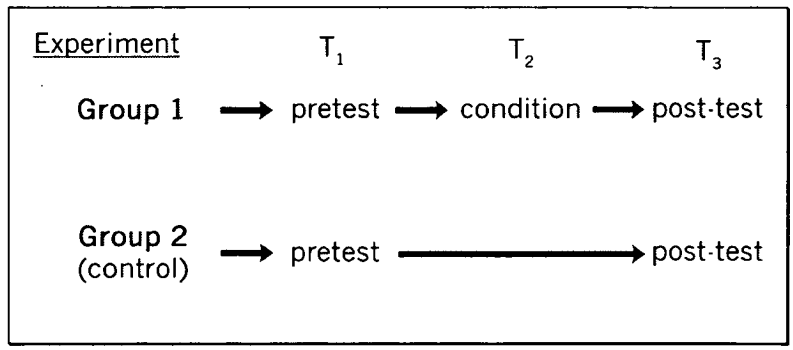


Figure 2.7. Typical Between-Subjects Experiment

A more economical design is the repeated-measures experiment (Figure 2.8). All participants will try all versions of a system being examined. Each participant controls their own unsystematic variations and can be thought of as serving as their own control. By reducing unsystematic variability, the repeated-measures design has greater power to detect effects with fewer participants. The order in which participants experience the conditions can influence the effect instead and is randomised in the design (Preece et al, 2002; Robson, 1983). It is sometimes of interest to balance for individual groups, e.g. different age groups and genders, where participants are thought to be similar in a way that might influence the results (avoiding confounding). Grouping allows these factors to be compared as well as the main independent variable (Moore, 2001; Landauer, 1988). Participants can also be asked comparative questions in such designs having gathered more than one experience in the session (Tohidi, 2006; Coolican, 1990; Whiteside et al, 1988). The result is that each participant will have a prescribed set of conditions, in a set order, with set tasks to perform, questions to respond to and measures collected.

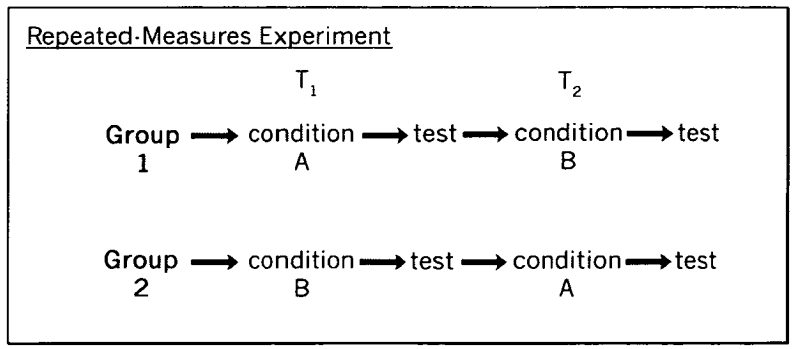


Figure 2.8. Repeated-Measures Experiment

The number of participants needed in these experiments depends upon the amount of segmentation required in the population and whether statistical methods are to be used. Studies on the Web require larger groups of users in order to evaluate any design sufficiently (Hudson, 2001, Spool & Schroeder, 2001).

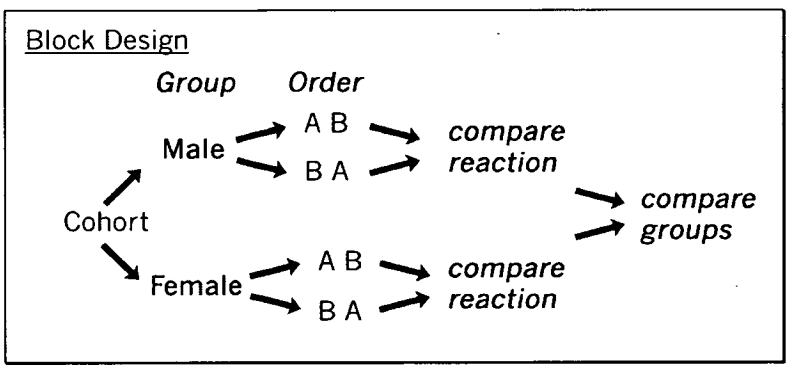


Figure 2.9. Group/Block Design

Sufficient numbers are needed in each key sub-group to lessen the impact of individual differences (Faulkner, 2003; Landauer, 1988) and to allow statistical tests to be completed. For example, for the variables gender (2 groups, male/female), age group (3 groups, <35, 35-49, 50+), interface order (2 groups, AB/BA), task allocation (2 orders, T1T2/T2T1), to be compared, the sample size should be a multiple of N = 2 x 3 x 2 x 2 = 24 and typically an experiment would recruit at least three people (and as many as possible given economy and time-constraints) to each of these sub-groups such that a sample of 72 participants would be the minimum sought, balanced in these factors. An example of this group or block experiment design is shown in Figure 2.9.

The experiment collects qualitative and quantitative data. Quantitative data can be recorded on different levels of measurement:

- Nominal data are collected where measures are simply indicative of the names of a set of responses, e.g. gender, where the responses have no intrinsic order.
- Ordinal data are measures on a scale which have a well-defined ordering in terms of direction, e.g. preference rankings. Ordinal scales do not define the size of the difference between points - simply ranking three interfaces in order of preference does not indicate the size of difference in preference between each pair.
- Interval data are similar to ordinal data, in that the rank order of categories is

implied, but in addition the intervals between points are equal, thus age, height and weight are all true interval data (Field, 2005).

- Binary data is treated as a special case, dichotomous with one of only two options being possible, e.g. Yes or No (McInnes, 2005[2]).

2.3. Statistical Analysis

Statistical analysis is used to describe the data collected and draw conclusions with respect to the experimental hypotheses. Different statistical tests are appropriate depending on the variables and type of data collected. A statistical test of significance is used to decide whether the variations in the collected data are due to some genuine variation between the sets of variables or if they are due to chance only. In the collection of experimental data there will be two types of variation in the data samples. Systematic variance due to the genuine effect between variables is attributed to controlled differences in conditions. Unsystematic variance is due to individual differences and natural differences between people in samples. Test statistics establish if the model being fitted to the data reasonably represents the larger population.

The general form of all test statistics is:

$$\text{test statistic} = \frac{\text{variance explained by the model}}{\text{variance } \textit{not} \text{ explained by the model}}.$$

Each test statistic, e.g. t, F, and χ^2 (see Section 2.3.2) has a defined distribution which allows the probability of obtaining the resulting value to be calculated. The probability indicates the likelihood of finding the experimental result in the true population. Statisticians have rules for deciding whether hypotheses can be rejected to ensure that true hypotheses are rejected only occasionally, and that hypotheses that are false are rejected as often as possible. Probability levels of .05 and .01 are useful conventions for these levels but they can also be specified by the experimenter (Moore, 2001; Fisher, 1971; Cochran & Cox, 1957). The

[2] Internal communication

statistical significance of an experimental result is defined as the probability that the experiment would produce the result by chance alone.

For significance testing of experimental data, the following basic procedure is followed:

- Define the null hypothesis (H_0) that the mean difference is zero. The hypothesis is formed as a testable statement. The statistical test will show the strength of evidence against this hypothesis.
- Specify a significance level, the probability of rejecting the null hypothesis if it is in fact correct.
- Run the statistical test to yield a probability value, *p*, the probability of observing data this different from what would be expected, assuming that the null hypothesis is true. As the probabilities get smaller, there is stronger evidence against H_0.
- Examine the size of *p* and if it is below the designated significance level then the result is termed significant enough to reject H_0.

The resulting *p* values are generally described as significant when below .05, highly significant at below .01 and very highly significant when lower than .001. When a test for a particular effect shows a *p* value of .0015, it means that there would be a 0.15% chance of obtaining evidence as strong as this (or stronger) for that effect, if the effect did not actually exist. Reporting the actual *p* values allows the reader to deduce the results at different significance levels (Moore, 2001). The two-tailed test looks for deviations at both ends of the distribution. In a one-tailed test a specific direction of change is predicted and focused on, see Figure 2.10.

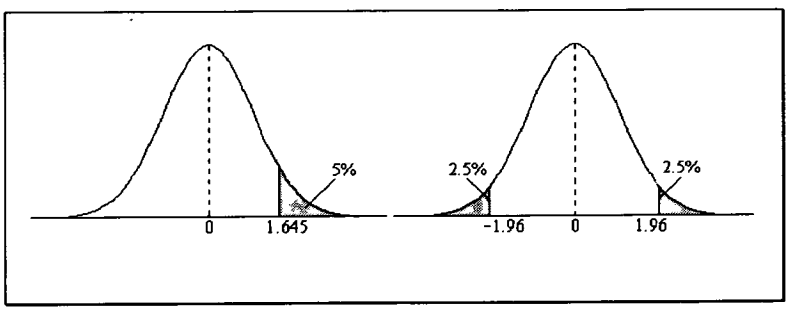


Figure 2.10. One/two tailed test illustration of a normal distribution

At the standard .05 significance level, there is a 5% probability that a result is due to chance. So, there is a 1 in 20 chance that the result could be wrong. There are two types of error that can effect conclusions: Type I errors occur when a genuine effect is found in the sample yet there is no real effect in the population; Type II errors are the opposite, that no effect is found in the sample where one really exists in the population. These errors are traded-off in the defining of significance levels and as a consequence this means that a statistically significant result is not certain but is made in reasonable confidence (Field, 2005; Moore, 2001).

2.3.1. Descriptive Statistics

Basic summary statistics describe the overall pattern of distribution in a collected sample of data: centre, spread and shape. The arithmetic mean is a measure of central tendency, the average of a set of scores. The variance is used to measure the dispersion of the sample, the average distance of the observations from their mean. Degrees of freedom (df) refers to the number of observations that are free to vary (Moore, 2001; Hoaglin, 1983).

Other measures of central tendency are the median (the central score when all scores are ranked in order of size) and the mode (the most frequently occurring value). The standard deviation is another measure of the data spread (square root of variance).

The Gaussian (normal) distribution is a bell-shaped distribution, symmetrical about a mean value such that 99.73 % of the samples will lie within three standard deviations above or below the mean. Human height, for example is distributed normally (Robson, 1983), see Figure 2.11 (taken from Blakeslee, 1914).

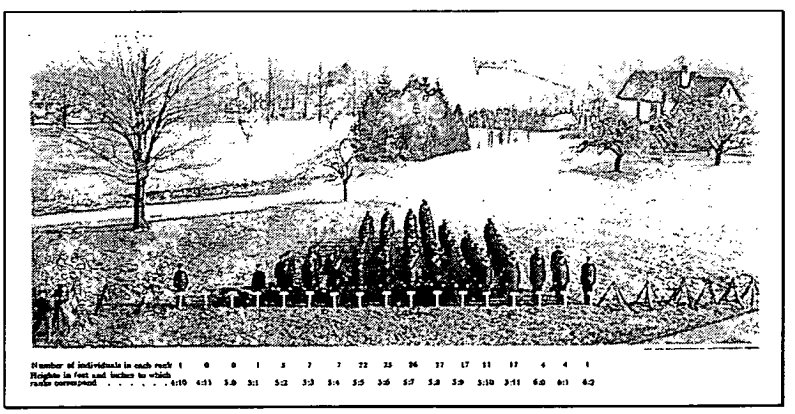


Figure 2.11. The distribution of human height

The standard error of the mean is the standard deviation of the means of a set of samples all taken from one population. Taking many samples from the same population enables a sampling distribution of the sample means to be constructed. Typically, there is much less variation in mean sample scores than individual scores within a certain sample.

Where samples contain more than fifty scores, the sampling distribution will always be normal, this is known as the Central Limit Theorem (Lane, 1999). A small standard error means than most samples from a population will have similar means. When the standard error is larger, the sample means can deviate much more from the population mean, and differences between sample pairs can be reasonably large by chance alone (Field, 2005).

2.3.2. Statistical tests

Different statistical tests are selected and applied depending on the level of measurement and adherence to basic assumptions about the sample and population. Practically, results obtained using the normal distribution can apply even when the distribution is different from normal (Robson, 1983). There are also cases where violations are not well tolerated by the statistical models and these will be discussed for each specific statistic.

Parametric tests are based on the assumption that the sample (and population) are distributed normally, assuming homogeneity of variance, interval data and independence - where the behaviour of one participant does not influence another (Field, 2005).

This research makes use of the Likert scale (Likert, 1932) for attitude measurement (described further in Chapter 3). Likert scales ask for a person's degree of agreement or disagreement with a statement using a range of options, typically 5, 7 or 9-points long. Although Likert scales are actually ordinal, the data can also be considered as interval, assuming that the differences between 'strongly agree', 'agree' and 'slightly agree' are all perceived as being equal (McInnes, 2001). In fact, only small errors result from assuming ordinal categories as interval, particularly when a 7 point rating is employed (Nunnally, 1978; Kerlinger, 1973; Labovitz, 1967). Thus Likert data can be analysed using parametric statistics such as the ANOVA and t-test. These statistics are actually robust to violations of normality and more powerful than using non-parametric equivalents – except when group sizes are very unequal (Kline, 2004; Kerlinger, 1973).

Some researchers suggest comparing the parametric and non-parametric procedures and examine results from both for consistency before assuming that the interval assumption is

valid (Anderson, 1961). This comparison has been made on the original usability attitude questionnaire for spoken dialogues (CCIR, 1992), where non-parametric and parametric tests both revealed the same significant differences from the question responses. Thus in the subsequent data analysis, parametric statistics are used on Likert response data.

The statistical tests in this research were all carried out using SPSS[3]. In a case where the hypothesis concerns the simple comparison of two means, either for two different groups (independent samples) or two repeated measures comparisons from each participant (paired samples), the t-test is applied.

t-test

Sample means are calculated for two different samples of data. The rational for the t-test is that the sample means should be roughly equal if they came from the same population, large differences should occur only infrequently. The t-test compares the difference between sample means to the difference expected based on chance alone, estimated by the standard error.

Data must be parametric and interval to perform the t-test, however the test is robust in respect to violations of the assumptions (Robson, 1983) with the exception of different sample sizes in the case of independent samples t-tests. The assumed homogeneity of variance for independent t-tests can be tested and appropriate adjustments can be made using Levene's test (Field, 2005).

The t-test output usually reports the sample mean and standard error, the test statistic, t, the df and p, the probability value of that test-statistic.

The t-test is limited to comparing two conditions and does not take into consideration all the variables usually requiring study in a single experiment. For this, the analysis of variance is conducted.

Analysis of Variance (ANOVA)

The ANOVA employs the F-statistic, named after R. A. Fisher (Field, 2005). ANOVAs analyse measurements where several different kinds of effects operate simultaneously to

---

[3] SPSS (Originally the Statistical Package for the Social Sciences) created by SPSS Inc., Chicago.

decide which effects are important and which to eliminate (Scheffé, 1959) and see how the variables interact (Field, 2005). The principle of an ANOVA is to divide the data into groups based on the within-subject and between-subject factors. F-statistics determine whether various sample means are the same by comparing the amount of systematic variance (caused by the experimental conditions) to the unsystematic variances (individual differences). F-statistics are produced for each group allowing both main effects of the experimental variables (between and within-subject) and their interactions to be analysed. There are various types of ANOVA. The simplest is a one-way univariate ANOVA, appropriate for testing the effect of a single independent variable (between participant factor forming at least two groups) on one dependent variable (condition). This extends to two or more between-participant factors, and further to repeated measures ANOVA's with within and between-participant factors.

The ANOVA assumes a normally distributed population, similar variances, independent observations and measurements on an interval scale. In fact, the ANOVA technique is very robust to most violations except the use of unequal sample sizes and independence of observations. For repeated measures experiments with three or more conditions there is an additional assumption (sphericity) which assumes variances across conditions are equal. Sphericity is tested using Mauchly's test (null hypothesis that variances are equal), therefore if significant (at .05) this condition is not met. The Greenhouse-Geisser correction is used to adjust the degrees of freedom in the calculation of the F-statistic, providing a more conservative estimate.

The ANOVA only highlights if there are differences between the various conditions. For cases of three or more conditions, it does not explain where exactly the effects lie between the alternatives. This is done in post-hoc routines based on multiple t-tests where pairs of conditions can be compared. When conducting multiple t-tests the Type I error is cumulative, for example: a significance level with 95% confidence for one test is actually equivalent to $(.95)^3 = .857$ over three tests. This raises the probability of making a Type I error above the acceptable criterion of 5% to 14.3%, so some adjustment must be made (Field, 2005; Cochran & Cox, 1957). There are several methods available, e.g. Bonferroni and Tukey. The Bonferroni test gives very good control of the Type I error for a small number of tests. Bonferroni adjustments are the most conservative, but perform well under small deviations from normally distributed data and violations of sphericity (Field, 2005).

The output for a repeated-measures ANOVA includes Mauchly's test and lists the F-values (and appropriate corrections) for the within-subjects effects, interactions with between-subject factors and the between-subject effects. The main ANOVA F-tests are used as a filter: if there is no significant overall effect, then post-hoc comparisons are not performed (Field, 2005; Cochran & Cox, 1957). If however, the main effect is significant, then within-subject effects and interactions can be explored using paired comparisons corrected for multiple tests.

The ANOVA reports the descriptives of the sample, mean and standard error, the results of Mauchly's test (if appropriate and any corrections to the df), the test statistic, F, the df, the error terms, and *p*, the probability value of that test-statistic. Examples are shown in Appendix A, p.279.

Chi-squared and Binomial tests

Chi-squared, χ^2, tests look at relationships where data is categorical. This is done by cross-tabulating two variables and comparing the frequencies observed in each category to those expected by chance. The test requires that each person contributes to only one cell of the table and that expected frequencies in each case are greater than 5 (Howell, 1997). The test can also be conducted on one variable, in this case testing that all categories are equally likely. The binomial test is a special case where there are only two levels of variable. It tests the hypothesis that the two values are equally likely (McInnes, 2005).

Pearsons' r

Correlation is a measure of the relationship between two variables. Pearson's correlation coefficient (*r*) is used on interval data and assumes normality. The correlation does not imply causality, nor does it explain the direction of causality. As it only refers to two measured variables, the relationship may be confounded by a third, unknown or untested variable. As for the other test statistics, the probability that the observed value of *r* occurred by chance can be found and the significance of that value reported, based on the distribution at the specific level of df. The size of the sample can have an effect on the significance of the value of *r*, such that at larger sample sizes, lower values of the correlation coefficient are significant (Bryman & Cramer, 2001).

The correlation coefficient, *r*, squared, R^2, is a measure of the amount of variability in one variable which is explained by the other. This is also known as the correlation ratio or effect size (Kline, 2004). Again, it does not indicate causality (Field, 2005).

Cronbach's alpha

This is a statistical procedure used to analyse reliability. It checks that individual items in a questionnaire set all produce results consistent with the overall questionnaire. Cronbach's method is an extension of the split-half technique, literally checking that each half of the dataset (arbitrarily partitioned) are highly correlated. Cronbach's method is equivalent to doing the split in every possible way, to avoid bias in the method of choosing the two partitions. The average correlation of all the computed splits is Cronbach's alpha, α. The convention is to seek reliability of above 0.7 or 0.8 (Field, 2005), and Likert attitude scales have been shown to have high reliability, with values of 0.9 not unusual (Oppenheim, 1992).

The analysis also gives the value of alpha if each item were deleted. If the values of α increase above the overall α, this indicates that the item might be removed to increase the scale reliability. Otherwise all items are thought to be positively contributing to the reliability of the overall questionnaire (Likert, 1967).

Similar techniques can compute the reliability of subgroups of questionnaire statements (Field, 2005).

2.3.3. Drawing conclusions and recommendations

Conclusions drawn from statistical analysis are by no means certain. The results represent measurements taken from a sample and are used to estimate likely values in the population as a whole. Gaining a valid sample is key to this process, but is fraught with difficulty. Results from larger samples are less variable than for a smaller set of responses, therefore samples should be as large as practical. Control and randomisation in the experiment design can reduce bias and confounding and allow several comparisons to be made in one experiment. The ANOVA allows inferences to be made about the main effects of variables and their interactions, drawing more precise conclusions, however unknown confounding variables may also contribute to the effects found.

Measurements of attitude toward usability (as for other attitude measurement) have yet to be proved accurate in predicting behaviour with respect to an object, and some argue that instead a person brings their attitude in line with their behaviour (Fishbien, 1967). Thus effects are typically considered two-way rather than causal.

Question wording has a potential to bias responses and must be carefully piloted in order to mitigate this effect. Perceived usability scores measured using mean attitude scores are not substantially useful taken in isolation. Although, this is where a benchmark, or another overall metric may prove helpful. In the comparative arena of repeated measures experiments, usability attitude scores provide a powerful basis for statistical analysis of differences measured between conditions. Using robust statistical tests, conclusions can be carefully drawn. Reporting the actual p-values associated with analysis is done on the understanding that the criteria for significance is a trade-off between rejecting a true hypothesis and not rejecting a false one. There is no definitive answer, just increasingly strong evidence as the probability decreases (Moore, 2001).

The statistical insights are typically augmented by qualitative data gathering, allowing recommendations to be formulated. Recommendations from usability experiments typically take a leap from the experimental data into the realm of user-interface design – thus redesigns may in fact cause more usability issues (Nielsen, 1993b). Therefore the experiment should be considered as only one part of a larger user-centred and iterative design process.

2.4. Summary: The Usability Evaluation Methodology

The controlled experimental approach to usability evaluation therefore requires:

- A set of product variants – the alternative design conditions to be used in the experiment.
- A set of hypotheses – the statements which the experiment will test, formulated as null hypotheses of no differences.
- A group of participants – representing members of the target population of users, as large as practical.
- User criteria – aspects of the users and their behaviour used to measure the effectiveness of the product for various key groups (demographics and technographics).

- A set of tasks – representing real-world tasks that must be completed using the design or designs, usually presented in the context of a fictitious persona (see Appendix D) and scenario to aid experimental control.
- Measurement techniques – used to measure the usability of the user interface against the criteria selected (e.g. usability attitude questionnaires, task performance measures).
- Qualitative data collection – used to complement measurements
- Statistical analyses – techniques to describe and draw conclusions from the collected data.

The experiments described in later chapters were applied in the field of Banking and the Internet, with a view to informing the user-interface design for successful Web Portals and self-service eBanking functionality.

2.5. Introduction to eBanking

eBanking services are Web-based user-interfaces that allow users to manage their bank accounts and transactions remotely. Banking has changed dramatically from the traditional branch, firstly by employing self-service in branches (e.g. Automated Teller Machines or ATMs), then with banking services through the telephone, the Internet since the 1990s (Shaw et al, 1997) and now banking on the move with mobile technology.

Banks are continually looking for innovative services to offer their users, hoping to provide differentiation between their offerings and that of their competitors (Lee & Kim, 2002), eBanking services offer a range of account management facilities to users who register for online access. However, banking tasks are generic in nature, which explains the apparent functional convergence in different eBanking services (Schubert & Dettling, 2002). The financial sector has always been quick to respond to new technology (Stamoulis et al, 2002) and as the market for eBanking and mobile banking grows (National Statistics, 2006) there is still scope for Banks to attract new users and hold on to existing accounts.

Financial institutions migrating to the Internet do so for the potential rewards: cost savings are by far the most often cited benefits (e.g. Jayawardhena & Foley, 2000; Sohail & Shanmugham, 2003). There are considerable differences in cost between different channels (Gopalakrisnan et al, 2003). Internet transactions cost much less than either ATM or

telephone services but research suggests that Internet users are looking for these cost savings to be passed on to them (Tan & Teo, 2000), predicting that Internet users will not expect to pay for services that are otherwise provided free of charge (Jayawardhena & Foley, 2000).

Yet the Internet also poses a threat to businesses from competitors: users have quick and easy access to a wide range of financial information and services, with competitors only a click away (Siddarth & Chattopadhyay, 1998). In this way, Banking on the Internet has much in common with eCommerce. Yet eBanking has not become a true eCommerce application as users cannot use an eBanking service from a bank without opening an account first, in contrast to the way that shoppers can try many competing eCommerce Websites before going ahead with purchases. Whilst eBankers are reluctant to switch banks (BBC News, 2004), banks will continue to have an advantage over retailers in this respect. However, there is some evidence that switching banks is becoming simpler and more common (BBC News, 2005). Aspects of eBanking that have a potential to drive switching behaviour are important in understanding adoption, loyalty and repeated usage. To produce useful service applications on the Internet, designers must understand the motivation behind initial and continuing use. For eBanking, core drivers for the adoption cycle are convenience (Lichtenstein & Williamson, 2006; Centeno, 2004) and control (Shih & Fang, 2004).

As users adopt and become frequent users of a service, they begin to demand better ease of use, something which is increasingly true of eBanking as the sector matures (Hudson, 2002). Ease of use and user friendliness are growing concerns from the bank point of view as well: as they rollout increased self-service functionality, greater demands are put on the user to understand financial transactions and complete them in an error-free and timely manner. As users become more experienced with online services they will desire extensive functionality (Shneiderman & Plaisant, 2005). This creates an opportunity to provide eBanking adopters with both core and innovative services. To retain these users, banks must work to meet the users' requirements. Usability and functionality are two aspects of design which are known to influence real-world usage (Whiteside et al, 1988). Where use of a system is discretionary, innovation can attract users.

Human factors and usability specialists point to human needs as the driver for innovation (Shneiderman, 2002). New computer systems should be designed with improved processes for completing current tasks, rather than simply emulating inefficient methods. Indeed, many online services are still driven by technology, rather than focusing on the needs of the users (Holland & Westwood, 2001; Zwass, 2003). In addition to knowing how a task is currently

performed, the underlying functional reasons also need to be exposed (Nielsen, 1993a). From such functional analysis innovative design solutions may become apparent.

Usability engineering can play an important role in the design of eBanking functionality and user-interface design: the idea simply is to "know the user, know their task" (Faulkner, 2000). From the users point of view, eBanking is a convenient way of performing transactions outside normal opening hours and without physically visiting the branch (Sheth & Sisodia, 1997). User requirements and expectations for eBanking have been widely documented in marketing and usability literature (i.e. Jamal & Naser, 2002). It is generally agreed that that control, accuracy, security, speed, user-friendliness and convenience (Jayawardhena & Foley, 2000; Liao & Cheung 2002) are key attributes that users desire. Thus designs have to be carefully constructed to provide information conveniently, be tolerant of errors, easy to use, efficient, innovative, functional and enjoyable.

The body of research relating to usability and eBanking services is not as extensively documented as for Graphical User Interface (GUI) design or Internet (Web) interface design. This is mainly due to the commercially sensitive nature of eBanking and the fact that the field is still relatively new. To date, several key design issues relating to eBanking usability have been defined: users tend to be very private about their finances and are concerned about eBanking security (Furnell, 2004); error prevention and recovery are very important with regard to self-service account management (Liao & Cheung, 2002); consumer trust in eBanking services has an impact on levels of adoption (Aladwani, 2001). Of these, the issue of trust is the most common focus of published research (Grabner-Kräuter & Kaluscha, 2003; Yousafzai et al, 2003; Suh & Han, 2002; Kim & Prabhakar, 2000; Kim & Moon, 1998). Other research focuses on user requirements for eBanking (Jayawardhena & Foley, 2000), and studies of the factors affecting eBanking adoption (e.g. Lai & Li, 2005; Centeno, 2004; Tan & Teo, 2000). Research has identified a range of usability methods that would be appropriate to apply to eBanking services (e.g. Wenham & Zaphiris, 2003). Although usability studies typically emphasise hands on usage and the collection of both performance data and opinions, much of the eBanking work is currently based on surveys and expert opinion (e.g. Kim & Prabhakar, 2000). There are relatively few published studies observing actual users performing banking tasks online (Nilsson et al, 2005).

The distinct lack of empirical research evaluating interaction with different eBanking interface designs lead to the work in this thesis (Weir et al, 2007; Weir et al, 2006). This work will also empirically research the relationship between usability metrics, interface

preferences and usage intentions. The first step is to construct the usability metrics using the extended definition of usability in the Banking context, specified as shown in Figure 2.12.

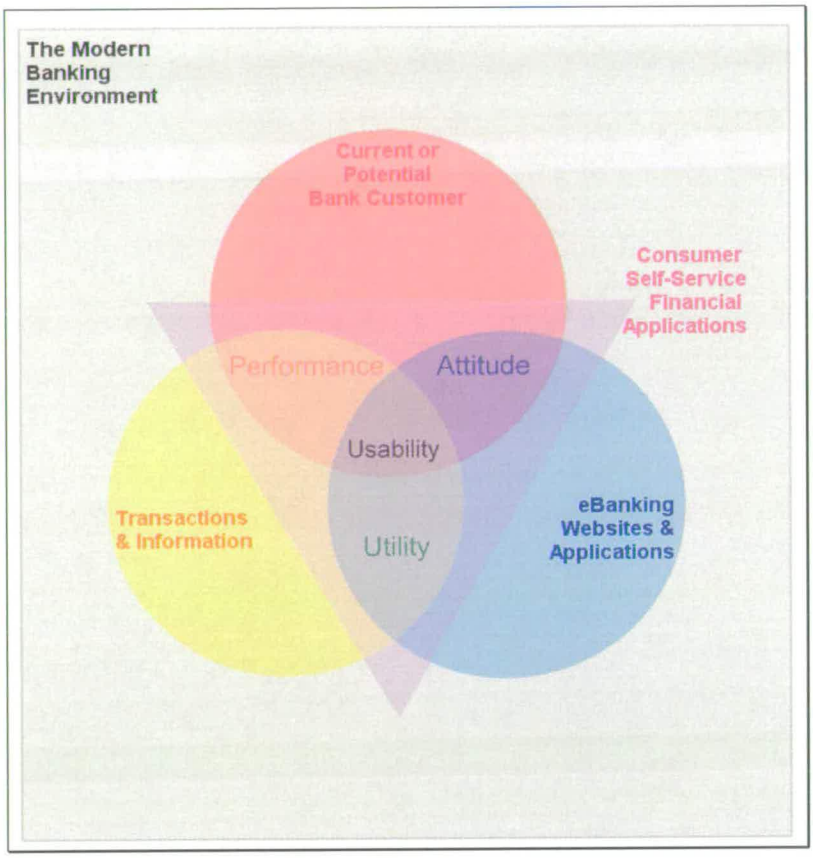


Figure 2.12. eBanking Usability Specification

In this figure, the target *User* is a bank customer (or potential customer), the *Interface* is the banking Website, the *Tasks* are banking transactions and information gathering. The target users interact within a larger modern banking environment, and in this case performing tasks in a self-service environment. These considerations result in *attitudes*, *performance* and *utility* measures for this particular set of circumstances. Altogether, these aspects of the user's interaction combine to explain the usability of the service they experience.

Chapter 3. Usability Metrics for Banking Online

In order to quantify the usability of an artefact, the concept of usability must be defined in terms that can be measured. For Web sites, this means that important characteristics of the Web-interface which may contribute to usability must be determined. Two subjective measures of usability were created to explore alongside task performance. Firstly a proven usability attitude questionnaire (see Appendix B2) was adapted to relate to Internet interfaces (such as banking Websites). In addition, a quantitative and relative preference scale was designed to elicit comparative ratings from alternative versions of Web designs experienced in a repeated measures exposure. The comparison between objective usability (performance), subjective usability (attitude) and preference (quality rating) can then be made with a view to better understanding of how these concepts relate to each other in a general Web context of information retrieval, and in relation to performing banking transactions online.

This work began by considering the online presence of the Case Bank: Web sites as a medium for presenting information accessible to all Internet users. Subsequently, the research focused on eBanking services, where customers access and manage their accounts online. There are key differences between the two types of banking Website. The Website aims to promote the provider brand and offer information on products and services. This information is targeted at a wide range of Internet users, and one of the main considerations must be information-retrieval: locations, contact points or phone numbers, account rates and charges etc. The eBanking service is restricted to registered customers of the Bank and serves to display specific customer-account information, allow transactions to be performed and money matters to be managed. This service is targeted at Internet-savvy customers, the functions being self-service account activities: balance enquiries, transaction details, transfers and payments etc.

First, the general case of Web-based information retrieval is considered, which could be applied to many Web sites including banking, commercial and governmental. The Web usability questionnaire developed for this purpose was based on published work (Dutton et al, 1993 and 1999; McBreen et al, 2000; Marshall et al, 2001). The development of the attitude questionnaire will be described in this chapter.

3.1. Measuring Usability

Studies of user interactions in experimental setting involve developing measurements of usability. Two of the three pillars of usability (as defined by the ISO) can be related to quantitative measurements: efficiency measures such as timing participants provide a continuous dependent variable, while effectiveness can be measured quantitatively by counting the number of errors made, the number of clicks or whether tasks are completed, providing interval data.

The third well-established aspect of the ISO definition of usability, satisfaction, is not so easily quantified. Early measures, for example, consisted of recording verbal and physical expressions of frustration during the interaction. However, such records give only a crude indication of satisfaction. Moreover, they rely on participants consistently displaying these outward signs throughout their interactions, and highly skilled observers are required to record and document the signals reliably.

Fortunately, research efforts in psychology have long included methods of measuring attitudes (Fishbein, 1967). These theories have been successfully applied to formulate quantitative attitude measures for peoples' perceptions of usability. Questionnaires are highly useful devices for collecting large amounts of data quickly and easily, and are commonly used in usability evaluation. They are presented after direct experience.

> *"...Asking users about the value of some proposed change without giving them experience of it is an essentially useless guide to their satisfaction with it in practice."*
>
> *(Root & Draper, 1983)*

One such usability attitude questionnaire relating to pre-defined usability aspects has been widely reported, extended and validated (Love, 1997; Love et al, 1994; Dutton et al, 1993; Jack et al, 1993; Love et al, 1992). The questionnaire was developed to provide a standard set of usability attributes and has been successfully used in a wide range of experiments with various technologies and interfaces. The original questionnaire attributes were used in a between-subjects controlled experiment comparing one group who, after using an automated telephone service, were asked to pick and rank their six most important attributes from a proposed attribute list and add any additional items to the list. A control group completed the same procedure being asked for six attributes without access to the proposed list. The results showed strong similarities between the attributes identified by the two groups. No additions were made to the list by the treatment group, and the spontaneous responses from the control group matched items in the proposed list (Dutton et al, 1993). This research resulted in the development and refinement of a robust questionnaire for evaluating user perceptions of automated telephone services. The questions have been modified successfully to evaluate synthetic agents in multimodal interactions (McBreen et al, 2000) and interactive television usability (Dutton et al, 1999; Marshall et al, 2001). The questionnaire attributes and wording were also adapted to evaluate the usability of the Web (Internet) and visual (screen-based) interfaces (Weir et al, 2006; Weir et al, 2007), as will be described in this chapter.

An attitude questionnaire typically consists of a set of twenty to twenty-five basic statements (Edwards & Kenney, 1967). For a subjective usability metric, the statements are selected to represent the key attributes for evaluating the usability of interfaces. Each questionnaire item relates to an issue that has been documented to potentially enhance or reduce usability (e.g. Shneiderman & Plaisant, 2005). The questionnaire is structured in the Likert (Likert, 1932)

format using a 7-point response scale ranging from strongly disagree (1) to strongly agree (7). The use of attitude measurement scales allows for the application of scientific processes to the study of peoples perceptions (Thurstone, 1967). The scale includes a neutral response. There is much discussion in psychology of the benefits (or not) of using a neutral point on the scale (Guildford, 1967). In this application, it was considered helpful to include the neutral response as it can indicate items that were less relevant to participants in terms of their perceptions of usability. If individual scores toward an attribute are consistently neutral (4) on the scale, this will contribute to evidence that the item is superfluous to usability attitude measurement (Oppenheim, 1992).

The resulting list of statements is typically counterbalanced (Coolican, 1990; Oppenheim, 1992) using an equal number of positive and negative phrased attributes, and presented in a randomised order. This technique controls for acquiescence in the response set, the tendency to agree rather than disagree to any question statements (Colman, 2001; Guildford, 1967). For the statistical analysis, polarity is normalised such that a score of 7 (on the 7-point scale) is consistently indicative of a strong positive attitude. The result is a Likert summated rating scale (Edwards & Kenney, 1967). Overall attitude toward usability is determined for each interface by calculating the mean of all the usability question responses from all participants. Individual statements are also analysed separately to identify any specific areas in each interface where usability could be improved (Agarwal & Venkatesh, 2002).

The Likert-based approach was selected for its simplicity and high reliability (Edwards & Kenney, 1967). Participants prefer them to other question formats and find them simple to respond to (Oppenheim, 1992). Mean scores of greater than 4 (on a 7-point scale) indicate a relatively positive response to usability (or to an individual attribute), whilst scores of below 4 indicate potential problems in the interface.

In the studies presented here, amendments to question statements were conducted under rigorous conditions. First, interface issues and potential statements were brainstormed in a session comprising a number of experts familiar with the Internet and with reference to the interfaces under evaluation. Then, relevant literature was reviewed to ensure consistency with published results and criteria. Questions were then constructed into unambiguous statements, again using a panel of experts (Moore, 2001; Guttman & Suchman, 1967). Finally, the use of pilot testers ensured that the questions were considered relevant and were comprehended (Likert, 1967). The addition of any extra questions relating to task-specific,

technology-specific or interface-specific qualities for each set of interfaces being tested also followed the same procedure.

3.1.1. Attitude Questionnaires in Usability Studies

Many different attitude questionnaires have been developed to assist in researching user perceptions and behaviours when interacting with modern technology. In addition to the questionnaire presented in this thesis, other research organisations have also developed usability questionnaires and modified them for various technology evaluations (Lewis, 2002; Kirakowski, 1994; Chin et al, 1988).

Many published usability evaluations employ home-grown attitude measures developed for individual studies, for example Hornbæk (2006) reviewed over one hundred studies and found only twelve of these used questionnaires based on previous work. Hornbæk (2006) also comments on the wide range of different questionnaires and measurements used to relate to user satisfaction, for example, Hornbæk lists a set of words commonly seen used in usability questionnaires across published literature in the area, as shown in Figure 3.1.

Accessible, adequate, annoying, anxiety, appealing, boring, clear, cluttered, comfortable, competent, comprehensible, conclusive, confident, conflict, confusing, connected, convenient, desirable, difficult, dislikeable, dissatisfied, distracting, easy, effective, efficient, embarrassed, emotional, engaging, enjoyable, entertaining, enthusiasm, excellent, exciting, familiar, favourable, flexible, flustered, friendly, frustrating, fun, good, hate, helpfulness, immediate, important, improving, inefficient, intelligent, interested, intuitive, involved, irritation, learnable, likable, lively, loved, motivating, natural, nice, personal, plain, pleasant, preference, presence, productive, quality, quick, relevant, reliable, respect, responsive, satisfied, sensate, sense of control, sense of success, simple, smooth, sociable, social presence, stimulating, successful, sufficient, surprising, time consuming, timely, tiring, trust, uncomfortable, understand, useful, user friendly, vexed, vivid, warm, and well-organized.

Figure 3.1. Diverse Range of Expressions Used in Usability Evaluation (from Hornbæk, 2006)

The consequence of such diversity is that it is difficult to make comparisons between different studies. Further, there is less confidence in the reported usability results (Hornbæk, 2006). In more recent work, researchers have recommended the use of standard questionnaires as they typically demonstrate high reliability (Hornbæk & Law, 2007). Therefore the Web usability questionnaire was adapted from an attitude questionnaire designed to assess the usability of spoken dialogue systems (Dutton et al, 1993).

3.2. Formulation of the Web Usability Questionnaire

In the development of the Web usability attitude questionnaire, the original statements were adapted from referring to listening and interacting using voice and/or the telephone keypad to the visual aspects more typical of the Internet. A system running on the Internet is characterised by interacting with hyperlinked text and graphics on a screen, or perhaps using a multimodal interface. The interface 'dialogue' which describes the communication between the human and the computer will be perceived and evaluated in slightly different terms in these different cases. For example, audio systems have aesthetic and clarity contributions from the voice, whilst for Web pages it is the 'look and feel' that matters. Yet many basic usability attributes apply to almost all interfaces. Therefore modification of the telephony usability questionnaire is legitimate, as is common in commercially available questionnaires such as SUMI (Software Usability Measurement Inventory) and QUIS (Questionnaire for User Interaction Satisfaction) (Shneiderman & Plaisant, 2005).

In the majority of cases the original question items needed only minor adjustment in wording to be more appropriate to the Internet medium, these are indicated by a star (*) next to the attribute e.g. "I felt under stress when using this ATS[4]" was altered to be "I felt under stress when using this Web site".

In a few cases, the questions were made more specific in order to better diagnose interface concerns, such as "I found the service confusing to use" was altered to stress page layout "I found the page layout confusing to use". Similarly, the attribute *concentration* additionally specified reading pages, *frustration* was related to the site organisation and *complication* concerned moving about within the Web site. These statements are indicated by two stars (**).

---

[4] ATS – Automated Telephone Service.

Certain questions required more extensive changes due to the medium, e.g. "I liked the voice" was altered to be more general, "I liked using this Web site"; "I thought the ATS was efficient" was amended to the more specific "I could quickly find what I wanted on this Web site"; "The ATS was easy to use" was amended to "The pages on this Web site were easy to understand". Similarly, *politeness* was replaced by attractiveness; *voice clarity* was replaced by page layout clarity. These amendments are indicated by a superscript sharp next to the attribute ($^{\#}$).

Two of the original questions were discarded: "The prompts were too fast for me" – highly relevant for audio interfaces, but with no real equivalent in terms of interactions online where visual information is typically static and the user controls the time they spend locating and selecting between different options; and "I would prefer to talk to a human being" – again highly relevant when a customer is making a telephone call and possibly expecting a human operator rather than an automated service, but not usually the case for customers going online for information. Instead of the discarded items, some additional attributes were also considered in the pilot tests as described below. All the final attributes relate to the documented user-interface and user perceptions which are considered to influence usability. Several aspects are considered semantically grouped: Visual appearance, Content, Integrity, Structure and flow, Affect and Quality.

Visual Appearance

In visual design, aspects such as colour, typography, layout (proximity and alignment), whitespace and contrast ensure the clarity, unity and flow of work. These are key graphic design principles, and have similarities to Gestalt Principles of visual perception (Similarity, Proximity, Continuity and Closure) concerning how elements of a whole are grouped and organised (Colman, 2001; Lidwell et al, 2003). Webpage layout can thus either contribute or detract from information comprehension and discovery (Lewis, 2002; Marcus, 1995) and clarity is typically highlighted as important (Nielsen, 2000; Spool et al, 1997; Ives et al, 1983). Guidelines suggest pages be designed with minimal clutter (Nielsen, 2000; Cooper, 1999; Neilsen, 1993a). In addition, visual interfaces are often judged on aesthetic attractiveness and graphical content (Lindgaard et al, 2006; Lidwell et al, 2003; Lewis, 2002; Nielsen, 2000; Spool et al, 1997; Baeker et al, 1995; Marcus, 1995). Eye movement data can also inform page layout, item placement and expectations from previous experiences

(Tzanidou et al, 2005). These considerations resulted in the following modifications and additions relating to visual attributes:

- Confusion with page layout**
- Clarity of page layout#
- Attitude toward appearance#
- Degree of clutter
- Attitude towards graphics and pictures

Content

In order to supply useful data, pages and content should be understandable (Lewis, 2002; Thimbleby, 1990), and the interface should ensure legibility in selecting the default size of text (Nielsen, 2000; Ives et al, 1983). Thus the following content related attributes were included:

- Ease of understanding#
- Legibility

Integrity

Web pages in the context of a service offering, particularly when considering banking data, should contain helpful information (Bevan, 1995; Kirakowski, 1994; Chin et al, 1988; Ives et al, 1983). Important for a wide range of eCommerce and banking activities, the application should also be perceived as reliable (Chin et al, 1988; Ives et al, 1983) and trustworthy (Kim & Moon, 1998).

- Helpfulness
- Reliable*
- Trustworthy

Structure and Flow

The Internet uses hyperlinks forming a web of information linking different pages. Thus, navigation of these pages has an effect on usability, therefore many usability guidelines refer

to creating simple, visible navigation design (Lidwell et al, 2003; Lewis, 2002; Chin et al, 1988). Similarly, in terms of completing a linear process or a complex task requiring several steps it is important to know what to do next, to know where in the process you are – the process flow (Raskin, 2000; Kirakowski, 1994; Chin et al, 1988; Norman, 1988) and to be able to locate items efficiently (Lewis, 2002; Nielsen, 2000; Kirakowski, 1994; Chin et al, 1988). Navigation and hyperlinks within the site and within the content must be visually salient in order to be discoverable (Norman, 1988), and be labelled clearly and understandably (Rosenfeld & Morville, 2002).

Often cited by researchers (and generally considered to be of vital importance) is to give the user a feeling of control over their interaction with the interface (Nielsen, 2000; Karat, 1997; Bevan, 1995; Kirakowski, 1994; Thimbleby, 1990; Whiteside, et al, 1998). In a similar vein, it is also important to match the users' expectations (Kalback & Bosenick, 2003; Lewis, 2002; Karat, 1997), remembering that they are likely to spend more time on other interfaces, so consistency with what they have previously learned and experienced can provide usability cues (Neilsen, 1996). This also indicates the potential role of prior experience in usability evaluations. Therefore the questionnaire includes these issues of structure, flow and control:

- Procedural knowledge*
- Frustration with organisation**
- Complication of navigation**
- Orientation
- Speed finding#
- Link labelling
- Visibility of links
- Degree of control*
- Match expectations

Affect

Cognitive and emotional impressions are often described by users in verbal reports, such as requiring too much concentration in reading high density pages (Spool et al, 1997; Chin et al, 1988; Ives et al, 1983), feeling flustered or stressed when using the interface (Nah & Davis, 2002; Baeker et al, 1995; Bevan, 1995; Kirakowski, 1994). Finally, some general satisfaction

questions regarding perceived user-friendliness, attitudes to using and enjoyment (Lewis, 2002; Spool et al, 1997; Bevan, 1995; Chin et al, 1988; Ives et al, 1983) are typically applicable in usability studies.

- Concentration reading**
- Flustered*
- Stressfulness*
- User-friendliness*
- Liked using#
- Enjoyment*

Quality

A disposition to reuse (in discretionary-use systems) is indicative of a quality and usable interface (Cooper, 1999) and was hoped to indicate potential to use in real-life, behavioural intent (BI), one of the key adoption concepts. If users perceive that much improvement to the interface is needed (Bevan, 1995), this would be a potential concern. Thus these final interface quality attributes are also included:

- Would use again*
- Improvement needed*

3.2.1. Web Usability Attitude Statements

This resulting attributes were carefully formed into statements for the proposed Web usability questionnaire (Figure 3.2). The usability attitude questionnaires used in the research experiments are all based on this pilot question set. Questions were randomised to be in a different order in each case of presentation to avoid any contextual effects.

The attitude questionnaire formed the basis of the subjective dimension of usability (the users' perceptions of their experience). The questionnaire was always completed after direct experience. When coupled with objective performance metrics (e.g. task completion and errors), qualitative comments and observations, this offers a wide-ranging evaluation of potential usability, ease of use, usefulness and user experience aspects in Web service usage (Hartson, 1998).

Web Usability Statements

I found the page layout on this Web site confusing

I found the layout of the pages on this Web site very clear

The pages on this Web site were attractive

The pages on this Web site were very cluttered

This Web site needs more graphics and pictures

The pages on this Web site were easy to understand

The text on this Web site was too small

I felt that the Web site was helpful

I felt this Web site was reliable

I could quickly find what I wanted on this Web site

I found this Web site trustworthy

When using this Web site I didn't always know what to do next

I found the organisation of this Web site very frustrating

Moving about this Web site was too complicated

I always knew where I was on this Web site

The links on this Web site provided a clear indication of their content

The links I needed were always visible on the screen

Reading the pages on this Web site took a lot of concentration

I got flustered when using this Web site

I felt under stress while using this Web site

I found this Web site 'user friendly'

I liked using this Web site

I did not enjoy using this Web site

I felt in control when using this Web site

The options available did not match my expectations

I would not use this Web site again

I feel that this Web site needs a lot of improvement

Figure 3.2. The Web Usability Questionnaire Developed for the Pilot Study

3.3. Formulation of the Preference Metric

In repeated-measures experiments, participants gain experience of several alternative interface designs. One important benefit to this approach is that the options can also be compared. It has been shown that in presenting a variety of alternatives, participants are more likely to explicitly accept or reject certain designs (Tohidi et al, 2006). This is thought to be due in part to the ability to compare and contrast aspects of designs. However, it has also been suggested that the exposure to various options reinforces the idea that the design is not fixed, and that participants input will be used to inform the process and hopefully in refining the most appropriate design towards production.

There are several different methods of determining user preference. The most behavioural method is to allow participants to choose to use various interfaces or interaction methods (search vs. browse, etc.) and log their choice, or the amount of time spent using each. However, this method is not compatible with performing controlled experiments which desire each participant to have a comparable experience to reduce the influence of confounding variables. Instead, participants in a controlled, repeated-measures experiment could be asked to give a preference amongst several interfaces they have used. This method produces a count for each interface option. Another approach would be to request a full ranking of all the experienced interfaces, giving a weighted approach to preference measurement and obtaining a proportional tally for each alternative interface.

In this work, a related approach was used which ascertains an overall quality and preference rating for each interface (specifically a rank and distance measure). Preferences were recorded using a 0-30 point linear scale labelled "Worst" at 0-cm and "Best" at 30-cm with a marker relating to each alternative interface experienced. The markers are all placed on the scale simultaneously at the end of the experiment to make comparative judgements of alternative interface quality, as well as an inferred preference ranking.

The preference metric provided the following data:

- An overall rating score (to a maximum of 30) for each interface.
- The relative score difference between pairs of interfaces.
- A rank order of preference.

The preference metric is illustrated in Figure 3.3, where A, B and C are examples of interfaces experienced, to be rated on the scale. Using letters instead of naming interface options they experience avoids leading participants.

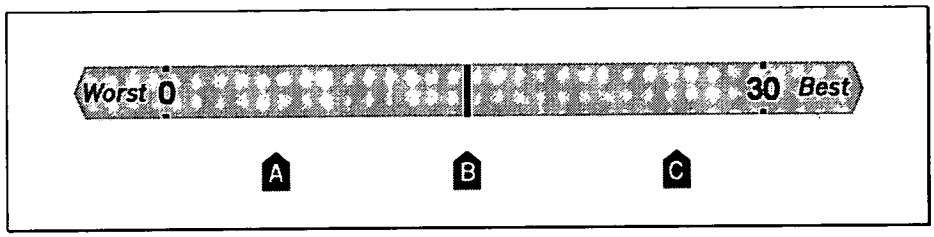


Figure 3.3. Illustration of the Preference Metric

3.4. Qualitative Usability Data

Quantitative data can be compared statistically and provide robust evidence. However, a numerical approach alone cannot provide insight into solutions or assist in making redesign recommendations. Some form of qualitative data collection is typically included in a usability evaluation as the main aim of the project is usually to feedback into the design cycle to optimise the more successful interface. For example, participants may be asked to talk through their interactions with the system to give clues about what they are thinking, expecting to happen or looking for at each point – a process known as *think aloud.*

The think aloud approach has drawbacks as some participants find it difficult to verbalise their thoughts, or explain what they are doing. Some researchers have suggested that it may even alter their behaviour. Certainly, different individuals will likely have varying abilities to concurrently 'think aloud' and as such it is very likely to interfere with collecting precise timing information (Shneiderman & Plaisant, 2005; Van Den Haak et al, 2003; Dumas & Redish, 1993; Karat, 1988). There are also questions about whether concurrent or retrospective reports are equivalent, and which is more appropriate (Van Den Haak et al, 2003; Teague et al, 2001; Dumas & Redish, 1993). When considering the approach, it may depend on the stage of development, the novelty of the interface, the importance of evaluating speed rather than investigating user behaviour. Depending on the situation, the evaluation can either concentrate on the collection of robust quantitative measurements for statistical analysis or focus on building a picture of how users and groups approach interfaces and options, whilst in some cases both types may be collected concurrently.

For the experiments described here, the evaluations occurred very early in the design process. They aimed at informing strategic directions for interface design – getting the right design (Tohidi et al, 2006) which would then be further developed, tested and iteratively improved towards eventual rollout. At this stage of a design process, the need for qualitative remarks about the interfaces, task domains and options available outweighed the requirement to collect precise timing data.

Qualitative comments can offer more insight into how users perform novel tasks (Molich et al, 2004), such as the self-service banking tasks much of this research will be investigating. As understanding of the task domain increases, better quality recommendations for design changes or innovative solutions can typically be generated.

Each participant received the same instructions for the experiment and the think aloud protocol. This maintains experimental control. The introduction to experiment facilitation and the verbal protocol used in the experiments is presented in Appendix C (adapted from Boran & Ramey, 2000; Dumas & Redish, 1993), this includes some prompts designed to encourage participants to talk during their experiences and which were to be used consistently by facilitators throughout the experiments. The facilitators were all trained and experienced in conducting usability experiments, see Appendix C, p.290 (Kuniavsky, 2003; Dumas & Redish, 1993).

In the final stage of an experiment session, a structured one-to-one interview can be used to debrief participants on their experience. Questions are ordered carefully to avoid contextual effects (Oppenheim, 1992) and cautiously worded to avoid bias or careless prompting (Moore, 2001). If the order and wording of the questions are the same for all participants, it is more likely that they will be responding to the same stimuli. Then different answers to the same questions will indicate real differences between the participants, rather than reflect differences in the way questions are phrased (Guttman & Suchman, 1967; Moore, 2001). The responses of different participants and participant groups can be compared – and in the case of closed questions, non-parametric statistical analysis can be performed.

Direct observations about the behaviour of the participants can also be collected during experimental evaluations (McGrath, 1995). Outward expressions of frustration, pleasure and detailed notes are taken about errors, or any other incidences where the participant appeared to be struggling or slow to complete a task. These observations can highlight problem areas as well as where interfaces perform very well. This type of data provides insight that can help explain how people interact with various computer systems enabling redesign to focus

on target areas as well as helping to formulate design recommendations (McGrath, 1995; Whiteside & Wixon, 1985). These qualitative techniques are all highly useful in practical usability evaluation, where the search for evidence is combined with a need to make changes and iterate the interface toward production. By considering both the quantitative scores with qualitative remarks redesign options and solutions can be offered.

3.5. Performance Metrics

Performance is an objective measure of usability. Efficiency and effectiveness are part of the ISO definition of usability and are traditional dimensions of evaluation. The theory being that better usability should correspond with more efficient task completion times. Similarly, task performance (completion and error rates) is reflective of system usability. Therefore the timing of tasks or subtasks, completion and error counts are typically recommended (Dumas & Redish, 1993).

Efficiency has great weight in the provision of technology in a mandatory situation, for example an office program for employees or call centre software. An eBanking application, or a Website offering information about Bank products and services, is substantially different from work-oriented software in this respect. Efficiency is a key driver of customers to online services, including eBanking and Bank Web sites (Weir et al, 2005; Dandapani, 2004; Wright, 2002). However, it should be noted that efficiency savings gained by using the Internet are often harder to gauge. In terms of eBanking and eCommerce, efficiency (and convenience) is high compared to visiting a bank or shop (queuing, arranging appointments) as information is available 24 hours a day, 365 days a year with the click of a mouse from any computer. This factor is potentially as important as time-on-task efficiency measures in these contexts (Helander & Khalid, 2000).

Due to participants in the evaluations being required to use a number of different interfaces, usually for the first time, and in fairly short hands-on sessions (typically 10-20 minutes per interface was timetabled), timing their performance was not appropriate. In such cases, timing data cannot be generalised as it is typically subject to a distinct learning curve (e.g. Weir et al, 2005; Ziefle, 2002).

Therefore the performance usability metrics designed for these experiments focused on task completion, errors, logs of user behaviour (clicks, paths etc.) and verbal reports as objective and subjective indicators of efficiency, rather than recording precise timing data. Although

less robust than timing data, they do highlight efficiency as defined more broadly as time and effort savings compared to more traditional commerce, product comparison and transaction tasks.

Performance measures collected in this research were therefore confined to those which relate to:

- Whether the task can be performed by self-service (without out of channel assistance, like calling a Helpdesk).
- Whether errors are made, noticed and corrected (or unnoticed and potentially troublesome, depending on the error and any subsequent action required correcting it).
- Deviations from optimum paths, lengthy click-streams.
- Verbal reports of efficiency and effectiveness.
- Perceptions of efficiency and effectiveness as measured in attitude questionnaire statements (e.g. speed to find, helpfulness).

3.6. Web Usability Metrics Summary

The basic metrics proposed to measure Web usability in the thesis consisted of a combination of qualitative and quantitative, subjective and objective measures:

- Task completion
- Interaction logs
- Usability (attitude) questionnaire
- Preference (quality) ratings
- Think aloud comments
- Observations
- Interview responses

3.7. Relationships between Alternative Metrics

Difference aspects of usability are often competing for optimisation in a design, and usability engineering is concerned with making appropriate trade offs. When considering a quality such as usability, which is thought to involve several dimensions, it is important to examine relationships between the different concepts and understand them in a variety of contexts. This will assist in better understanding what constitutes usability in different applications. Researchers and practitioners need guidance to select appropriate measures to evaluate usability for various interfaces and designs (Hornbæk & Law, 2007).

Various previous studies have computed correlations between common usability metrics, such as errors, time, satisfaction and preferences. A review of these values are summarised in Table 3.1. In addition to that reported in Table 3.1., correlations were also computed between alternative measures of the same usability dimension. Performance measures of task completion and errors had strong negative correlations, - .384 (Sauro & Kindlund, 2005a) and - .411 (Sauro & Kindlund, 2005b), more errors lead to less completion. Satisfaction measures such as attitude questionnaires and preferences had a correlation of .245 (±.281) (Hornbæk & Law, 2007). Considering the confidence interval (CI), this range covers weak negative, weak positive and strong positive correlations.

The literature review suggests that, in theory, strong and positive correlations should be expected. In fact correlations between different types of satisfaction measures (e.g. usability attitude questionnaires and subjective ratings of overall usability or preference) have been used to validate novel satisfaction instruments (Argawal & Venkatesh, 2002). This study reported very high correlations (.71 to .93 across a variety of products and tasks) between their multi-item usability scale based on Microsoft's Usability Guidelines (MUG) and a three-measure score combining ratings for overall usability, overall design and overall experience on linear scales.

There have been attempts to combine usability metrics into a single score (Sauro & Kindlund, 2005b). Using typical usability metrics (time on task, number of errors, completion of task and average satisfaction scores) and evaluating competing products (4) with a sample of 129 users and 57 tasks (Sauro & Kindlund, 2005a), correlations are included in Table 3.1. Similar correlations were reported when extending the study to 238 users, 88 tasks and 6 products (Sauro & Kindlund, 2005b).

r	Effectiveness vs. Efficiency	Effectiveness vs. Satisfaction	Efficiency vs. Satisfaction
Negative (strong)	-	- .348**[d] (Sauro & Kindlund, 2005a) - .334**[d] (Sauro & Kindlund, 2005b)	- .478**[f] (Sauro & Kindlund, 2005a) - .324**[f] (Sauro & Kindlund, 2005b)
Negative (weak)	- .268[a] (Sauro & Kindlund, 2005a) - .297[a] (Sauro & Kindlund, 2005b) -.156**[b] (Frøkjær et al, 2000)	- .25 (Walker et al 1998)	-
Positive (weak)	-	.164 ± .062 .196 (±.184)[d] .243 (±.158)[d] (Hornbæk & Law, 2007)	.196 ± .064 .145 (±.129)[f] (Hornbæk & Law, 2007)
Positive (strong)	.316 (±.070)[c] (Hornbæk & Law, 2007) .517**[c] (Sauro & Kindlund, 2005a) .553**[c] (Sauro & Kindlund, 2005b) .49[c] (Walker et al 1998)	.454** (Sauro & Kindlund, 2005a) .381** (Sauro & Kindlund, 2005b) .44[e] (Neilsen & Levy, 1994)	.309 (±.146)[g] (Hornbæk & Law, 2007)

Table 3.1. Summary of Published Correlations between Common Usability Metrics

Notes:
Strong correlations are where $r = .3$; weak correlations where $r < .3$
Where significance levels were published: ** $p < .01$
Hornbæk & Law, 2007 reported values of r and the 95% Confidence Interval (CI) of the correlation value

Metric types:
[a] Completion vs. Time
[b] Time vs. Quality of solution
[c] Errors vs. Time
[d] Errors vs. Satisfaction
[e] Errors vs. Preference
[f] Time vs. Satisfaction
[g] Time vs. Preference

Another study evaluating information Websites using a subjective attitude questionnaire found that Relevance correlated highly with Comprehensiveness ($r = .80$). The Relevance category contained questions relating to usefulness while Comprehensiveness concerned content sufficiency and precision (Elling et al, 2007), suggesting these different aspects of usability are also highly associated.

Other researchers who have studied performance and perception rankings have found no correlation between the two metrics suggesting that users find it hard to predict how well they will perform with an interface (e.g. Dillon, 2002). Studying Web (homepage) design, significant correlations were found between perception rankings and the presence of common elements (e.g. email addresses, external links, graphics etc.), $r = .95, p < .01$ (Dillon and Gushrowski, 2000), again highlighting the importance of meeting user expectations.

A frequently cited work in addressing the range of metrics required to conduct usability evaluations found negligible correlation between task completion time and solution quality (efficiency and effectiveness) in information retrieval tasks (Frøkjær et al, 2000). The correlation reported between time and quality of solution was very highly significant (see Table 3.1), however it corresponded to less than 2.5% (R^2) of the variance in one being explained by the other. The study also noted that satisfaction was not simply correlated with performance measures although no values were reported.

In a meta-analysis of 57 published HCI studies, a positive association between preference and performance was found in most cases (Neilsen & Levy, 1994), see Table 3.1. The correlation between objective performance and subjective preference suggests that users' preferences may be reasonably successful in selecting an appropriate interface at an early stage. However the authors caution that there were still many cases where preferences were towards interfaces where performance was measurably worse. In fact, for 25% of the cases, users did not prefer the more efficient system (Frøkjær et al, 2000).

Bailey reviewed seven studies which also illustrated the potential mismatch between performance and preference (Bailey 1993). One example cited was a study of accessing information via hypertext or on paper (hard copy), which found that people were faster using paper, but much preferred the hypertext versions (Nelson & Smith, 1990). In other examples, users had strong preferences even when no performance differences between various treatments were apparent. In comparison he reported on one study showing that user performance mirrored preferences.

Bailey went on to report on two studies conducted to assess this relationship. In the first, participants were asked to compare four interfaces to create a shopping list. They selected which interface they thought would be fastest, however that design did not turn out to offer fastest performance times. In the second experiment, participants compared four interface widgets. Again, participants did not select the fastest (objectively measured) option when asked to rate which would be fastest (in advance of use). The studies conclude that users were unable to accurately assess their performance with different interfaces by subjective means (Bailey, 1993).

Bailey's paper is widely cited in regard to the assumption that preference does not always follow performance. However, the participants in the experiment were actually computer professionals and interface designers, not typical users which may mean that their results do not generalise.

Similarly, various studies of adoption and IT acceptance have published correlations between the different constructs measured. The relationships between common metrics in the TAM are summarised in Table 3.2.

Only one study logged actual usage (Szajna, 1996), the others relying on self-report measures. In this study, the results from actual use logs showed generally lower correlations than self-reports. In fact the correlation between self-report & actual logged usage was only weakly positive, .26 ($p < .001$). Similarly, BI and actual usage logs were reported at $r = .29$ and $r = .25$; PU and actual usage logs at $r = .12$ and $r = .06$; and PEOU and actual usage logs, $r = .22$ and $r = .16$ (Szajna, 1996).

Less commonly measured in technology acceptance research is the construct of satisfaction, although this item was measured in several studies using different modifications of the basic TAM. In addition to the correlations reported in Table 3.2, the following correlations between satisfaction (attitude) measures and self-report usage have been published:

- $r = .28$ (Baroudi et al, 1986);
- $r = .12^{**}$ and $r = .24^{**}$ (Igbaria et al, 1994)
- $r = .19^{*}$ (Roberts & Henderson, 2000)
- $r = .39^{**}$ (Igbaria & Tan, 1997)

Finally, the correlation between satisfaction (attitude) and PU was also reported at $r = .30^{**}$ (Roberts & Henderson, 2000).

r	PU & SR Usage	PEOU & SR Usage	BI & SR Usage	PEOU & PU	PU & BI	PEOU & BI
Positive (weak)	.09 (Szajna, 1996) .26** (Pikkarainen et al, 2004)	.25 (Davis, 1989) .27** .12 (Davis et al, 1989) .14 (Szajna, 1996) .09 (Pikkarainen et al, 2004)	.28* (Szajna, 1996)	.25 (Davis, 1989) .10 .23** (Davis et al, 1989) .29** (Yi & Hwang, 2003)	-	.29* (Szajna, 1996)
Positive (strong)	.56** .68** .63** .71** .59** .85** (Davis, 1989) .65** .70** (Davis et al, 1989) .38** (Szajna, 1996) .33** (Roberts & Henderson, 2000)	.32** .48** .45** .47** .59** (Davis, 1989) .32** (Szajna, 1996)	.35 .63 (Davis et al, 1989) .57** (Szajna, 1996) .44 to .57 (Venkatesh & Davis, 2000) .53** (Sheppard et al, 1998)	.56** .69** .64** .38** .56** (Davis, 1989) .48** .30* (Szajna, 1996)	.72** .31* (Szajna, 1996)	.40** (Szajna, 1996)

Table 3.2. Summary of Reported Correlations between Commonly Used TAM Metrics

Notes:
Strong correlations are where $r = .3$; weak correlations where $r < .3$
Where significance levels were published: * $p < .05$, ** $p < .01$
SR = Self Report measure of usage

3.7.1. Piloting the Metrics

The usability and preference metrics proposed for this work were assessed in two pilot studies examining different Web site designs for accessing general information on banking products and services. The pilot studies explored empirically the relationships between subjective and objective measurements of usability. Thus the suitability of the questionnaire in selecting the most appropriate interface design for the content could be determined and the expected relationship between usability and preference studied. Results allow further understanding of the role of different usability metrics in interface design.

Chapter 4. Pilot Studies of the Usability Metrics

The construction and validation of the usability attitude questionnaire was the focus of the pilot investigations. By comparing questionnaire results to the other collected metrics, the pilot studies ensure that the attitude statements proposed for the Web usability questionnaire were appropriate for evaluating the Web medium. The pilot tests also examined the additional metrics of task performance, the preference rating and use of the think aloud protocol. The results demonstrate the relationship between different aspects of usability for the general context of information-retrieval in a public-access Website. Finally, they begin to provide guidance as to what constitutes a usable and potentially preferred Website.

4.1. Pilot Study 1: The Usability of Web Portals

The first pilot study involved comparing the usability of two different Web Portal designs for the Case Bank: A and B. The designs provided access to the same range and type of information. However, the characteristics of the two designs were different, as shown in Figures 4.1 – 4.4. The most notable differences were in terms of:

- The Information Architecture (*IA*) – such as the use of A-Z index pages or hierarchical structure.
- The amount of graphical content (*Clutter*) – such as graphical banner adverts and non-informational content (pretty pictures).
- The placement of menus (*Persistence)* – such as the use of fixed navigation options situated in frames.

Design A was characterised by indexes, low clutter and persistent menus. Design B was hierarchically categorised, cluttered and without persistent menus. The aim of the experiment was to compare the two Web Portal designs in terms of usability and preferences of Bank customers and non-customers in information-seeking tasks.

4.1.1. Hypotheses

Data were collected to test the null hypotheses. Statistical analysis using the analysis of variance (ANOVA) was performed to examine how the various controlled factors attributed to performance and attitude measures. The level of significance was selected as .05 in a two-tailed test.

Pilot Study 1: The Usability of Web Portals

The null hypotheses tested were:

Hypothesis H_0 P1a: The different Web Portal designs will not result in different usability attitude and performance scores.

Hypothesis H_0 P1b: The different Web Portal designs will not result in different user perceptions of preference.

Hypothesis H_0 P1c: There will be no relationship between the usability measures of performance, attitude or preference.

4.1.2. Participants

The participants in the experiment were recruited to be computer savvy, having previously used a computer at home or work. This criterion ensured the sample would model current and future users of the Internet who had the potential to look for banking information online. By recruiting computer-savvy participants, it was ensured that the usability experiment sessions focused on the competing interface designs, rather than the use of computers in general. Participants were asked whether they recognised or thought they may have used the Web sites before (after each experience), none indicated that they had.

All participants undertook a set of defined tasks on mirror copies of two Web portal designs during the experiment. Participants were grouped by gender and age group, with two age groups selected: 18-29 and 30 years and over. These categories were selected based on previous studies and data about the typical ages of Internet users and the range of willing participants available during recruitment (National Statistics, 2006; Nichols et al, 2001). Grouping variables of age and gender were balanced in the experimental design along with the order of experiencing the two interfaces.

4.1.3. Tasks

The tasks were devised by considering what information real-users might seek online, constrained only by the type of information available on both sites. Two task sheets were created and matched in terms of type and involvement with the interfaces. Each task sheet contained four information retrieval (single-fact) tasks and one comparison of fact question (Spool et al, 1997).

For single-fact retrieval questions, e.g. "How many Cashpoint machines are there in the UK?" participants were asked to report back their findings to the facilitator.

The comparison of fact questions required participants to search for more than one piece of information, compare and come to a conclusion between them. These tasks focused on the comparison of rates on different accounts and also required the participants to report back on their decision.

Participants were encouraged to try each task in the order presented; however, they were able to give up on individual tasks, skip or return to any task. In some cases, if a participant was observed to be struggling, facilitators were allowed to intervene and suggest that they move on to the next task, reassuring them that it was the Website being tested, not the participant themselves. A list of approved remarks was offered to the facilitators to ensure that such prompts were well-phrased and controlled; these are shown in Appendix C, p.290

The tasks were tested on several test participants to ensure they were acceptable. Full details of all the tasks are available in Appendix D, p.297.

4.1.4. Dependent Variables

As utility in both Web Portals was comparable, usability was measured using performance and attitudes. Other data collected included comments, observations, comparative preference scores and interview question responses.

Usability – Performance

The performance component of usability was measured using a tally of task completion. Single-fact retrieval tasks were counted as complete if the appropriate fact was correctly found and reported to the facilitator. For the comparison of fact questions, completion was counted if the participant used at least two facts in drawing their conclusions. Task

completion was only recorded as a binary response – complete or failed. Pages visited and positions of links clicked were also logged automatically.

Usability – Attitudes

The attitude toward usability was measured using the Web usability questionnaire previously described. After using each design, facilitators introduced the questionnaire: "This questionnaire relates to the Website you have just used. Please tick the box which most closely represents how you feel about each of the following statements".

Success – Preference Rating

As a final quantitative measure, participants were asked to rate the two interfaces on the preference rating scale (0-30 points). This occurred after both hands-on sessions such that both experiences could be compared. This rating was also reduced to a rank order of preference.

Qualitative Measures and Interview Questions

Observations of the types of difficulties, errors or problems encountered (including technical difficulties) were taken during the hands-on sessions as well as any comments resulting from the 'think aloud' protocol. Additional qualitative data were also collected in a debriefing interview: participants were asked about what they liked, disliked and what suggestions for improvements they could offer.

As a final question, participants were asked whether they would be encouraged to use the Internet to obtain information on banking and financial services (see Appendix D).

4.1.5. Experiment design

The pilot experiment investigated two alternative Web Portal designs using a repeated-measures, within subjects design. A sample of 24 participants was recruited, balanced for age, gender and presentation order of the interfaces. The dependent variables were the responses to individual items in the usability questionnaire, the preference rating and deduced rank order of preference. Task completion and verbalisations were also recorded.

4.1.6. Experiment Design Summary

Pilot Study 1 – Usability of Web Portals

Design: Two cell, repeated measures, within subjects

Independent Variables: Design A (Indexed)

Design B (Cluttered)

Participant Independent Variables: Age (2 levels)

Gender (2 levels)

Order of experience (2 levels)

Dependent Variables: Attitude questionnaires

Task completion

Quality ratings & preferences

Confounding variables: Researcher Bias (randomised)

Room (randomised)

Task sheet (balanced & matched task per task)

Other data: Think aloud remarks

Researcher observations

Interview questions

Sample size: 2 orders x 2 genders x 2 age groups x over-sampling 3 = 24

Honorarium: £20

Session Time: 1 hour

4.1.7. Interfaces

Figures 4.1 and 4.2 illustrate the general look and feel for Web Portal Design A (Indexed).

Figure 4.1. Homepage of Design A (Indexed)

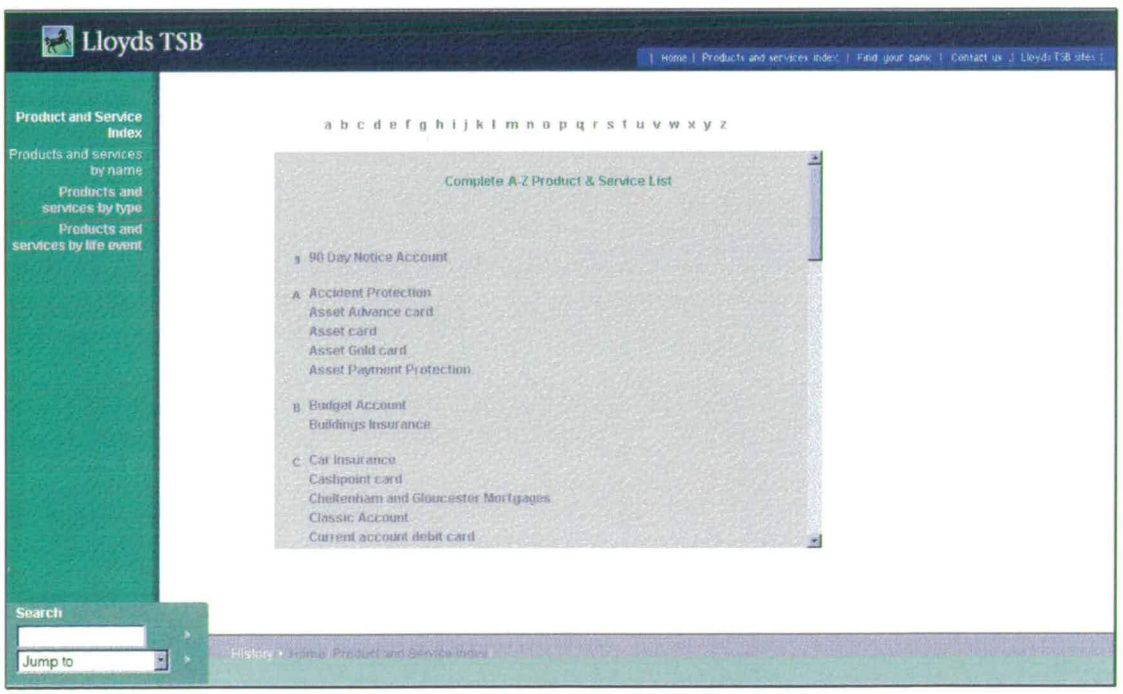


Figure 4.2. A-Z Index Page of Design A (Indexed)

Figures 4.3 and 4.4 illustrate some typical pages from Web Portal Design B (High clutter).

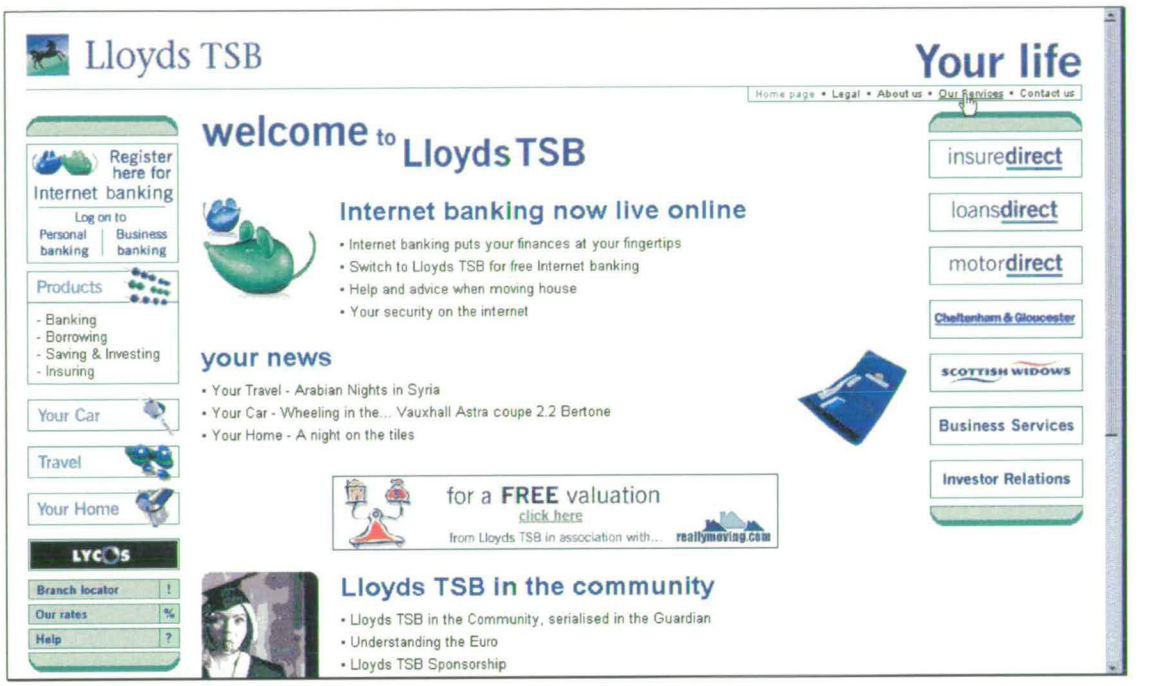


Figure 4.3. Home Page of Design B (Cluttered)

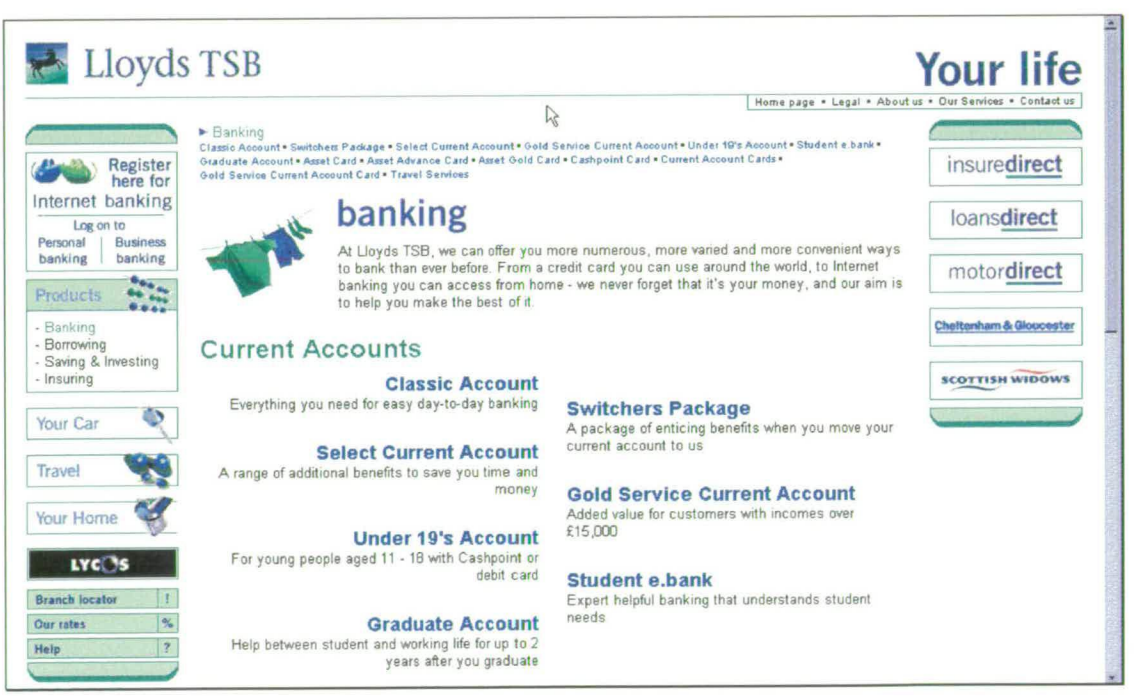


Figure 4.4. Category Page of Design B (Cluttered)

4.2. Pilot Study 1: Results

4.2.1. Participants

The sample was balanced in terms of gender: 9 (37.5%) were under 30 years old, the other 15 (62.5%) were 30 or older, with the oldest participant over 70 years old. Most volunteers were aged between 20 and 40 years, therefore the 30 year split was considered appropriate although the groups were not well matched in size. The order of experience for the two interfaces was also balanced overall, and across the participants of differing ages and genders as much as possible. Participating users made frequent use of the Internet with 11 (45.8%) using it daily.

Some 8 (33.3%) participants used eBanking services to manage their accounts online. Additionally, 4 (16.6%) of the sample were customers of the Case Bank (approximately representing the market share). A further 16 participants (66.6%) had used the Internet for shopping purposes.

4.2.2. Performance

Task completion data was summed for the five tasks and recoded into a measure of performance as a rate of completion. For this experiment, task completion rates were fairly moderate, as shown in Table 4.1. These scores indicated there was scope for performance improvement in both Web Portal designs.

Design	Performance	St. Dev.[a]	N[b]	Lower Bound CI[c]	Upper Bound CI
A (Indexed)	3.500 (70.0%)	1.216	24	2.9866 (59.7%)	4.0134 (80.3%)
B (Cluttered)	3.542 (70.8%)	1.103	24	3.0761 (61.5%)	4.0072 (80.1%)

Table 4.1. Task Completion Sum (Rate) for the Alternative Web Portals

Notes:
[a] Standard deviation
[b] Number of participants in the sample
[c] 95% CI of the mean

A paired sample t-test on the two mean completion scores for each interface (no between subject factors) found no significant difference between them, $p = .857$. Repeated-measures

ANOVAs also concluded that there were no age or gender differences in terms of the rates of task completion.

Task completion for both task sheets was very similar for the alternative designs. Overall, the two task sheets were equally distributed amongst the two designs and the participants, balancing any effect of task difficulty within the different designs. Most participants made a good effort to find information, only giving up if after much searching they could not find what they wanted.

4.2.3. Attitude

Balance was not completely achieved between males/females and participants of differing age groups resulting in insufficient numbers in the between-participant groups to include all three variables in the analysis. Order effects were first used as the between-participant variable (in combination with both age and gender separately). There were no order effects; therefore, order was eliminated from the rest of the analysis. Similarly, the task sheet variable was also tested, found to have no effect and eliminated from subsequent analysis. The main analysis presented here focused on the between-subject factors of age and gender, Table 4.2.

There was a main effect between the two designs in terms of usability with Design A outperforming Design B in terms of mean usability questionnaire scores $F(1, 20) = 5.938$, $p = .024$ (see Appendix A, p.279). No within subject or between subject gender or age group effects were found. In fact, Design B obtained an overall mean score of less than 4 on the 7-point scale, indicating generally low attitudes to a range of usability considerations.

Design	Mean Score	St. Dev.	N	Lower Bound CI	Upper Bound CI
A (Indexed)	4.684	1.063	24	4.2346	5.1326
B (Cluttered)	3.944	.909	24	3.5607	4.3282

Table 4.2. Usability Questionnaire Means for the Alternative Web Portals

Despite very similar task performance rates using both designs, there were significant attitude differences. The mean usability questionnaire scores also indicate substantial room for improvement in both designs.

The analysis of individual attributes focused on age and gender differences as between-subject factors. The main results are presented in Table 4.3. The individual scores are illustrated in the chart, Figure 4.5., where higher scores indicate better usability (with negative statements polarised, see p.47). Generally, the figure shows that on a wide range of interface characteristics, Design A obtained higher attitude scores than Design B. Although Design B was evaluated more positively in terms of the graphical and pictorial content, all the remaining characteristics were in favour of the Design A. Stars indicate significant differences between scores (* $p < .05$ and ** $p < .01$).

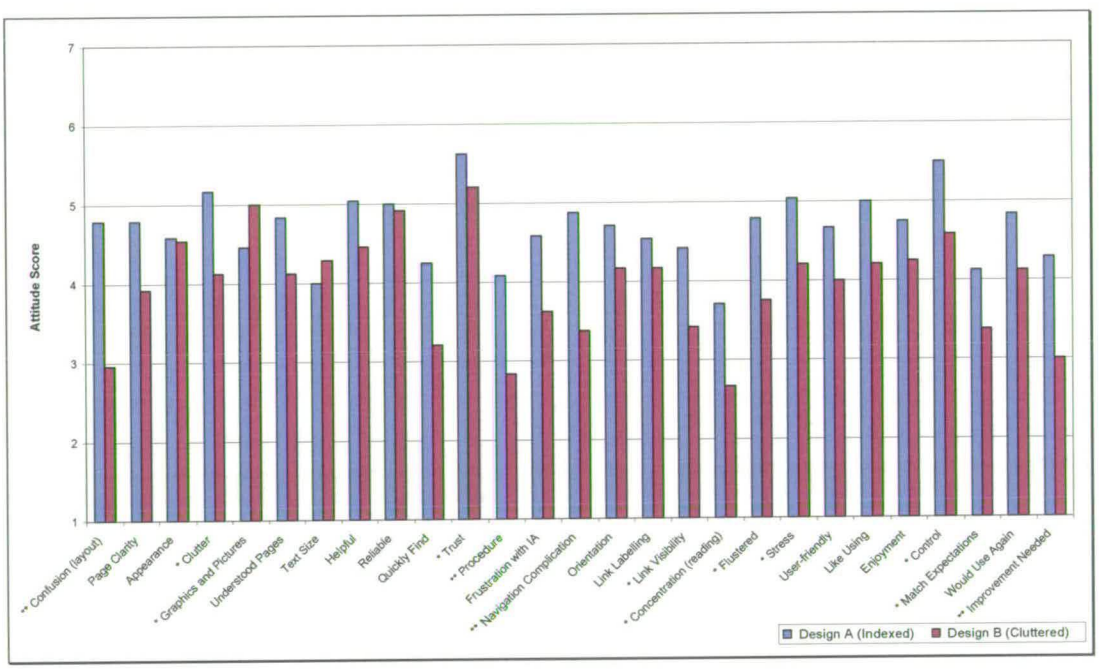


Figure 4.5. Comparison of Usability Attributes for the Alternative Web Portals

Some thirteen of these comparisons were significantly different favouring the original design, with four of those being highly significant with $p < .01$; these were layout, navigation, procedures and the need for improvement. Also significantly higher scoring for the original design were being in control of the interaction, having to concentrate less on reading the material, feeling less flustered and stressed during the interaction, matching expectations of the Website, promoting better trust in the interface, having higher link visibility and a less cluttered interface. Design B had much room for improvement; in particular the lowest scores suggest that redesign efforts focus on the structure, organisation and layout of the site content (IA), reducing clutter and increasing link visibility.

Attribute	Main effect (p)	Comparison	W-S Inter[a] (p)	B-S Effects[b] (p)
Confusion (layout)	.001	A > B	-	-
Procedure	.005	A > B	-	-
Navigation complication	.007	A > B	-	Age/Gender .012 (F1 > others)
Improvement needed	.007	A > B	-	Age/Gender .053 (NS) (F1 > M2 > others)
Flustered	.010	A > B	-	-
Pictures	.018	B > A	-	-
Link visibility	.019	A > B	-	Age/Gender .014 (F1 > M2 > others)
In control	.020	A > B	-	-
Match expectations	.028	A > B	-	-
Concentration (reading)	.033	A > B	-	Age .014 (Y > O)
Clutter	.039	A > B	-	-
Stress	.040	A > B	-	-
Trust	.048	A > B	Gender .025 (F: A > B; M: B > A)	-
Liked using	.057 (NS)	A > B	-	Age/Gender .026 (F1 > others)
Frustrating organisation	.058 (NS)	A > B	-	Age/Gender .050 (F1 > M2 > others)
Quickly Find	.068 (NS)	A > B	-	-
Page clarity	.074 (NS)	A > B	-	-
Enjoyment	-	-	-	Age/Gender .002 (F1 > others)
Orientation	-	-	-	Age .006 (Y > O)
Text size	-	-	-	Age .021 (Y > O)

Table 4.3. Interactions & Between-Subject Effects on the Individual Attributes

Notes:

[a] Within-subject interactions with the interface design)

[b] Between-subject effects (regardless of interface design

NS Not significant, but a marginal p value (where p = ~.075) indicating a possible trend

F = Female; M = Male; Y = <30 years age group; O = 30+ years

M1 = Male, <30 years; F2 = Female, 30+ years; similarly for F1 and M2.

In terms of between-participant variables, age and an age/gender combination were the most differentiating characteristics of the user population. Distinct differences between older and younger participants were apparent in terms of their attitude towards text size, concentration reading and orientation within the information hierarchy. In each case it was younger participants whose attitudes toward these factors were higher than the older age group.

An age/gender effect in general attitudes toward using and enjoyment showed younger females were more positive than other age/gender groups. A similar pattern is noted in the navigation complications, with younger females less critical than other groups. For three further aspects (IA, improvement and link visibility) young females remained the most positive scorers, but older males also showed more positive results than young males and older females.

Trust attitudes were apparently significantly different between the two interfaces $F(1, 20) = 4.422$, $p = .048$, with Design A promoting trust slightly better than Design B. However, an interaction with gender, $F(1, 20) = 5.841$, $p = .025$ indicates that it was female participants who awarded significantly higher scores to Design A. As shown in Table 4.4, male participants actually gave slightly higher responses in terms of trust to Design B, although not very different from their trust scores for Design A. Women apparently had quite different impressions of trust from the two designs, where men did not. This indicates that trust perceptions formed by men and women may be influenced by different stimuli, although more research would be needed to explore and confirm this theory.

Design	Gender	Mean Score	St. Dev.	N	Mean Score	St. Dev.
A (Indexed)	Female	5.943	.231	12	5.565	.167
	Male	5.188	.242	12		
B (Cluttered)	Female	5.043	.309	12	5.146	.223
	Male	5.250	.323	12		

Table 4.4. Gender Differences in Trust Attitudes towards Alternative Web Portals

Most of the age and gender results were consistent on both interfaces, suggesting that age and gender differences may be general in the population of Internet users. Age and gender characteristics appear to be important in considering Banking Website usability. However, there may be other influencing factors such as Internet experience, experience and knowledge of the banking domain. These are possible confounding issues which were not

factors in this pilot study, but may be important to consider in a larger sample where such factors can be measured and balanced.

4.2.4. Preference Ratings

The preference ratings were analysed using order of experience as the between-subject factor. Order had no effect and was eliminated from further analysis. Age and gender were then included as the between-subject factors. The preference scores are presented in Table 4.5.

Design	Mean Score	St. Dev.	N	Lower Bound CI	Upper Bound CI
A (Indexed)	18.77	6.846	24	15.88	21.66
B (Cluttered)	13.00	7.077	24	10.01	15.99

Table 4.5. Mean Preference Scores for the Alternative Web Portals

There was a main effect between preference for the two Web Portals, Design A outperforming Design B in the overall rating, $F(1, 20) = 6.619, p = .018$. The within subject age/gender effect was significant at $F(1, 20) = 4.781, p = .041$, but there were no between subject gender or age group effects.

In fact, Design B obtained an overall mean score of less than the midpoint of the preference scale, indicating generally low perceptions of overall quality. However, Design A didn't perform much better, with the lower bound of the 95% CI only just above the midpoint.

Preference ratings for both designs followed very closely the pattern seen in the usability questionnaires.

4.2.5. Preference Rankings

Reducing the preference ratings to a rank order, Design A was clearly preferred (Table 4.6). By removing two participants[5] with no difference in rating scores, a binomial test on the dichotomous choice between the two designs remains. The binomial test, assuming both

[5] Although a Chi-squared test can be performed on the full data set, if the number of participants in each group is less than 5, the results loose statistical power, therefore reducing the set to those who made a choice provides a more accurate and reliable computation (Howell, 1997).

portals had equal chance of being selected as preferred, shows a Design A was significantly favoured, $p = .017$, N = 22.

Preferred Design	N	%
A (Indexed)	17	70.8
B (Cluttered)	5	20.8
No Preference	2	8.3
TOTAL	24	100.0

Table 4.6. Preference Rankings

4.2.6. Intention to Use Online Financial Portals

One question in the interview ascertained potential usage intentions by asking whether participants would be encouraged to use Web sites such as these for financial information. At this stage of the research, participants were not asked to rate this for each different design, just for the general idea. The results showed that 10 (41.7%) of participants responded positively, 7 participants (29.2%) responding negatively and a further 7 participants uncertain.

Generally these usage intentions were fairly moderate. There was scope for higher intention in the domain given the room for usability improvement highlighted by the study; metrics of performance, attitude and preferences all indicating general and specific design problems.

4.2.7. Qualitative Data

Comments and observations about Design A suggested that the navigation options and ease of finding detailed information were its best features. The A-Z index was praised and frequently used but occasionally poor nomenclature in the alphabetic index caused problems. The other criticism was the relatively dull appearance in terms of look and feel. Some links on the page were hard to notice or read due to text size or colour contrast. Participants expected coloured hyperlinks within the text to indicate their presence clearly.

The use of pop-up windows was a confusing aspect of Design A. Although many commented that they liked the windows, others lost them behind the main window. Pop-up windows were used to show rates and charges information for the different accounts, this

made comparisons difficult using this design as the information had to be found separately for the different accounts.

For Design B, there was a general feeling that banking pamphlets had been digitised and placed online, without much thought to cross linking information and with too many graphics and clutter. Many people missed navigation options which were sometimes situated on the right hand side of the screen amongst the banners and subsidiary brand logos. These contextually appearing menus (when noticed) were described as "very subtle". This menu also contained the link to rates and charges. If found this link made an easier task of comparing rates as they were all on the same page although no comparison feature was offered.

Generally, Design B was more confusing to use, with menu positions spread around the left, top-right and right hand sides, with the addition of a top centre list of links that appeared contextually and duplicated the links in the content pane.

4.2.8. Hyperlink and Menu Usage Logs

Logs were examined to determine any patterns in the positions of links clicked. Participants generally used links from the content pane: this accounted for 93% of links clicked using Design A and included the main links on the home page, links within all the index lists and also links on the information pages. This number was reduced for Design B to just 54%, again referring to the main links in the centre of the homepage, products in the category pages and hyperlinks within the content pages. Participants using Design B made use of the left hand menu also, some 23% clicking on these links. The rest of their clicks were the rest spread between the links along the top of the content pane (contextually), top right (persistent) and the contextual right hand side menus. Qualitative comments and observations indicated that this variety of locations to find more information, or related content made navigation in this site a lengthier process. The contextual right hand side menu in Design B was not well used across the cohort, thus some participants missed important links.

4.3. Analysis of the Web-Usability Questionnaire

4.3.1. Questionnaire Reliability

When designing an attitude questionnaire, it is important to investigate the contributions made by individual statements in relation to the averaged overall metric. This is performed using reliability statistics such as Cronbach's alpha (see section 2.3.2). Reliability scores above .8 for Likert attitude scales would be expected.

The overall alpha (α) calculated for usability questionnaire (27 questionnaire items, 24 participants) over the individual interfaces showed high inter-item reliability: .966 for Design A, and .942 for the Design B.

Examining the values of α for each item if it were eliminated from the questionnaire can determine if there are any likely candidates for deletion or whether all the individual items positively contribute to the overall questionnaire mean.

For Design A, only one item (if deleted) results in a higher α, the difference being $\alpha = .967$ removing the item referring to the amount of *graphics and pictures*. Since this is a very small change in α, there is no strong evidence to remove this item.

Similarly, for Design B, there were no strong candidates. For five of the twenty-seven attributes very slight increases were observed: $\alpha = .943$ (removing *reliable*), .944 (removing *concentration reading*, *enjoyment* or *link labelling*) and .945 for *graphics and pictures*.

This is not strong evidence to remove the items. However, the fact that the item *graphics and pictures* was noted in both designs may indicate that this is not as consistently acting as a typical usability trait.

4.3.2. Analysis of Neutral Responses

Another possible indicator of superfluous attributes would be the tendency to obtain many scores of 4 on the 7-point scale, indicating a 'neutral' attitude, or more fully, to 'neither agree nor disagree' with the statement.

Examining the counts of these neutral responses towards the various attributes shows that generally, participants did not select the neutral response. Out of a total of 648 responses to questions for each interface (27 questions, 24 participants), Design A was given a neutral score 13.4% of the time, 17.4% for Design B.

For Design A, the highest frequency of neutral scores (8) was for the attribute *Appearance* (representing one third of the participants). The second highest frequency was for *Reliable* which was scored neutral 7 times (by 29.2% of participants).

For Design B, the highest frequency of neutral scores (9) was for the attribute *Reliable* (representing some 37.5% of participants). The second highest frequency (8) was associated with attributes *Enjoyment* and *Stress*, scored neutral by one third of the cohort.

Overall there is no strong evidence to remove any items from the questionnaire set. In considering the overall totals, irrespective of design, reliability did appear to be one of the harder aspects to form an attitude about. This may be due to the short time frame involved in the hands-on sessions. It may also indicate that reliability was not considered a vital issue by participants. Yet as one of the original question items, it has proven important in other technologies. Further, it represents one of the main elements of the Banking Code, therefore of importance in studies of Banking Websites. For this reason it was not removed.

4.4. Relationships between Metrics

4.4.1. Usability – Attitudes and Preferences

Two key subjective usability metrics were collected: attitude questionnaires and a comparative preference (or quality) rating. Although both metrics represent user perceptions, they were collected at different times, and in different contexts: the attitude questionnaire directly after each experience (and therefore subject to positional effects due to the sequence experienced); preference ratings at the conclusion of the hands-on sessions when both alternative options had been experienced. Whether a relationship exists between these two metrics was of interest to explore.

The differences between the two designs in terms of mean usability questionnaire scores and preference ratings (in the direction of the majority preference, A-B) were computed. Subsequently, the individual designs themselves were considered in isolation.

There was a strong correlation between the differences in usability attitudes and preferences. A Pearson correlation analysis showed a highly significant positive correlation, $p < .001$. These data are illustrated in Figure 4.6, Pearson's $r = .830$, $R^2 = .689$ (a large effect size). The points are also categorised by participants' preference rankings.

For Design A considered alone, the correlation between the mean usability attitude score and the preference rating was also strongly positive, $r = .727$, $p < .001$, $R^2 = .529$ (a large effect size). Similarly, for Design B the correlation between the mean usability attitude score and preference rating was slightly lower, but still strongly positive, $r = .685$, $p < .001$, $R^2 = .469$ (a medium-large effect size).

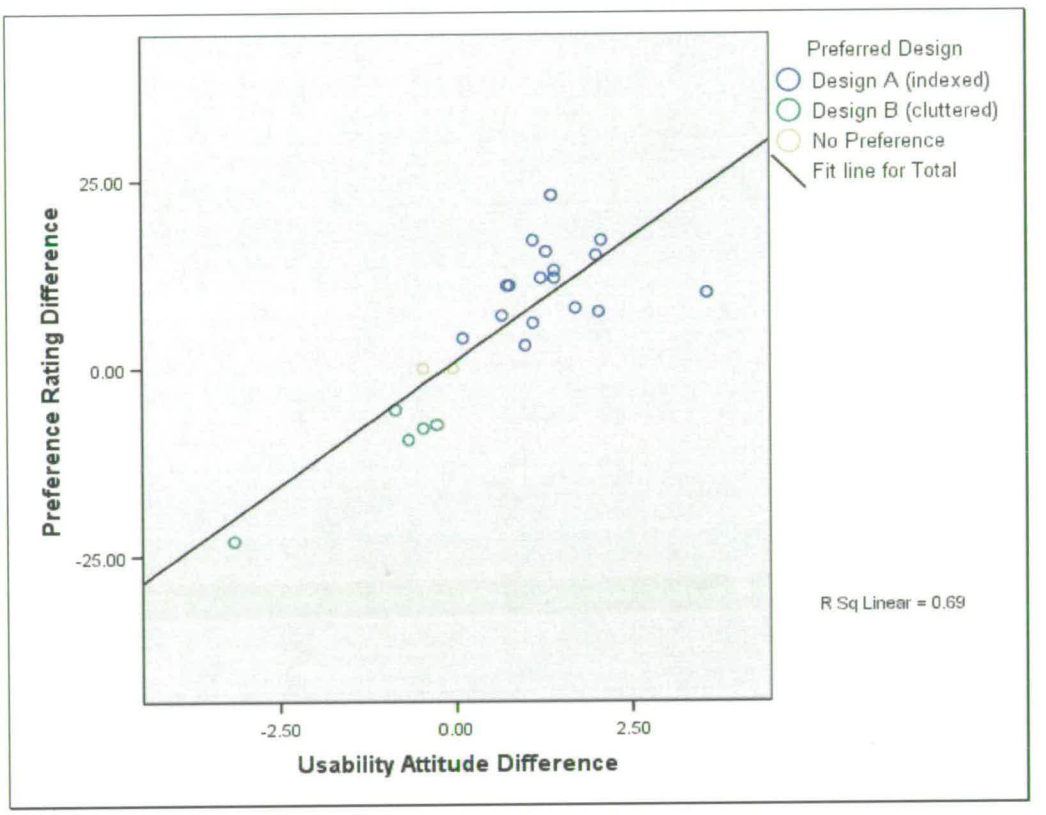


Figure 4.6. Usability Attitude – Preference Relationship (Differences between Designs)

Generally, the correlation appears strong, significant and positive between the mean of the usability questionnaire and the comparative preference rating score. As attitude towards a range of usability factors increase, so does the level of the preference rating. The preference rating in this situation could be considered an overall quality evaluation between best and worst. Then this could be restated as increasing usability attitude being related to increasing quality rating.

Individual Attributes in the Usability Questionnaire

In terms of individual items in the usability questionnaire, it is possible that some usability or interface characteristics are more salient in relation to people's preferences than others. Therefore the correlation matrix of each individual attribute and the final preference rating was computed; again the differences in scores between the two interfaces were considered first, then individual interface scores. The correlation coefficients are summarised in Figure 4.7, ordered by the magnitude of r for the difference between interfaces.

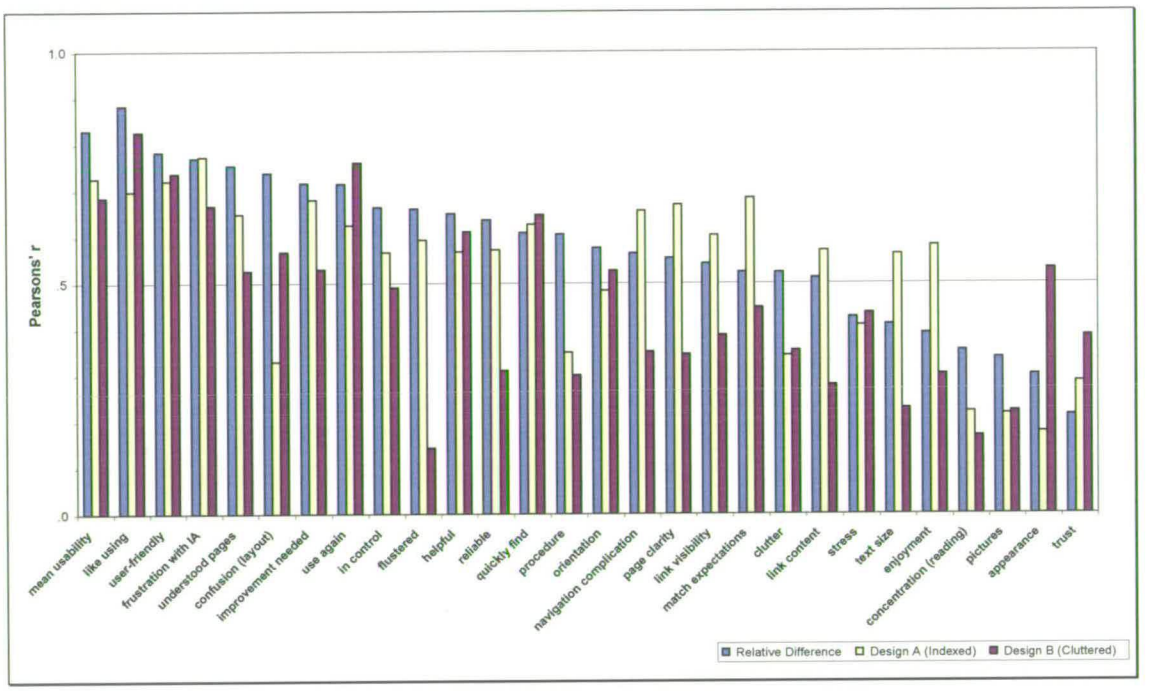


Figure 4.7. Correlations for Usability Attributes and Preference Ratings

As can be seen in the chart, many of the correlations are high (Pearson's $r > .5$), however there are some differences between the two interfaces. The significant correlations are detailed in Table 4.7, where highlighted items indicate strong and significant correlations for all three computations.

Attribute	Difference A-B		Design A		Design B	
	r	*p*	*r*	*p*	*r*	*p*
like using	***.883***	***<.001***	***.698***	***<.001***	***.827***	***<.001***
user-friendly	***.783***	***<.001***	***.721***	***<.001***	***.737***	***<.001***
frustration with IA	***.770***	***<.001***	***.773***	***<.001***	***.667***	***<.001***
understood pages	**.754**	**<.001**	**.649**	**.001**	**.526**	**.008**
confusion (layout)	.738	<.001	-	-	.567	.004
improvement needed	**.716**	**<.001**	**.680**	**<.001**	**.530**	**.008**
use again	***.715***	***<.001***	***.625***	***.001***	***.760***	***<.001***
in control	.664	<.001	.566	.004	.490	.015
flustered	.661	<.001	.593	.002	-	-
helpful	**.651**	**.001**	**.568**	**.004**	**.611**	**.001**
reliable	.637	.001	.572	.004	-	-
quickly find	***.609***	***.002***	***.627***	***.001***	***.648***	***.001***
procedure	.606	.002	.349	.094	-	-
orientation	.576	.003	.483	.017	.527	.008
navigation complication	.564	.004	.656	<.001	.351	.093
page clarity	.554	.005	.669	<.001	.345	.098
link visibility	.542	.006	.603	.002	.387	.062
match expectations	.524	.009	.684	<.001	.447	.029
clutter	.523	.009	-	-	.354	.090
link content	.511	.011	.570	.004	-	-
stress	.426	.038	.408	.048	.435	.034
text size	.410	.047	.562	.004	-	-
enjoyment	.391	.059	.581	.003	-	-
concentration (reading)	.353	.090	-	-	-	-
appearance	-	-	-	-	.531	.008
trust	-	-	-	-	.385	.063
mean usability score	.830	<.001	.727	<.001	.685	*<.001*

Table 4.7. Significant Correlations between Usability Attributes and Preference Scores

Notes:
Blank cells indicate non-significant (nor marginally significant) correlations at $p < .05$.

For information retrieval and item comparison tasks, the attributes most highly correlated with preference included those concerning the speed of finding information, the perception of the IA, page content comprehension and helpfulness. Other more generic usability factors (Marshal et al, 2001) were also salient: whether participants *liked using* the Web site, found

it *'user-friendly'*, thought it *needed improvement* and whether they thought they would *use it again*. As *use again* represented future usage intention, that it highly correlates with the preference (or quality) rating is encouraging in terms of relating intentions and preferences. Choice has been used before as a indication of potential usage behaviour (Szajna, 1994), and these results also show the benefit of using repeated measures evaluations rather than studying Websites in isolation.

Other attributes came close to the $r > .5$ level, and were highly significant at this sample size, these included the degree of control, orientation in the interface, matching expectations and stress.

Attributes which generally did not correlate well with overall preferences included the attitude to graphics and pictures, the appearance of the site (although this did show a strong, positive relationship to preference scores for Design B), and trust in the site. The lack of association between appearance and preference was contrary to recent research has found that people quickly form opinions about Web sites based on their visual appeal (Lindgaard et al, 2006).

Individual correlations all indicate that more positive attitude towards a usability attribute is related to stronger preference or higher overall quality on the preference scale.

In comparison to a recent meta-analysis (Hornbæk & Law, 2007), the magnitude of correlations found for these Banking Websites are generally higher than the upper bound previously published $-.036 < r < .526$ (see p.60). The usability questionnaire appears to be providing a good overall metric which is highly related to participants' preferences and overall quality rating levels for different designs.

4.4.2. Usability – Attitudes and Performance

Correlations between different metrics can also indicate how aspects of subjective and objective measures of usability relate. Correlations between individual attitude questions (and mean scores) and performance (task completion rates) were computed in the same way as for the preference scores. The difference between interfaces on mean usability questionnaire scores correlated significantly with task performance, $r = .438, p = .032, R^2 = .192$. This is a medium to strong positive correlation indicating that as performance increased there was a corresponding attitude score increase. For Design A, $r = .519, p = .009, R^2 = .269$ – another strong correlation; but for Design B, $r = .255, p = .230$ – a weak

and not significant relationship. Participants overall attitudes were related to their performance when considering using Design A and for interface differences, but not so for Design B. For Design B, attitudes were very low (compared to Design A), despite similar and relatively successful performance levels for both designs, this may have contributed to the lack of a strong or significant correlation for this particular Web design.

Individual questionnaire attributes which highly correlate with performance for the difference between interfaces A-B included: Trust, $r = .587, p = .003$; Frustration with IA, $r = .516, p = .010$; Matched expectations, $r = .457, p = .025$; Use again, $r = .442, p = .031$; Link visibility, $r = .410, p = .046$.

For Design A: Flustered, $r = .631, p < .001$; Matched expectations, $r = .620, p = .001$; User-friendliness, $r = .566, p = .004$; Frustration with IA, $r = .538, p = .007$; Quickly find, $r = .526, p = .008$; Link visibility, $r = .494, p = .014$; Trust, $r = .470, p = .021$; Reliable, $r = .465, p = .022$; Understood pages, $r = .440, p = .032$; Enjoyment, $r = .438, p = .032$; Page clarity, $r = .431, p = .035$; Orientation, $r = .429, p = .036$.

For Design B, there was only one significant correlation, as would be expected from the lack of correlation with mean usability scores. The attribute related to performance scores was the degree of Control, $r = .547, p = .006$.

Whilst these scores were higher than published correlations between efficiency and satisfaction ($.102 < r < .454$), there was less agreement between different interface designs.

4.4.3. Usability – Performance and Preferences

The relationship between performance (task completion) and preference is explored in the same way. There have been few published results on this relationship, one in the field of biometric technologies for security purposes (Toledano et al, 2006), a very different application to these Website evaluations.

Table 4.8 displays the correlations for the differences, then each interface individually. There were no significant correlations, although the relationship for Design A is marginally significant. Values are positive, generally supportive of the idea that preference may increase with better performance, however not at the levels of variance seen as attitude scores increase.

	Differences (A-B)	Design A	Design B
r	.237	.377	.150
p	.264	.069	.483

Table 4.8. Correlations between Performance (Task Completion) and Preference

4.4.4. Comparisons of Metrics

The mean usability attitude scores, performance rates and preference or quality ratings are displayed comparatively using normalised scales (as a fraction of the total score possible) in Figure 4.8. The error bars indicate the 95% CI of the means. The chart clearly shows no obvious differentiation between the two interface designs in terms of performance metrics, although less spread in the CI for Design B. However, the subjective measures of usability attitude and preference show clear tendencies for Design A to be favoured. In fact, when it came to the comparative preference metric, the CIs do not overlap, indicating very definite differences in responses to the two interfaces.

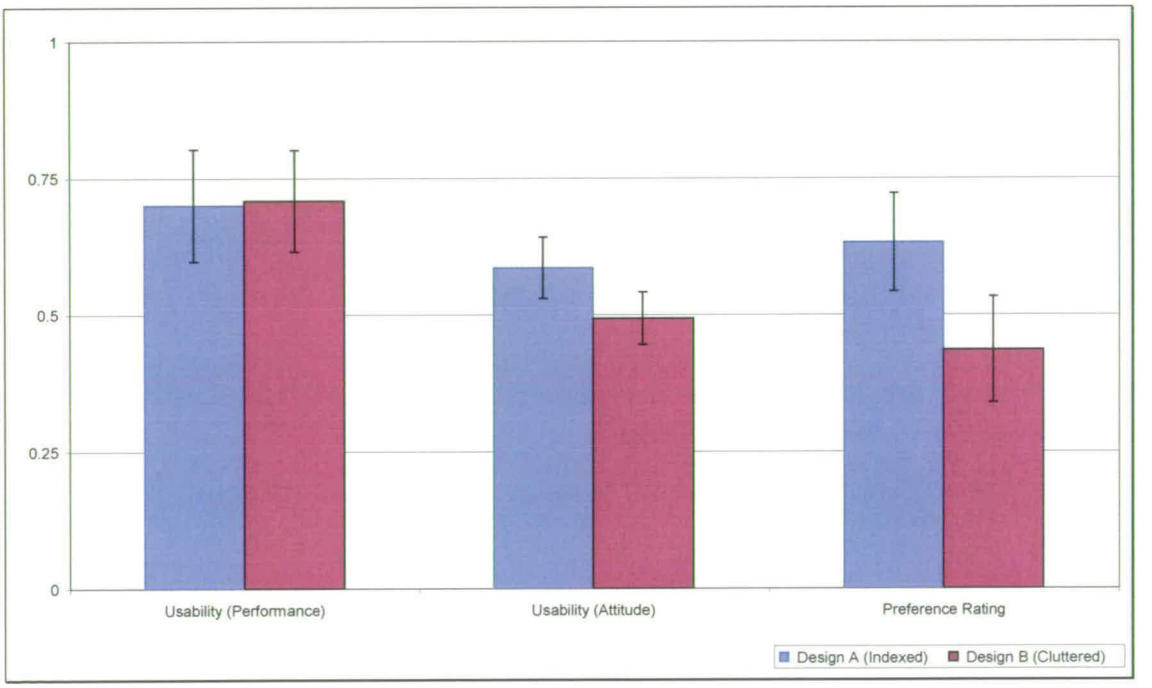


Figure 4.8. Comparison of Normalised Usability Metrics for Two Web Portal Designs

4.5. Pilot Study 1: Discussion

The small pilot study tested the format, implementation and presentation of a set of metrics designed to evaluate usability in the context of Banking Websites. The metrics were designed to generically apply to Websites providing information. In evaluating two alternative Banking Portals, the metrics provided information from two crucial aspects of usability:

1. Attitudes to a range of statements relating to a modern, broad definition of usability and interface characteristics, indicate where improvements need to be focused in the development process. In addition, overall preferences (or quality) scores which are used to confirm the findings of the usability (attitude) questionnaire scores.
2. Task completion as the main performance measure, to ensure that realistic tasks can be carried out by a range of users using the different interfaces.

In this study, performance measures did not differentiate between the two interfaces. In fact, both designs included features which contributed to problems in completing certain tasks. The error bars were wide, showing the different abilities of participants to find the information required.

The usability attitude questionnaire showed a clearer difference between the two designs than the performance metric, but not as clear as the final comparative preferences. Most aspects of Design A scored positively whilst aspects of Design B showed heavily depressed scores, particularly for attributes relating to the structure, layout and visibility of links and information density. The questionnaire had good reliability (>.9), see p.82. The position and contextual appearance of certain important links, lack of differentiation between hyperlinks and text in the content pane and reduction of visual clutter were highlighted as potential focus points for any redesign of interface B. Categorisation mismatches were highlighted for Design A.

Preference scores mirrored and exaggerated the overall subjective usability results. However, as can be seen from the error bars, there was much more variation in preference compared to the mean of the usability questionnaire scores. Nevertheless preference ratings provided a helpful counterpoint to summarise the main subjective usability characteristics which appeared to drive preference and therefore potential success.

A strong positive correlation found between the differences in usability questionnaire and preference scores for the two interfaces (highly significant), with 69% of the variance in one score explained by the variance in the other.

Attitudes towards interface user friendliness, like using, improvement needed or would use again were very highly correlated with preference. Page content and organisation (IA) were also highly related to preferences.

Usage intent as measured in the interview was low. This was subsequent to experiencing mediocre (~70%) performance rates, with corresponding low-medium attitudes toward a broad range of usability attributes (at well-matched utility) on two different interfaces. In contrast, the measure of usage intent in the attitude questionnaire (*use again*) was strongly correlated with preference in all cases. However, *use again* only correlated strongly with performance for the relative differences between the two designs (p.86). Although rated differently for each Portal design at 4.83 for Design A and only 4.13 for Design B (see Figure 4.5, p.76), the statement was not a differentiator between the two designs, as scores were very widely ranged.

A relationship was apparent between reuse intentions and preference. The broad usability questionnaire also strongly relates (highly significantly) with preference (p.86). It would be of interest to determine whether this relationship holds for different interfaces designs.

4.5.1. Limitations

The sample size (N=24) was relatively small for a formative usability evaluation considering participants of different ages and gender. The final cohort was not precisely balanced for age group which resulted in slightly different sample sizes in the between subject groups. In addition, order of experience could not be included as a main factor in the analysis due to the lack of balance and participant numbers. Although counterbalanced and eliminated from being a main effect in the analysis, there could be order interactions that may not have been accounted for in these results.

In the analysis, a minimum of four participants and a maximum of eight made up the groups for the ANOVA procedure. At these sizes, there were wide ranges of performance and perception differences between participants. This may affect the statistical reliability of the conclusions. The effects noted even at this small sample size were very highly significant in most cases, minimising these concerns. However, more precision in balancing participants

into groups to include age, gender and order of experience as main factors in the analysis, and a higher over-sampling ratio is desired to ensure that groups were equal and adequately sized. This would increase confidence in the conclusions of statistical analyses.

A final limitation is the lack of a controlled variable to compare between these designs, as several aspects were different. Therefore the results determined cannot be attributed to any particular feature in the designs. A more controlled comparison of interface designs would allow this.

4.5.2. Outcome and Further Work

To further assess the reliability and consistency of the proposed metrics more participants and other interfaces should also be considered. The inclusion of all the three highlighted between-participant groups (age, gender and order) in confirming these findings required a larger sample. It would also be of benefit to introduce further interface designs; in particular a controlled variable comparison could offer insight for creating usable Websites.

There appears to be a strong relationship between attitudes toward reuse of an interface design and preference ratings. This relationship held for the differences between each interface as well as each design separately. The evaluation of further interface designs exploring these metrics may confirm that this relationship generally applies.

Design B clearly required further improvement. One fairly superficial change which had the potential to increase usability was the reduction of clutter. In theory the remaining content should be easier to scan and navigate.

4.6. Pilot Study 2: Reduction of Web Site Clutter

The second pilot study involved comparing the usability of a low clutter version of Design B, following the look, feel and IA but with reduced graphics and elements (adverts, graphics and other 'pretty' pictures). The result was more whitespace in the design and better visibility of inline hyperlinks (within the content pages) and menus, see Figures 4.9 and 4.10. Otherwise, both designs provided access to the same range and type of information, the only difference being the presentation form. This comparison was more controlled due to the matching of the bulk of the design in terms of content, menus, structure and overall style. The experiment otherwise proceeded exactly as the first pilot study.

4.6.1. Hypotheses

Data were collected to test the null hypotheses, with the level of significance selected as .05 in a two-tailed test. Significant results can be attributed to the variation in clutter on the designs as all other factors were balanced or controlled by randomisation.

Pilot Study 2: Reduction of Graphical Clutter in Web Usability.

The null hypotheses tested were:

Hypothesis H_0 P2a: The reduction of graphical clutter will not result in different usability attitude and performance scores.

Hypothesis H_0 P2b: The reduction of graphical clutter will not result in different user perceptions of preference.

Hypothesis H_0 P2c: There will be no relationship between the usability measures of performance, attitude or preference.

4.6.2. Experiment Design and Materials

Participant variables such as age and gender were balanced in the experimental design along with the order of experiencing the two interfaces and the task sheet presented with each.

The cohort recruited for this experiment excluded anyone who had participated in the first pilot study, to ensure they had no experience with Design B in advance of their session.

By retaining the procedure and materials from the first pilot, comparisons can be made between the two studies. Therefore materials were reused in this larger pilot study.

In one slight change in procedure for this study, participants were asked before and after the experiment about their usage of financial information on the Internet. This was an attempt to determine whether using test interfaces had any influence on intentions. This relates to the diffusion of innovations theory, that amongst other factors, trialability (the ability to test out a new innovation) will influence adoption. Therefore expected rates of usage before the experiment begins would be expected to be lower than intentions after trials of the two designs.

Figures 4.9 and 4.10 illustrate Design C (Reduced Clutter).

See previous Figures 4.3 and 4.4 (p.73) for illustrations of Design B (Cluttered).

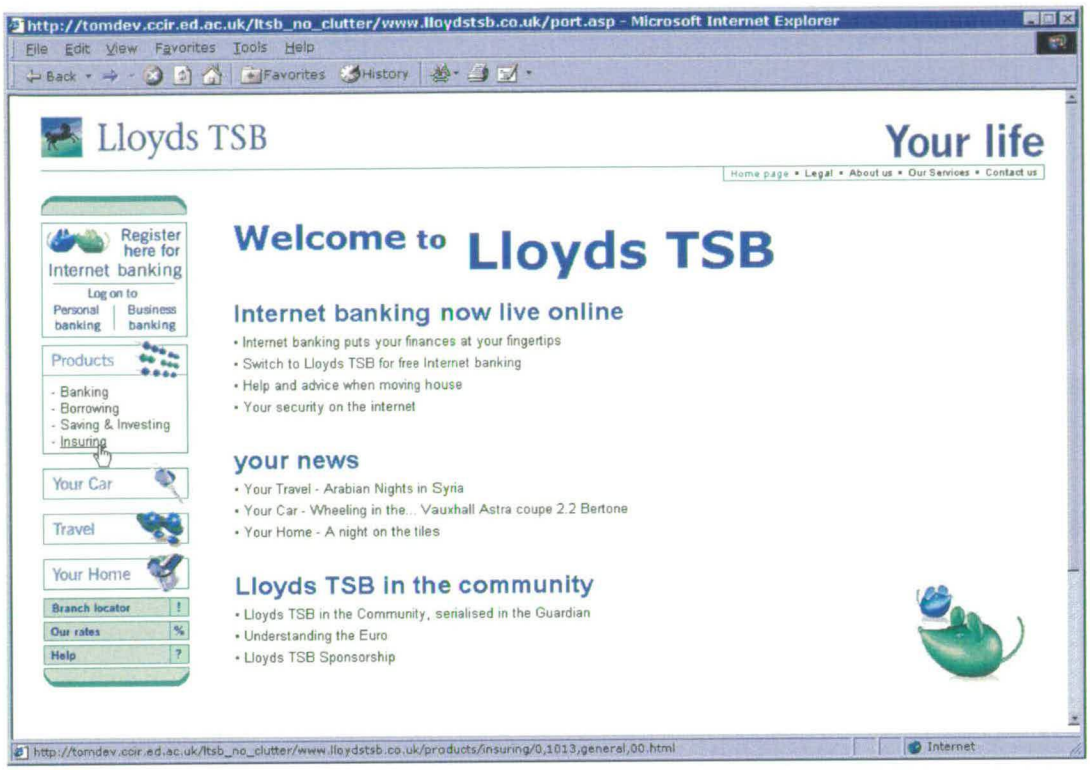


Figure 4.9. Homepage of Design C (Reduced Clutter)

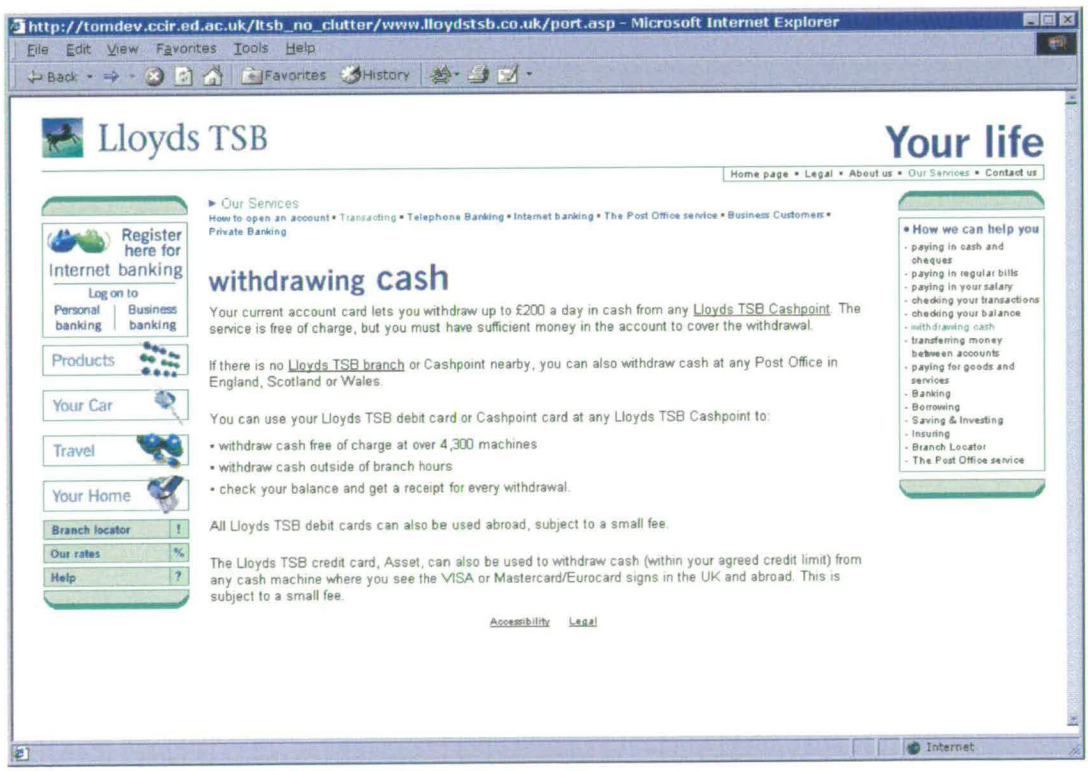


Figure 4.10. Contents Page of Design C (Reduced Clutter)

4.6.3. Experiment Design Summary

Pilot Study 2 - Reduction of Graphical Clutter in Web Usability

Design: Two cell, repeated measures, within subjects

Independent Variables: Design B (Cluttered)

Design C (Reduced Clutter)

Participant Independent Variables: Age (2 levels)

Gender (2 levels)

Order of experience (2 levels)

Dependent Variables: Attitude questionnaires

Task completion

Quality ratings & preferences

Confounding variables: Researcher Bias (randomised)

Room (randomised)

Task sheet (balanced/matched task per task)

Other data: Think aloud remarks

Researcher observations

Interview questions

Sample size: 2 orders x 2 genders x 2 age groups over-sampling 6 = 48

Honorarium: £20

Session Time: 1 hour

4.7. Pilot Study 2: Results

4.7.1. Participants

The sample was closely balanced in terms of gender, with twenty (45%) female and twenty-four males (55%). The age split this time was 50:50 from the age groups under 30 and 30+. Most volunteers were aged 20-29 years (20 participants, 45% of the cohort), there was also substantial numbers in the 30-39 years old group (11 participants, 25% of the cohort). The oldest participant was in the 60-69 age group. The order of experience for the two interfaces was also balanced. The resulting cohort was divided into fairly equal groupings for these key between-subject factors.

Participating users made frequent use of the Internet (a recruitment criterion) with 24 (54.5%) using it daily, and others using it 2-3 times a week (resulting in a cumulative percent of 82% of the cohort (36 people) using the Internet 2-3 times a week or more frequently).

Some 11 (25%) participants used eBanking services to manage their accounts online. Additionally, 5 (11.4%) of the sample were customers of the Case Bank (slightly less than the previous recruitment). A further 29 participants (66%) had used the Internet for shopping purposes.

4.7.2. Performance

In this experiment, task completion rates were fairly moderate, as shown in Table 4.9. These scores indicated there was scope for improvement in both Portal designs.

Portal Design	Performance	St. Dev.	N	Lower Bound CI	Upper Bound CI
Design C (Reduced Clutter)	3.73 (74.6%)	1.065	44	3.40 (68.0%)	4.05 (81.0%)
Design B (Cluttered)	3.52 (70.4%)	.976	44	3.23 (64.6%)	3.82 (76.4%)

Table 4.9. Task Completion Sum (Rate) for the Alternative Web Portals

Task completion for both task sheets was very similar for the two Web Portal designs. Overall, the two task sheets were equally distributed amongst the two designs and the participants, balancing any effect of task difficulty within the different sheets. A paired

sample t-test on the two mean completion scores for each interface found no significant difference between them, $t(43) = 1.354$, $p = .183$, although in this case there is a clearer numerical difference favouring Design C (c.f. p.74). A repeated-measures ANOVA run on the performance rates showed no significant differences between men and women, or for participants of different age groups or orders of experience.

4.7.3. Attitude

There were enough participants in this sample, balanced in terms of the three important between subject factors of age, gender and order of experience to include all these in the analysis, Table 4.10.

Portal Design	Mean Score	St. Dev.	N	Lower Bound CI	Upper Bound CI
Design C (Reduced Clutter)	4.145	1.065	44	3.821	4.469
Design B (Cluttered)	3.843	.937	44	3.558	4.128

Table 4.10. Usability Questionnaire Means for the Alternative Web Portals

There was a main effect between the two designs in terms of usability, Design C outperforming the Design B in terms of attitude toward usability, $F(1, 36) = 6.139$, $p = .018$. No within subject gender or age group effects were found. However, there was an order effect, $F(1, 36) = 7.075$, $p = .012$, strongly influencing the usability scores of each interface: where Design C was used first, both interfaces scored similarly (C: 3.92; B: 3.95), however where Design C was seen second, a much higher rating was offered compared to the Design B (B: 3.80; C: 4.52). Thus there appeared to be a learning effect, probably due to the matched IA, structure and content of the sites, but one which saw those struggling with Design B finding the usability much improved when using the less cluttered version (Design C). However no learning effect was apparent in the opposite direction.

A between-subjects age effect, $F(1, 36) = 4.648$, $p = .038$ showed younger participants giving significantly higher scores to both designs than their older counterparts: 4.30 for the younger group compared to 3.79 for the older. Also an age/gender between-subjects effect, $F(1, 36) = 5.838$, $p = .021$ showed older males giving very low scores to both designs compared to the other groups (3.3 c.f. 4.2-4.4).

Similarly to the previous pilot, Design B obtained an overall mean score of less than 4 on the 7-point scale, indicating general usability problems over a range of considerations.

Different attitudes in the usability questionnaire were found despite very similar task performance on both designs. Similarly to performance scores, the mean attitude scores also indicate substantial room for improvement in both designs. Individual attribute analysis was performed to identify specific areas to focus redesign efforts, and to further explain the perceived differences between the two Portal designs.

The individual attribute results are presented in Table 4.11. The individual scores are illustrated in the Chart, Figure 4.11. Stars indicate significant differences between scores (* $p < .05$ and ** $p < .01$).

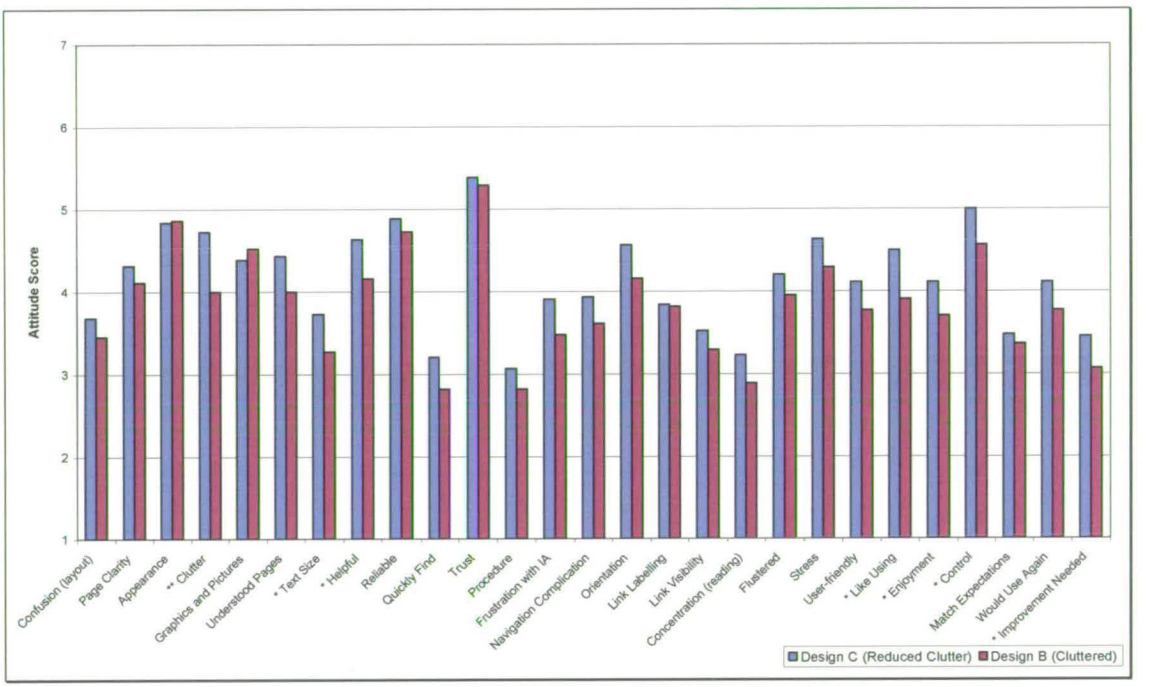


Figure 4.11. Comparison of Usability Attributes for Alternative Clutter Levels

From the chart it is clear that Design C obtained some substantially higher scores compared to Design B, but that in general the pattern of the problems remains in the IA, layout and link labelling and visibility aspects of the design, causing problems in getting task procedures completed and finding the appropriate information.

Attribute	Effect (p)	Direction	W-S Inter[a] (p)	B-S Effects[b] (p)
Clutter	.004	C > B	-	Gender .029 (F > M)
Liked using	.013	C > B	Order .039 (CB: C ≈ B; BC: C >> B)	Age .021 (Y > O)
Helpful	.015	C > B	-	-
Text size	.019	C > B	-	-
Enjoyment	.020	C > B	-	Age .008 (Y > O)
In control	.039	C > B	-	-
Improvement needed	.046	C > B	Order .006 (CB: C < B; BC: C >> B)	-
Concentration (reading)	.055 NS	C > B	Order <.001 (CB: C < B; BC: C > B) Gender .021 (F: C > B; M: C = B)	-
Quickly Find	.066 NS	C > B	Gender .013 (F: C > B; M: C = B)	Order .029 (CB < BC); Age .011 (Y > O); Age/Gender .002 (M2 < others)
Frustrating organisation	.074 NS	C > B	Order .034 (CB: C ≈ B; BC: C > B)	Age .016 (Y > O); Age/Gender .012 (M2 < F1, F2 < M1)
Navigation complication	.075 NS	C > B	-	Age .038 (Y > O); Age/Gender .001 (M2 < F1 < F2 < M1)
Use again	.079 NS	C > B	Order .003 (CB: C < B; BC: C >> B)	Age/Gender .007 (M2 < F1 < F2 < M1)
Procedure	.080 NS	C > B	Order .033 (CB: C ≈ B; BC: C > B)	Age/Gender .013 (F1, M2 < F2, M1)
User-friendly	-	-	Order .001 (CB: C ≈ B; BC: C > B)	Age/Gender .054 NS (M2 < others)
Flustered	-	-	Order .020 (CB: C ≈ B; BC: C > B)	-
Confusion (layout)	-	-	Order <.001 (CB: C < B; BC: C > B)	Age .035 (Y > O)
Link clarity	-	-	-	Age/Gender .010 (M2 < F1 < F2 < M1)
Orientation	-	-	-	Age/Gender .032 (M2 < F1 < F2 < M1)
Understood pages	-	-	-	Age .008 (Y > O); Gender .031 (F > M); Age/Gender .026 (M2 < others)

Table 4.11. Interactions & Between-Subject Effects on the Individual Attributes

In fact only seven attributes are significantly different favouring Design C, $p < .05$. These attributes relate to the control, enjoyment, attitude towards using, helpfulness, text size, clutter and overall amount of improvement needed. Of these, only the clutter attribute is highly significant. There were also some marginal results and a range of between-subject effects for the different participant groups in this comparison.

It is clear that the reduction of graphics did not improve the perception of other characteristics in the design, such as the IA, structure, content and position, clarity and visibility of links within and in menus. Thus it can be concluded from these results that although reducing clutter does increase the perception of text size, enjoyment, control and helpfulness of the content, it does not significantly improve the findability of items in a poorly structured design. Similarly, the degree of clutter did not impact task performance. The recommendation for redesign is the focus on IA, but it may be appropriate to keep a minimally cluttered look overall as attitudes toward clutter was certainly higher as it reduced.

Although balanced in the design, the order of experience did show an overall learning effect when clutter was reduced on the second experience. With only moderate task completion rates, the chance of a fatigue effect could have been high; however with these two designs exhibiting the same IA and much of the structure, links and content being identical, a learning effect was the more likely. Overall the order effect on the usability questionnaire mean exaggerates the superiority of Design C (Reduced Clutter) to the Design B (Cluttered). In the individual attributes, this effect is mirrored frequently: *use again*, *improvement needed* and *liked using*, see Table 4.12 for an illustration of scores for *use again*. Five of the twenty-seven attributes show highly significant interactions between interface and order, another four show significant interactions.

Order	Portal Design	Mean	St. Dev.	Upper Bound CI	Lower Bound CI	N
CB	Design C (Reduced Clutter)	3.650	.265	3.112	4.188	23
	Design B (Cluttered)	3.942	.295	3.344	4.539	23
BC	Design C (Reduced Clutter)	4.811	.282	4.238	5.383	21
	Design B (Cluttered)	3.743	.314	3.107	4.379	21

Table 4.12. Interactions between Interface and Order for *Use Again*

Individual differences between participants also had strong effects in this experiment; there were various between-subject age effects where younger participants were awarding higher scores to interfaces (both designs) than their older counterparts. Younger participants thought interfaces were more enjoyable, liked using them more, were more positive towards being able to quickly locate items, understood page content better, and were less critical of the IA, page layout and navigation.

There were a few variables where between-subject gender effects were noted, women were more concerned about the degree of clutter on the interface and they gave higher scores towards understanding page content.

There were several age/gender effects on the scores to both interfaces, a few variables showed older males giving much lower scores than other groups – this was mostly toward speed of finding information, understanding page content and overall user-friendliness. In other attributes, older males scored lowest, followed by young women, old women and then young men scoring the most positively, this was seen for navigation, orientation, IA, link clarity and whether the interfaces would be used again.

Again there have been a range of age, gender and order of experience aspects which have explained differences in attitudes towards the alternative interfaces, suggesting that in the banking domain, these user characteristics are important to balance and include in controlled analysis.

4.7.4. Preference Ratings

Order of experience, age and gender effects were also considered in reference to the preference (or overall rating of quality on the scale best-worst), the resulting analysis is presented in Table 4.13.

Portal Design	Mean Score	St. Dev.	N	Lower Bound CI	Upper Bound CI
Design C (Reduced Clutter)	14.56	6.917	44	12.45	16.66
Design B (Cluttered)	11.91	6.368	44	9.97	13.85

Table 4.13. Mean Preference Ratings for the Alternative Web Portals

There was a main effect between the two Portals in terms of preference, Design C outperforming Design B in the overall rating, $F(1, 36) = 8.551$, $p = .006$. The within subject order effect was significant at $F(1, 36) = 16.544$, $p < .001$, indicating that those who used the interfaces in the order B then C saw elevated scores for the Reduced Clutter version (C), similar to the order interactions with attitude questions. There was also a marginally significant between subject age effect, $F(1, 36) = 3.944$, $p = .055$, showing younger participants had a tendency to give higher scores than their older counterparts.

In fact, neither design obtained an overall mean score greater than 50% of the 0-30 preference scale, indicating generally low perceptions of overall interface quality. Although one design (the reduced clutter version) was preferred, there was much room for improvement in both. The preference ratings followed closely the pattern seen in the usability attitude metrics for these two designs.

4.7.5. Preference Rankings

The overall rankings of the two interfaces, as derived from the preference rating scores are shown in Table 4.14.

By removing the undecided participant, a binomial test on the dichotomous choice between the two portal designs was computed assuming both portals had equal chance of being selected as preferred. The result shows a significant bias towards preferring Design C, $p = .032$, $N = 43$.

Portal Design	N	%
Design C (Reduced Clutter)	29	65.9
Design B (Cluttered)	14	31.8
No Preference	1	2.3
TOTAL	44	100.0

Table 4.14. Preference Rankings

4.7.6. Intention to Use a Financial Portal

Before the experiment, 12 participants (27% of the cohort) responded that they had used the Internet for banking purposes, this included 11 participants who also used eBanking services

for account transactions. Similarly to the first pilot study, the final question asked participants whether they would be encouraged to use financial Portals on the Internet in the future. For this experiment, 20 participants (46%) responded positively in terms of intention to use a financial portal. This matches the expectation suggested in the diffusion of innovations theory, that the ability to test interfaces will influence intent and adoption. However, 23 participants (52% of the cohort) were not encouraged to use such Web sites, with one participant undecided.

4.7.7. Qualitative Data

Comments and observations were very similar to those mentioned in the first pilot study in reference Design B (as this was the basis for both designs testing the degree of clutter on usability metrics). The main problems were seen in terms of the information structure and the variety, location and some text size in the menus. The opinions were that these aspects made options difficult to find, read and notice on the screen. Some of these issues (e.g. menu and hyperlink salience) were more of a concern with Design B than Design C, with less elements competing for attention. People using Design C also tended to click the right hand side menu more frequently as it was clearer when it appeared (compared to being amongst a persistent menu consisting of links to other group brands as was the case in the original redesign).

The appearance of the sites, white backgrounds and simple clean colour schemes were appreciated for both designs

4.8. Analysis of the Web-Usability Questionnaire

4.8.1. Questionnaire Reliability

The overall alpha (α) calculated for usability questionnaire (27 questionnaire items, 44 participants) over the individual interfaces showed high inter-item reliability: .959 for the Design C, and .948 for the Design B. These values are similar to those found by the smaller cohort in Pilot A.

Examining the values of α for each item if it were deleted again reveals higher α's for Design C (only .001 or .002 higher than the α) for *graphics and pictures*, *appearance*,

reliable and *trust*. Since these are very small changes in α, there is no strong evidence to delete these items.

Similarly, for Design B, there was only one candidate: with α increasing to .950 removing the attribute *trust*. This is not strong evidence to delete this item.

The usability questionnaire has shown high inter-item reliability, with individual items all positively contribute to the overall questionnaire sum and mean. This result is also fairly consistent between two groups of participants evaluating three different interfaces in various comparisons.

4.8.2. Analysis of Neutral Responses

Out of a total of 1188 responses to questions for each interface (27 questions, 44 participants), Design B was given a neutral score 15.2% of the time, 15.5% for Design C.

For Design B, the highest frequency of neutral scores (17) was for the attribute *graphics and pictures*. This was scored neutral by 38.6% of the participants. The second highest frequencies were for *stress* and *enjoyment* which scored neutral 13 times (by 29.5% of participants).

For Design C, the highest frequency of neutral scores (14) was for the attribute *reliable*. This was scored neutral by 31.8% of participants. The second highest frequency (12) was associated with *appearance*.

This analysis points to two aspects which might be perceived less important in terms of satisfaction and usability in the case of Web Portals: the attitude to the amount of *graphics and pictures* used in the design, and the issue of *reliability*. However, these attributes scored neutral around a third of the time, which does not provide strong evidence to remove them.

4.9. Relationships between Metrics

Comparing the different aspects of usability (objective and subjective) can lead to a greater understanding of the concepts and characteristics of interface design which leads to usability and potential success. Again these relationships are explored using correlation analysis.

4.9.1. Usability – Attitudes and Preferences

Taking the differences between the two designs using the direction of overall preference (C-B), there was a strong correlation between perceived usability and quality (or preference), see Figure 4.12. Pearson's $r = .676$, $R^2 = .457$, a highly significant positive correlation, $p < .001$.

For Design B, the results were similar with a highly significant positive correlation $r = .555$, $p < .001$, $R^2 = .308$. Finally, for Design C, the correlation was also highly significant and positive $r = .466$, $p = .001$, $R^2 = .217$.

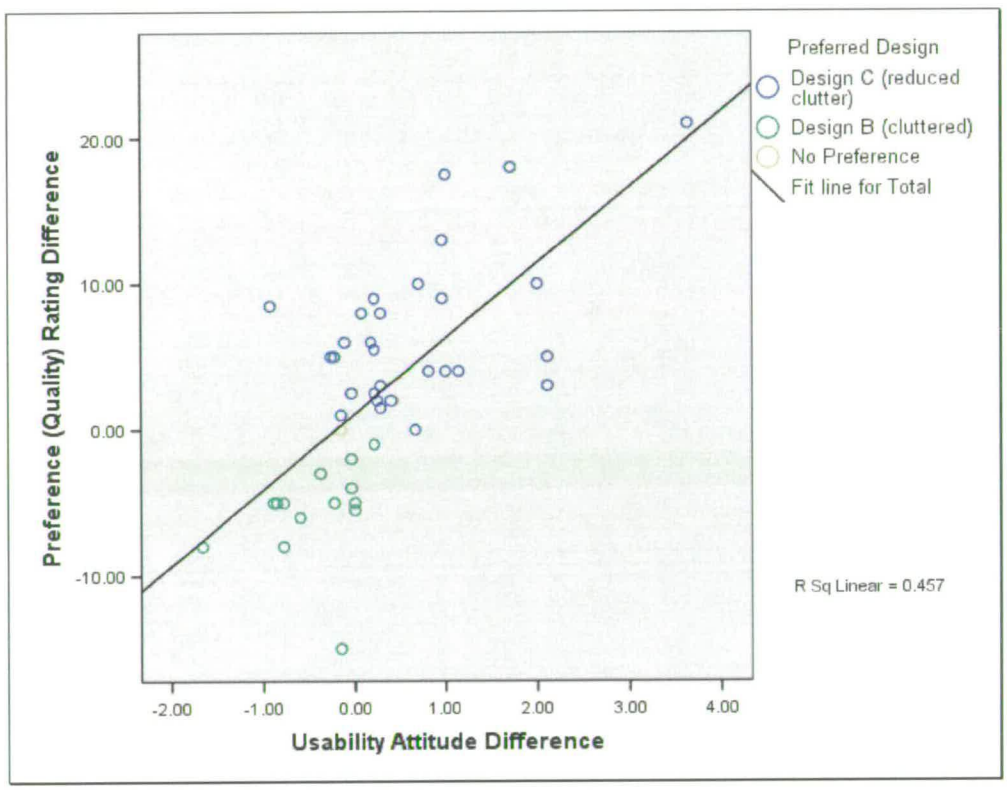


Figure 4.12. Usability Attitude – Preference Relationship (Relative Differences)

With an additional sample nearly twice the size of the first cohort, the correlations are still highly significant, although they dropped from the range .685-.830 to the range .466-.676, this is why examining the significance level of r is important (Kerlinger, 1973).

Nevertheless, the correlations are all highly significant at the $p < .01$ level. The mean of the usability metric is positively related to the preference score each interface obtained.

Individual Attributes in the Usability Questionnaire

The correlation coefficients (*r)* for the usability attributes and preference scores are compared in Figure 4.13. The significance levels are presented in Table 4.15. Highlighted items indicate strong correlations for all three computations (relative difference between interfaces and each individual design). *Use again* and *improvement needed* were strongly correlated for all three with values of r > .5 indicating a strong relationship.

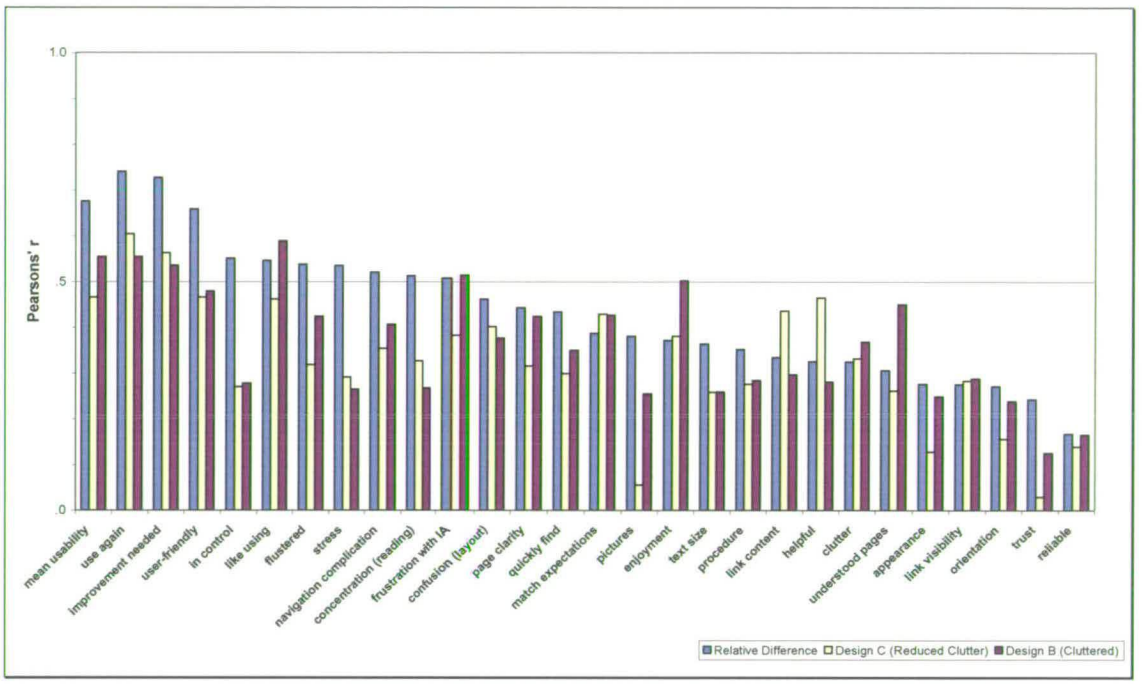


Figure 4.13. Correlations for Usability Attributes and Preference (Quality) Ratings

Strong and significant correlations with preferences (or quality ratings) related specifically to navigation and IA components, some other components related to general ease of use and usability traits and also to future usage intentions. This may be due to the task domain concentrating on information retrieval and comparisons.

Attributes which did not generally relate well to preferences included reliability, trust and appearance. Some of these issues may prove to be less relevant to Web usability than previously thought. Appearance, for example, appears to have little relationship to preference in the Banking Websites studied so far. It may be that in Banking tasks and interfaces, appearance is secondary to functionality, with aspects of look and feel being

supporting characteristics rather than being allowed to dominate preference forming activities.

Attribute	Difference C-B		Design C		Design B	
	r	*p*	*r*	*p*	*r*	*p*
use again	***.741***	***<.001***	***.604***	***<.001***	***.555***	***<.001***
improvement needed	***.728***	***<.001***	***.563***	***<.001***	***.536***	***<.001***
user-friendly	**.659**	**<.001**	**.466**	**.001**	**.479**	**.001**
in control	.551	<.001	.271	.076	.279	.067
like using	**.546**	**<.001**	**.462**	**.002**	**.589**	**<.001**
flustered	**.538**	**<.001**	**.319**	**.035**	**.424**	**.004**
stress	.535	<.001	.292	.055	.265	.082
navigation complication	**.520**	**<.001**	**.354**	**.018**	**.407**	**.006**
concentration (reading)	.513	<.001	.327	.030	.268	.079
frustration with IA	**.508**	**<.001**	**.383**	**.010**	**.514**	**<.001**
confusion (layout)	.462	.002	.402	.007	.377	.012
page clarity	.443	.003	.316	.036	.424	.004
quickly find	.434	.003	.299	.049	.350	.020
match expectations	**.387**	**.009**	**.429**	**.004**	**.427**	**.004**
pictures	.381	.011	.056	.720	.255	.094
enjoyment	.372	.013	.381	.011	.502	.001
text size	.364	.015	.258	.090	.259	.089
procedure	.352	.019	.276	.070	.285	.061
link content	.335	.026	.436	.003	.297	.050
helpful	.326	.031	.465	.001	.281	.065
clutter	.325	.033	.332	.028	.368	.015
understood pages	.306	.044	.261	.087	.450	.002
appearance	.276	.069	-	-	-	-
link visibility	.275	.071	.283	.063	.288	.058
mean usability score	***.676***	***<.001***	***.466***	***.001***	***.555***	***<.001***

Table 4.15. Significant Correlations between Usability Attributes and Preference Scores

Notes:
Blank cells indicate non-significant (or marginally significant) correlations at $p < .05$.

Again, these correlations are generally higher than previous publications would suggest, with the most highly significant ($p < .01$) correlations being in the range $.387 < r < .741$ (Table 4.15).

The mean scores from the usability questionnaire appear to offer a good summary of the characteristics, with 14 highly significant attributes (use again, improvement, user-friendly, control, liked, flustered, stress, navigation complication, concentration reading, frustration with IA, layout confusion, page clarity, quickly find and matched expectations) and 8 significant aspects (graphics, enjoyment, text size, procedure, link content, helpfulness, clutter and understanding of pages) of the interface differences having high relationships with final preference difference scores. Similar numbers and attributes were significant for the individual designs as shown in Table 4.15. However, there is growing evidence that appearance issues are not as crucial as IA factors, affect and control in the formation of preferences in Banking Websites.

4.9.2. Usability Attitude and Performance

Data from the first pilot on the relationship between attitudes and performance was not conclusive. For this data set there were also no significant correlations between task performance differences and mean usability questionnaire differences, $r = .150$ ($p = .331$), a positive direction but a very weak effect. This value is within the range found by Hornbæk & Law (Hornbæk & Law, 2007) and follows their supposition that effective performance has a positive association with satisfaction, even though it is not significant or strong. There were some individual aspects of usability which did significantly correlate with the task performance metrics when looking at the differences between the interfaces, *matching expectations*, $r = .397, p = .008$; *procedure*, $r = .359, p = .017$; and *navigation complication*, $r = .303, p = .045$.

For the individual interfaces, the results were very different. Design C (Reduced Clutter) had highly significant correlations between task performance and mean usability questionnaire score, $r = .503, p < .001$, a large effect explaining some 25% (R^2) of the variation in scores. In addition, nineteen of the twenty-seven attributes in the questionnaire also showed medium to strong correlations with the performance measure:

Matching expectations, $r = .523, p < .001$; *quickly find*, $r = .514, p < .001$; *understood pages*, $r = .488, p = .001$; *flustered*, $r = .472, p = .001$; *like using*, $r = .441, p = .003$; *helpful*, $r =$

.429, $p = .004$; *concentration reading*, $r = .402$, $p = .007$; *navigation complication*, $r = .397$, $p = .008$; *improvement needed*, $r = .397$, $p = .008$; *stress*, $r = .393$, $p = .008$; *user-friendly*, $r = .373$, $p = .013$; *use again*, $r = .356$, $p = .018$; *enjoyment*, $r = .352$, $p = .019$; *in control*, $r = .345$, $p = .022$; *procedure*, $r = .339$, $p = .024$; *link content*, $r = .325$, $p = .032$; *orientation*, $r = .319$, $p = .035$; *trust*, $r = .305$, $p = .044$; and *clutter*, $r = .302$, $p = .046$.

These results show a positive relationship between performance and many of the usability attributes involved with finding, retrieving and understanding content (performance-orientated attributes), along with general ease of use and satisfaction attitudes toward for the preferred interface. However, for the less preferred design, Design B (Cluttered), the relationships were different, with some negative relationships noted for individual attributes, and for the overall mean usability and performance rating the correlation was very close to zero: $r = -.003$ ($p = .984$). This indicated that for this design, although the majority of tasks were completed, attitudes toward the process did not follow performance. This is yet more evidence that attitude or subjective satisfaction may be more important than performance for understanding usability.

4.9.3. Usability – Performance and Preference

In the final set of relationships to explore, the preference (overall quality rating) was compared to performance metrics (task completion). In the first pilot, there were no strong relationships between these scores (p.89). Indeed this was also the case for the comparison of these two interfaces, no significant relationship was present. The differences between the two interfaces showed a weak positive association: $r = .232$ ($p = .130$); for the preferred interface (Design C) $r = .147$ ($p = .340$); and for the least preferred interface (Design B), $r = -.069$ ($p = .657$). This final correlation is very weak negative association, although performance levels were reasonable, preferences did not positively relate to that ability.

4.9.4. Comparisons of Metrics

The various metrics were then normalised (as fractions of their maximum possible scores) to compare in Figure 4.14. The error bars show the 95% confidence intervals (CI) of the means. While Design C had a slight edge in terms of attitude, preferences, and to some extent in terms of performance on tasks, there is no clear difference between the two designs. When

examining the CI data, it is clear that each metric has some overlap reducing certainty in the differences. However, for the sample size of N=44, some differences were found to be statistically significant, with *p*-values of .018 for the attitude questionnaire mean increasing to .006 for the preference rating. The differences were not significant for the objective performance measure.

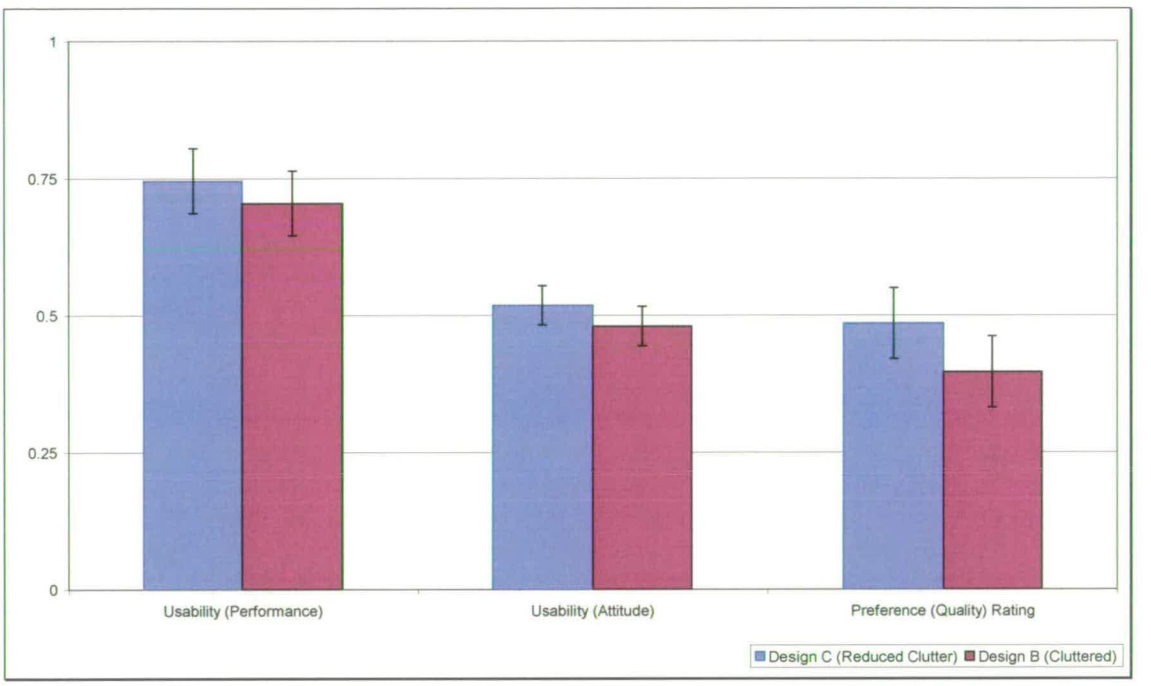


Figure 4.14. Comparison of Normalised Usability Metrics for Reduction of Clutter.

4.10. Pilot Study 2: Discussion

This slightly larger pilot study confirmed the value of the Web usability questionnaire for differentiating between designs and being related to preferences towards different designs. This was in the context of information retrieval tasks on public-access Banking Websites.

In evaluating designs differing only in the density of graphical clutter, they discovered a small benefit in creating lower density pages for Banking Portals. The weaknesses of both designs indicate that focusing efforts on improving the site structure, navigation and hyperlink/navigation salience would be important, and potentially more beneficial.

In this study, performance measures did not show significant differentiation between the two interfaces. The CI's were reasonably wide, showing a range of abilities in terms of locating and processing relevant information. Both Portals used identical IA, and task performance levels were similar and only mediocre (70-75% - see p.74 & p.97). The reduction of clutter tended towards offering improved performance, but not significantly.

Attitudes toward usability as measured by the questionnaire showed a clearer difference between the two designs. However, both designs showed some poor scores particularly for attributes relating to the structure, layout, findability and visibility of links and information density with regard to reading. Areas highlighted for redesign efforts included the position and contextual appearance of certain important links, lack of discrimination between hyperlinks and text in the content pane and categorisation mismatches to the users mental models of the information structure.

Preference scores again mirrored and exaggerated the usability attitude results. This metric provided a helpful point of reference to investigate usability and interface characteristics which appeared to drive preference decisions. A strong positive correlation was found between the differences in attitude toward usability and preference (highly significant) and 46% (see p.106) of the variance in one score was explained by variance in the other.

Future use intentions (*use again)* and *improvements needed* were the highest correlated aspects with preference. Satisfaction with the interfaces in general, *user friendliness*, *like using* and the lack of *fluster* and *stress* were also highly correlated with preferences. Site organisation (IA) and navigation were also very important in relation to shaping preferences.

In terms of usage intention (as collected in the interview), there was moderate interest in the use of the Internet to obtain financial information. This was higher than the initial, self-report of usage (taken at the start of the experiment). In contrast, the term in the questionnaire relating to intentions (*use again*) was strongly correlated with preference, but it was only moderately correlated with performance for the preferred Design (C) with low clutter (p.107). Although *use again* was rated differently for each Portal design (4.11 for Design C vs. 3.77 for Design B, see Figure 4.11, p.99), the statement was not a significant differentiator between the two designs. There was only a marginally significant difference indicating a possible trend in that direction, but an order interaction and a between-subjects age/gender effect also influenced the scores.

The relationship between *use again* and preference found in the first pilot was replicated in this comparison. The relationship was consistent for the study of three different Web portals,

although it did not always extend to being able to differentiate between the interface designs. Selection between the two designs was possible using mean attitudes to a range of usability and interface concerns. Differentiation was also possible with the preference metric alone. Yet only the usability questionnaire and qualitative data collection allowed redesign efforts to focus on problem areas, and for recommendations to be guided by behavioural observations.

The sample size of forty-four participants was a more substantial size for a formative usability evaluation. It also consisted of well balanced groups of participants of different ages, genders and orders of experience, more appropriate for statistical analysis.

In the case of this experiment, the significant order effect and the fact that participants were able to learn the structure of the information from one site to another, reduced the overall effect of interface design characteristics in one set of participants whilst boosting it in the other direction. The direction of these changes only served to reinforce the fact that reduction of visual 'noise' such as uninformative graphics, pictures, subsidiary logos interfered with the search and display of pertinent information.

4.10.1. Limitations

The main limitation to this study was the level of task performance which was only moderate for both designs. Low performance calls interface usability and acceptability into question. Corresponding usability attitudes and preference scores were also low. This may be due to the fact that it is hard to substantiate performance on the Web. Through using controlled tasks and specified goals for task completion the experiment may have been too artificial in measuring performance of Websites. The scenarios used may not have allowed a true representation of browsing behaviour on Banking Portals. In fact, task completion as a performance metric might not be of prime concern in a general information-retrieval setting and certainly in casual Web browsing where engagement with the brand may be the key reason to provide a Web Portal.

Different performance measures may need to be considered on the Web in general, depending on the context of study. Performance as measured by the ability to complete tasks could be an appropriate metric for more critical applications and tasks, such as eBanking. eBanking applications are typically critical in terms of quantitative performance metrics as they are self-service. In order to progress the study of attitudes, performance and

preferences, further studies focus on account-specific transactions in performance-critical tasks.

Similarly, moderate performance levels call into question whether usage intentions recorded in the interview and in the usability questionnaire (*use again*) really extend to real world usage. Although reuse intentions were higher for one design than another, these may only represent relative tendencies. The absolute preference scores, mean usability scores and *use again* scores were all low to medium indicating low levels overall. This was true of all the interfaces studied so far. Higher quality interfaces (over the midpoint of the preference scale) should provide more convincing data that interfaces would be used in real life. Lower performance contributed a little to the low levels of quality as shown in the weak but positive correlations with preferences and attitudes.

Although factor analysis or principle component analysis is often used to identify sets of dimensions within a questionnaire, the numbers of participants involved must be large compared to the number of items in the questionnaire. In these pilot studies, only small numbers of participants were used and the interest was in selecting the most appropriate and potentially successful design as well as finding aspects of the design which contributed highly to preferences. Factor analysis is not appropriate for these samples; however, an inspection of correlation matrices and inter-item reliabilities (Cronbach's alpha, p.37) can determine whether subgroups of related usability components are apparent.

4.10.2. Outcomes

The pilot experiment on reduction of clutter in Website design has succeeded in exploring the suitability of the differing designs and the effect of some between-subject factors. It has also shown the reliability of the Web usability questionnaire, the preference (overall quality) rating scale and the qualitative data collection. Metrics of task performance were less convincing in the Web context of information retrieval but should be helpful in eBanking tasks.

Subjective evaluations of a wide range of usability factors in the questionnaire were more helpful than objective measures. They provided rich evaluation information, redesign focus and recommendations. A preliminary categorisation of different usability attributes into potential subgroups of related components is proposed in Figure 4.15. The categorisation is based on several sources of information:

1. The original categories as defined for the usability questionnaire created for spoken dialogue interfaces.
2. The semantics of the attributes as they were modified for the Web interface.
3. Modifications to the categories accounting for the grouping of high and low value results during the three interface comparisons in the pilot experiments, with particular reference to the higher rated interfaces in each comparison.
4. Adjustments based on reliability analysis using Cronbach's alpha (p.37) within the proposed groups and adjustments based on the correlation matrices between different attributes and preferences.

The figure focuses on results for the two 'winning' designs (Design A with indexed content in Pilot 1 and Design C with reduced clutter in Pilot 2). Where a category of three or more attributes is proposed, the value of Cronbach's alpha is shown. Where only two items are grouped, the inter-item correlation (Pearsons' *r*) is shown instead. The correlation matrices are shown in Appendix D, p.300.

The result was four main categories: Affect (emotional and general satisfaction elements), Structure (elements relating to design of IA and flow), Page Design (layout and salience of page elements), Content (material contained in the Website and text attributes). These subgroups generally showed good reliability (>.8). Residual items which did not group with the above categories were found: *reliability* and *trust* were highly correlated and thus grouped together as 'Integrity'; *use again* and *improvement needed* were also highly correlated and considered here to be 'Quality' elements; *appearance* and *graphical content* did not highly correlate with any other group or category, similarly, they were not highly inter-correlated, thus are separated as miscellaneous visual attributes.

This grouping of attributes for Web usability relates to information retrieval and comparison tasks within a public Banking portal. Many of the usability characteristics measured were highly associated with preferences for one of the two designs used in each pilot experiment; however performance measures were not related to preferences.

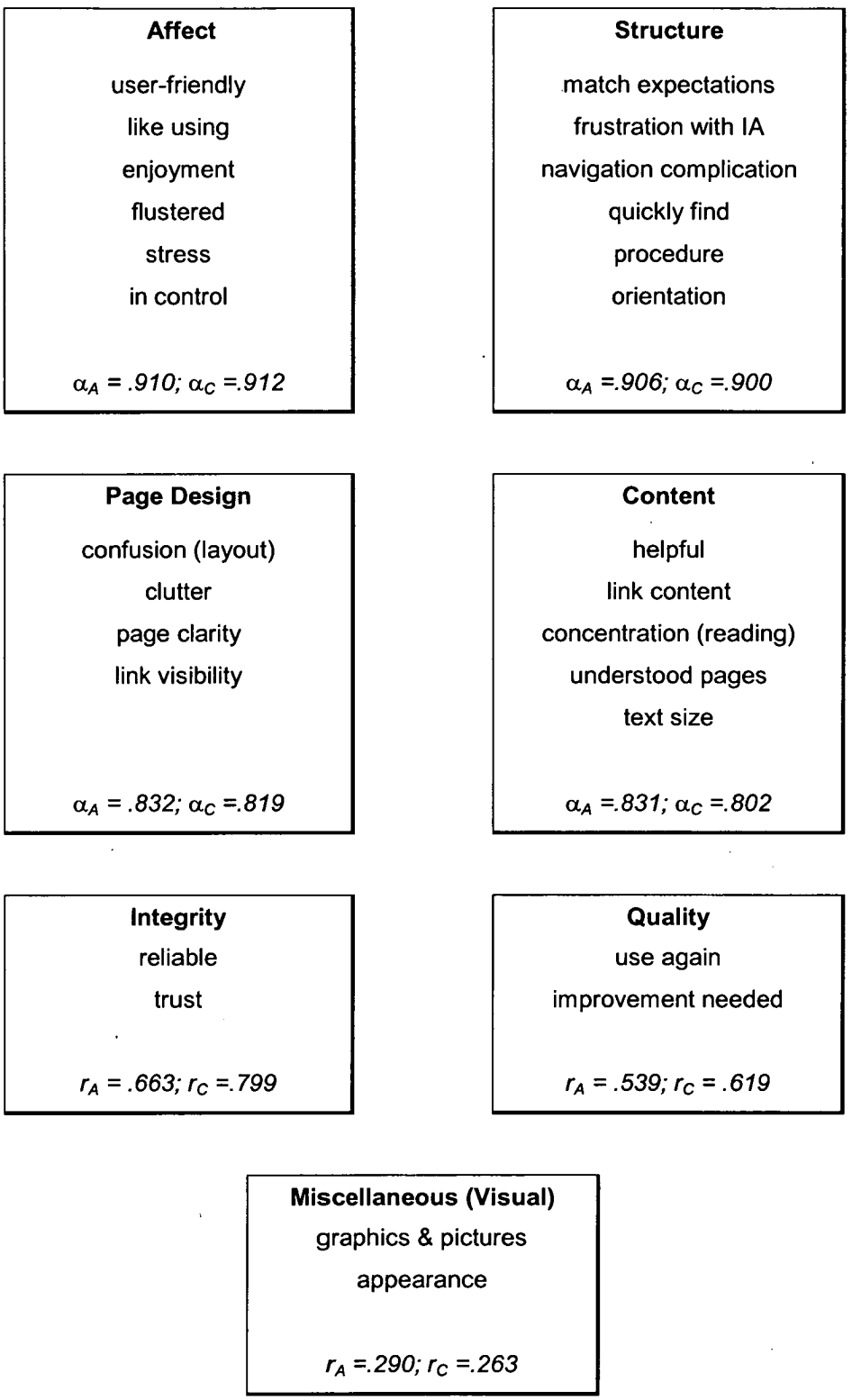


Figure 4.15. The Elements of Web Usability – Subjective Evaluation Components

4.11. Summary of Hypotheses and Evidence

Hypothesis and results for Pilot Study 1 – Usability of Web Portals

Hypothesis H_0 P1a: The different Web Portal designs will not result in different usability attitude and performance scores.

Partially Rejected: The different Web Portal designs resulted in significantly different user attitudes toward usability. However, they did not result in significantly different performance scores.

Hypothesis H_0 P1b: The different Web Portal designs will not result in different user perceptions of preference.

Rejected: The different Web Portal designs resulted in significantly different user perceptions of preferences for designs.

Hypothesis H_0 P1c: There will be no relationship between measures of usability performance, attitude or preference.

Partially Rejected:

There was a *consistently significant, positive and strong* correlation between measures of usability attitudes and preferences.

There was an inconsistent but generally positive association between measures of usability attitudes and performance.

There was consistently no significant correlation between measures of performance and preference.

Hypothesis and results for Pilot Study 2 – Reduction of Graphical Clutter in Web Usability

Hypothesis H_0 P2a: The reduction of graphical clutter will not result in different usability attitude and performance scores.

Partially Rejected:

The reduction of clutter in designs resulted in significantly different user attitudes toward usability. However, it did not result in significantly different performance scores.

Hypothesis H_0 P2b: The reduction of graphical clutter will not result in different user perceptions of preference.

Rejected: The reduction of clutter in designs resulted in significantly different user perceptions of preferences or overall quality in designs.

Hypothesis H_0 P2c: There will be no relationship between measures of usability performance, attitude or preference.

Partially Rejected:

There was a *consistently significant, positive and strong* correlation between measures of usability attitudes and preference.

There was an inconsistent association between measures of usability attitudes and performance.

There was consistently no significant correlation between measures of performance and preference.

4.12. Extending to Evaluate eBanking Services

The subjective attitude questionnaire related to the broad definition of usability described for this research proved highly related to subsequent preferences. Usability and preference were expected to indicate possible future success for a prototype interface, yet performance was not sufficiently high to consider them really successful. The Web usability questionnaire was constructed to evaluate information retrieval on Banking Websites in general. Nevertheless, as usability issues are similar for a wide range of technologies, the questionnaire should have extensions to a) other information retrieval tasks, b) eBanking where customers can perform account-related transactions or c) to other transaction interfaces, including eCommerce activities.

An area where Banks expect to gain the most significant cost savings by implementing Web-based services is eBanking. They rely on customers to use these services for recouping development and maintenance costs. It is of interest to confirm whether the strong relationship between usability in general and intentions to reuse (*use again*) and preference (overall quality) holds in this context. Thus the position of usability practices in a strategic development process for a service such as eBanking can be strengthened.

In the next chapter, the Web usability questionnaire was adapted to suit the evaluation of transaction tasks on eBanking services and determine which usability metrics were most influential in predicting preferences for different interface designs for the service.

Chapter 5. Metaphor & Language for eBanking User-Interfaces

In the eBanking service, customers' goals are to pay bills, transfer money, perform balance enquiries and get statements of transactions (Aladwani, 2001; Tan & Teo, 2000). This chapter presents the Web Usability Questionnaire tailored to evaluating different interface metaphors and dialogue styles in the provision of eBanking interfaces for performing transactions such as bill payments. The experiments evaluate and inform the design of new functionality in the service. Collection of subjective and objective metrics contributes to the understanding of relationships between different measures and dimensions of usability.

Finally, the study examines the contribution of usability and interface characteristics in relation to customer preferences and potential use.

5.1. Introduction

5.1.1. Third Party Payments

Third party payments (P3P) are funds transferred from a bank customer's account to an account at a different bank or to another recipient (Weir et al, 2006). As such they describe any payments made beyond the sphere of a customer's own accounts. These typically fall into two basic categories: bill payments (e.g. utility bills and credit card balances) and miscellaneous payments (paying small companies and private individuals as recipients).

One challenge for the deployment of effective eBanking interfaces is to migrate regular bill payments from other channels (such as telephone banking) onto the bank's Internet service. Thus eBanking services inherit payment arrangements from other channels and initially only displayed information to the user onscreen. The functionality for transfers between a users' different accounts were soon added to online services. By integrating payment functionality eBanking became a fully interactive service.

To design the P3P functionality interface a review of the characteristics of the eBanking service and relevant usability heuristics was undertaken.

5.1.2. The eBanking Service

The eBanking service used as a platform for this research typifies the standard interaction mode used on the Internet, that is linear form filling, and extensive use of banking terminology (jargon). The look and feel of the Website incorporated brand colours, while the structure comprised a top branded pane, a left hand menu, a right hand content pane and an account orientation bar (Figure 5.1).

The user interface metaphor mirrored real-world branch transactions by replicating standard forms with on-screen versions. This metaphor conforms to the principle of consistency (Shneiderman, 1992): The use of a form is familiar to a wide range of people both from transacting in the branch and using various Websites on the Internet. The 'Form' metaphor was therefore examined with respect to usability and human factors issues, with a view to

extending its usage into performing various payment transactions. It was also of interest to examine whether any alternative metaphors might be appropriate for implementing usable P3P functionality.

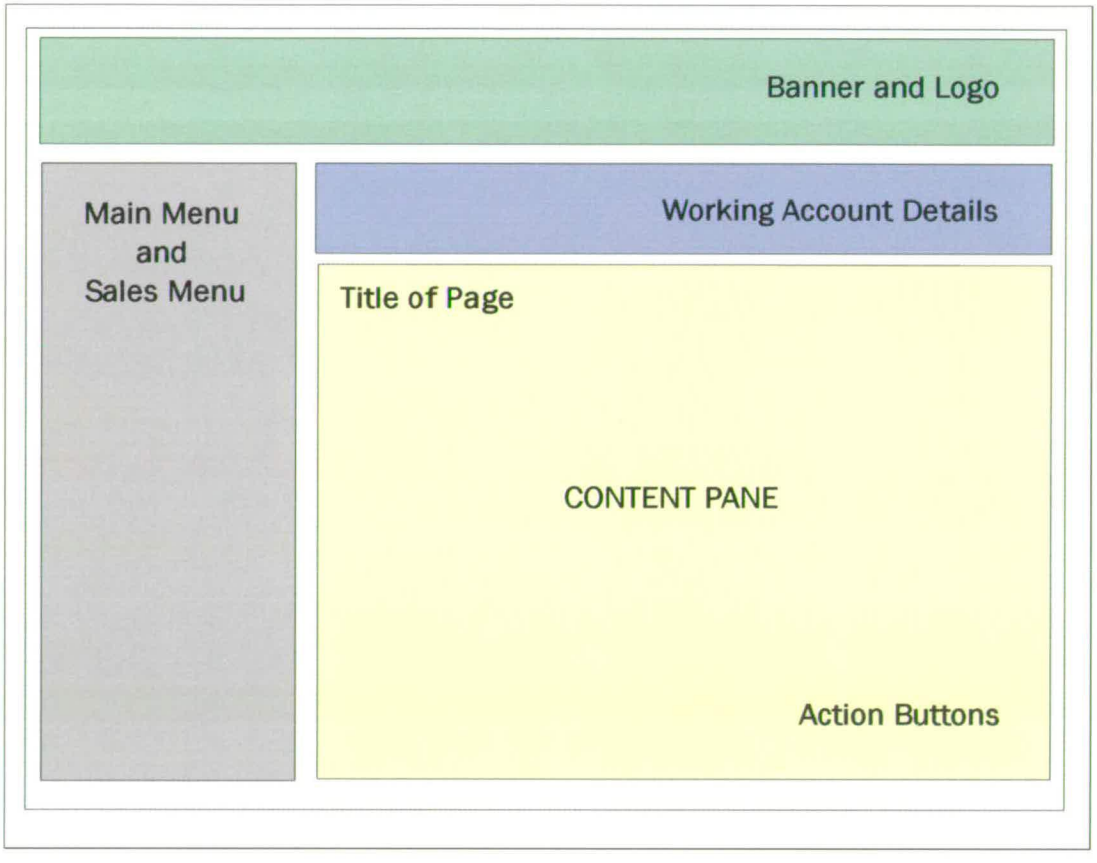


Figure 5.1. Diagram of Basic Website Design (Wireframe)

The use of Bank jargon in onscreen keywords, instructions and all interface dialogues also follows the principle of consistency. On the other hand heuristics suggest that automated services should employ the "user's language" (Nielsen, 1993a). This contradiction is a typical example of the limitation of design principles, guidelines and heuristics, and highlights the need for empirical data. To validate one approach or the other in an eBanking context, these two characteristics were explored to create experimental variables.

5.2. The Experimental Variables

5.2.1. Metaphors

Metaphors are widely used to aid the design and usability of human computer interactions, e.g. the desktop for an office or home computer screen, where a wastebasket icon for deleted files is now in common use by a range of operating systems (Alty et al, 2000) or the typewriter metaphor (blank page) of a text processing package (Carrol et al, 1988). A metaphor offers a model that allows a user's experience, knowledge and behaviour to be transferred from a familiar, real-world entity to, for example, a new computer interface. Metaphors, when well conceived, reduce the burden of grasping new concepts. However, in reproducing aspects of a real world process, the user-interface designer may not produce innovative or usable interfaces, but rather replicate inefficient procedures. An example is the use of manuals, where observations show people flicking through many pages - therefore providing online manuals with page turning and scrolling mechanisms, when really a better way to access the data (e.g. search facility) would be more helpful (Nielsen, 1993a). A formal task analysis can usually identify any problems with current procedures and can be helpful in identifying possible improvements for a new system.

The process of creating a metaphor usually involves, as a first step, understanding what has to be done by the system by analysing the task domain (Erickson, 1995). The resulting metaphor or metaphors must subsequently be tested with representative users to ensure improvement to the interaction design. The interface design must also make allowances for any discrepancies between the metaphor and the desired user actions (Carrol et al, 1988).

In designing an interface for eBanking payments, the form metaphor simply required the digitalization of branch banking forms. In branch, forms are typically completed by the customer and given to staff for actioning on their behalf. This metaphor is easily replicated onscreen and sending the form is analogous to handing it to Bank staff for processing. An alternative metaphor for the payment interface was derived from the common spreadsheet interface, widely used for numerical and financial purposes. These two basic metaphors provided two experimental conditions for usability evaluation in a controlled experiment.

5.2.2. Linear form fill metaphor

The linear form-fill mode utilises a step-by-step procedure that matches user expectations of both the Internet and the Bank. There are advantages to using linear form filling for types of eBanking tasks such as entering details of external accounts. Several details are requested in sequence and entered into form fields by the user. This method is quick to implement, clear and straightforward to use - the design has already been done on the paper forms. However, form filling can be a laborious task for the user. While each form pertains to a single transaction, multiple tasks are not conveniently supported. The eBanking interface also incorporates a common security feature: transaction authentication and confirmation by means of password entry. For multiple payment tasks, multiple transaction authentications are necessary.

This type of individual transaction authentication of is not unusual in eBanking. The password method is an approach used by many UK banks. Other common practices include asking for more than one password (or similar) for transactions. Although there are some eBanking interfaces which do not insist on such additional security measures, eBanking security lapses are frequently exposed and recently even more rigorous security procedures have been investigated (i.e. Nilsson et al, 2005). For these experiments, current security steps were retained and used consistently in each interface option. Evaluation focused on the interaction design for payment functionality.

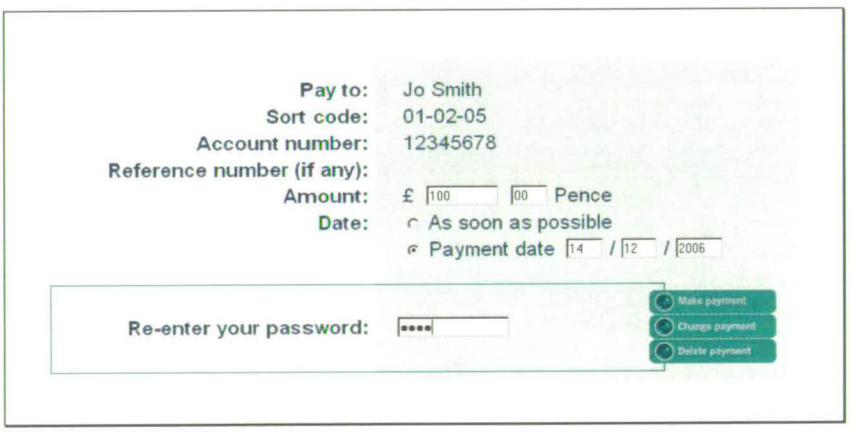


Figure 5.2. Simple Linear form in the Form Design (F)

The Form design (F) was characterised by separate forms for each payment requiring data input, amendment or deletion of details. Payment details could be altered using the forms

(Figure 5.2). Figures focus on the detail of the content pane (as described in Figure 5.1). Each form was accessed from the overview table using hyperlinks, and contained editable details for each payment order. Action buttons offered options to make, change or delete the selected payment.

5.2.3. Spreadsheet metaphor

As an alternative to the simple form, an innovative 'spreadsheet' version was created. Users completed a linear set up process, the payment details being stored in an editable array. Payment amounts, dates and customer reference information could be added, amended and deleted using the spreadsheet. The spreadsheet metaphor emphasised the big picture, control and efficiency by trading simplicity and step-by-step instructions for more complexity and flexibility. The spreadsheet interface allowed users to perform multiple payment tasks on one page, confirmed with a single password entry. Rigorous error checking procedures were employed, with the scanning of each payment row for missing or unrecognised inputs. Errors were fed back to the user and displayed on the screen. In the case of deletion, the dialogue prompted for confirmation whether to delete the pending payment or the entire arrangement (i.e. including the saved account details).

Figure 5.3. Editable Array in the Spreadsheet Design (S)

The 'Payments and transfers' page held a table of accounts to which money could be transferred and payments made. The bill and miscellaneous payments included fields that were directly editable, such as payment date and amount. In principle, multiple changes on

multiple arrangements could be performed at one time, and the password only needed to be entered once, at the end (Figure 5.3).

5.2.4. Dialogue style

The exchange of information between a user and a computer system is known as the human computer dialogue. It was clear that the eBanking interface had been designed using Bank jargon rather than simple vocabulary in the interface dialogue. eBanking interface heuristics offer little advice about whether a service should employ such jargon (to match official bank terminology used on other channels), or whether a more generic, everyday 'plain English' version would produce an interface with superior usability. The idea of removing jargon from interfaces is not new - it is a typical usability concern (Preece et al, 1994). Theory suggests that interface dialogues should contain simple, natural language: the user's language (Nielsen, 1993a). Whilst users of branch networks may be familiar with bank staff using banking terms and performing the appropriate transaction, in self-service the burden falls to the user to choose the correct transactions. Usability heuristics therefore seem to suggest simplifying the interface to match the users' capabilities. From the users' point of view they are simply making payments. However, having consistency in language and terminology between different Bank channels might be more important than using simple language. Whether banks should swap formal terminology for plain language is a debate that is still unresolved (BBC News, 2006; Davies, 1996).

There is a need to examine user understanding of banking terminology and how appropriate it is in self-service contexts. To explore this variable, two alternative human-computer dialogues were created based on the linear form interface design. All (user-initiated) transactions: payments, transfers and standing orders, were combined in a plain language interface. The interface used the linear set-up process, and form-fill metaphor, but described the transaction avoiding any Bank jargon (such as 'standing order' and 'beneficiary'). This design was compared with the use of standard, formal banking terminology in the Form style.

5.2.5. Formal dialogue style

The dialogue style for the formal style of interface was characterised by the use of traditional banking language and terminology. As such, each transaction was referred by its proper name: transfers indicated money movement within a single customers accounts at a bank; payments indicated money sent to external accounts; standing orders were regular arrangements to pay specific accounts on a habitual basis. These constitute user-initiated transactions, typically under the user's control in eBanking. In contrast, direct debits are arrangements that the user has no control over (pull transactions), as after initiation the user can only view the details or cancel the task, with no direct control over timing and amounts.

In addition to use of these terms, the dialogue used by this interface was formal, using terms such as 'beneficiary' to indicate a receiving account. As such, this dialogue is referred to here as the Formal style, an example instruction is shown in Figure 5.4.

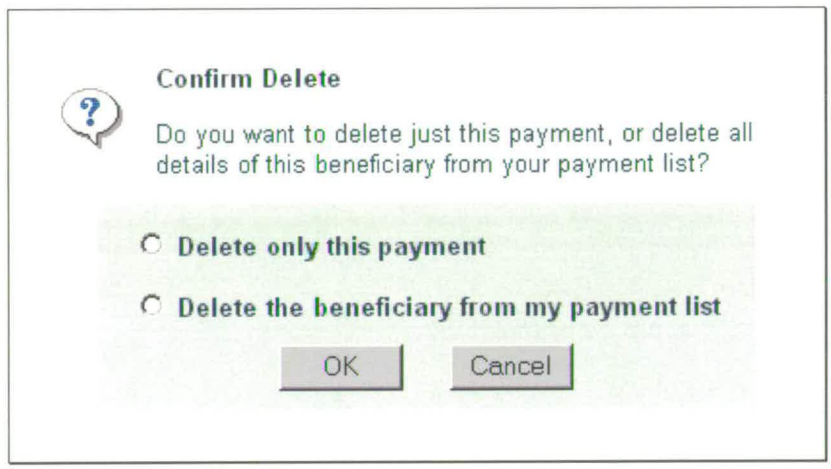


Figure 5.4. Instructions used in the Formal Dialogue Style (F)

5.2.6. Plain Language dialogue style

A panel of user-interface developers and designers with expertise in usability engineering conducted an informal critique of the proposed interface and dialogue designs to remove jargon and technical banking terms, recommending plain language alternatives.

The plain language dialogue style was characterised the use of simple, 'plain English' language in all instructions and labels. For example, instead of making a distinction between payments and standing orders in the menu and assigning alternative pages for these

transactions, the terms 'regular' and 'one-off' were used to describe payments instead. The list of user-initiated payments therefore included transfers, payments and standing orders in one overview table as shown in Figure 5.5. Otherwise, the interaction mode was the same as the Form metaphor and the Formal dialogue style.

Recipient	Amount	Date due	Frequency	Expiry date
CCIR's Savings Account (83-18-34 91889565) CCIR's Holiday Fund (83-18-34 91876490)				
Telewest British Gas J Smith	£45.00 £13.00 £200.00	23/01/2006 07/07/2005 23/06/2005	One-off One-off One-off	
Vodafone Ford Motor Finance (10346) Npower (1234567890)	£22.99 £560.00 £50.00	07/01/2005 23/06/2005 13/02/2006	Monthly Annually Monthly	

Figure 5.5. Overview Table of the Plain Language Dialogue Style (PL)

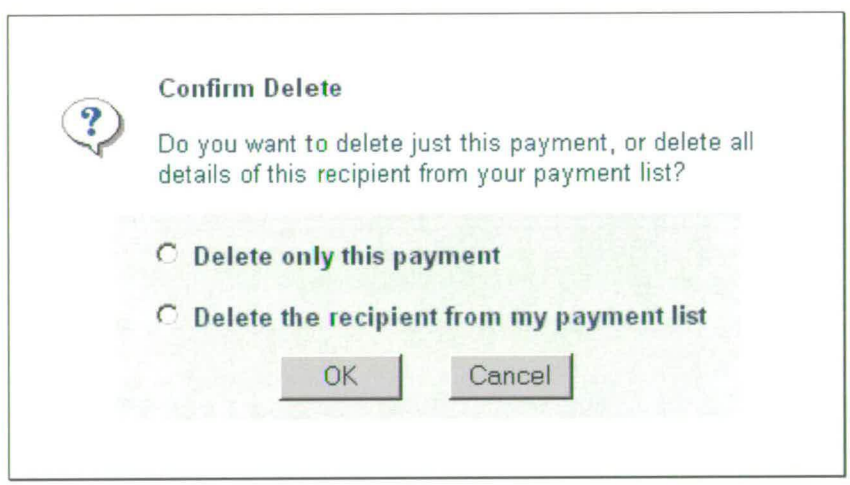


Figure 5.6. Wording changes in the Plain Language Dialogue Style (PL)

In addition, complicated dialogues were rewritten to promote scanning rather than reading and to give additional information regarding any choices available. The process of adding payment details to the interface involved choosing to specify account details or using predefined company accounts to pay (common) utility bills. This process was initially presented as a straight choice in the Formal dialogue style. In the Plain language style, the options were clearly distinguished and the dialogue offered further explanation of the two

alternative forms. Some words were flagged as being particularly inappropriate to the audience, including the term 'beneficiary'. The more commonly understood 'recipient' was used instead, as shown in Figure 5.6. A full list of substitutions is available in Appendix F, p.310.

5.3. Research Questions and Hypotheses

The two main considerations investigated in the experiments are firstly the interaction metaphor and secondly the dialogue style used in the human-computer interaction dialogue. Both of these design artefacts were hypothesised to have a direct impact on usability. Each interface tested was substantially different from the next and altered in one design feature only.

Three contrasting user-interface design variants were compared in two experiments:

- The first took the 'Form' metaphor that derived from traditional paper-based banking (F).
- The second took the 'Spreadsheet' metaphor from computerised numerical work (S).
- The third interface matched the look and feel of the 'Form' interface whilst employing simpler, 'Plain language' for on-screen instructions, transaction terms and dialogue. This dialogue style was an alternative to the more formal banking terminology (jargon) used throughout the 'Form' and 'Spreadsheet' interface (PL).

The aim of these studies was to provide quantitative and qualitative feedback on usability and user perspectives on the various interface designs proposed to perform eBanking transactions in order to make recommendations to the design of P3P functionality in eBanking and to establish validated eBanking design heuristics.

5.3.1. Hypotheses

Data were collected to test the null hypotheses. Statistical analysis using the analysis of variance (ANOVA) was performed to examine how the various controlled factors attributed to performance, attitude and quality (preference) measures. The level of significance was selected as 0.05 in a two-tailed test.

Experiment 1: Metaphor for eBanking Transaction Interfaces

The null hypotheses tested were:

Hypothesis H_0 E1a: The different interface metaphors will not result in different usability attitude and performance scores.

Hypothesis H_0 E1b: The different interface metaphors will not result in different user perceptions of preference.

Hypothesis H_0 E1c: There will be no relationship between the usability measures of performance, attitude or preference.

Experiment 2: Dialogue Style for eBanking User Interfaces

The null hypotheses tested were:

Hypothesis H_0 E2a: The different dialogue styles will not result in different usability attitude and performance scores.

Hypothesis H_0 E2b: The different dialogue styles will not result in different user perceptions of preference.

Hypothesis H_0 E2c: There will be no relationship between the usability measures of performance, attitude or preference.

5.3.2. Participants

The participants in the experiment were recruited as being representative of current and potential members of the target market for the Case Bank's eBanking service; they were customers of the Bank, who were 'Internet-savvy' (they were either Internet users, registered with or actually using the eBanking service). The resulting sample of current and potential users of the eBanking service had a wide range of experience with the Internet and eBanking services. By recruiting Internet-savvy participants, it was ensured that the usability

experiment sessions focused on the competing interface designs, rather than the usability of the Internet or computers in general. In addition, the participants represented the target audience for the eBanking functionality being proposed. All undertook a set of defined tasks on the prototype services for the experiment, which were not available on the 'live' banking site. None of the participants had previous experience of using either of the payment interfaces. Participant variables such as age and gender were balanced in the experimental design along with the order of experience.

Participants were grouped by gender and age group, with two age groups selected: 18-39 and 40 years and over. These categories were selected based on previous studies and data about the typical ages of Internet users and eBankers (National Statistics, 2006; Nichols et al, 2001).

5.3.3. Tasks

Current and proposed eBanking tasks were divided into the following categories:

- Passive viewing, such as checking balance information and statement items.
- Transactional interactions, defined as moving money to other internal or external accounts using existing payment orders such as transfers, bill payments and other miscellaneous payments.
- Informational interactions, defined as the process of inputting account details to create payment orders, but also extending to form-filling for opening new accounts (savings, loans etc.) online.
- Spring-cleaning, defined as deleting payment arrangements, payee account details that are no longer in use, inactive direct debits and other dormant account information.

Banks are keen for their customers to be able to fulfil all types of banking needs using the online channel, with its associated cost-savings for the bank and convenience for the user.

In the first experiment, the difference between the two competing designs was the payment interface metaphor. The payment interface offered a list of hyperlinks to individual payment forms or a spreadsheet array of payments. As such, the usability evaluation concentrated on tasks that highlighted the payment interface pages. Participants performed four tasks: one required deleting a future payment to a company; another required deleting an account from the list of recipients and two concerned amending amount details to pay certain accounts and

utility bills. The tasks required simple data input and editing. They were designed to maximise exposure to the two alternative metaphors (see Appendix E, p.304).

In the second experiment, the difference between the designs was in the language style, format and tone used for instructions and terminology. The language changes were distributed throughout the service. Three tasks were selected that highlighted the dialogue changes between the two interface designs. One task required creating a bill payment to pay money to a company (such as a utility or telephone bill). Another task required sending money to a third party account (such as a friend or relative). A third task concerned performing internal account transfers (such as from savings to a current account or vice versa). The tasks required following instructions and input of data into appropriate fields. They were designed to maximise exposure to the alternative dialogues (see Appendix F, p.314).

5.4. Dependent Variables

5.4.1. Usability - Performance

For this experiment, the main performance metric of usability – task completion, was expanded from a simple count of information found or not, to include instances of errors made and error recovery. In a transactional interface, there is more scope for errors to be made and in eBanking transactions it is important that any errors are spotted and corrected properly. Completion was counted when participants completed a task correctly with no errors, or when errors were made and easily recovered. Any errors (and if and how they were corrected) were noted by researchers. Errors were classified, for example, as missing data or a typographic slip. Errors were flagged as catastrophic when incorrect or incomplete banking tasks went unrecognised by the user. Such errors would have caused problems in real self-service banking applications, perhaps resulting in additional calls to human-operated, costly channels to remedy such mistakes.

The classification of errors allowed the following performance measures to be computed:

1. Count of right-first time transactions
2. Count of transactions where errors were recovered and tasks completed
3. Count of catastrophic errors (unnoticed errors resulting in wrong or incomplete transactions)

In each case, the counts were summed and averaged across task sheets to produce rates over the series of tasks performed. Individual tasks can also be studied to guide redesigns.

5.4.2. Usability – Attitude

The change in goals and environment when considering information-retrieval in a public space compared to the management and transaction of personal financial funds required a re-examination of the usability and interface characteristics. With these changes in mind, the Web Usability Questionnaire was inspected and alterations were considered. Adaptations made to the statements considered the new goals and environment as well as the different metaphors and dialogue style variables present in the interface designs within the two experimental groups.

In constructing an eBanking version of the Web Usability Questionnaire, the aim was to include as many relevant statements as possible, with reference to the literature and conclusions from the pilot studies, whilst ensuring the questionnaire was as short as possible. The length of the questionnaire was an important consideration in these two experiments, as the prescribed tasks were fairly limited and the experiences of each interface would therefore be fairly short. It was important in this case not to overburden users with excessive numbers of questionnaire statements. This ensured that the questionnaire would not take longer to complete than the tasks on the interface.

The following items from the Web Usability Questionnaire were highlighted for removal:

To match expectations

eBanking was in an early stage of development when these experiments were carried out. The Case Bank was one of the first to add a provision for third party transactions to their service. As other Banks develop and roll out their eBanking services, matching the expectations of switchers and new customers may become more important, but it was not relevant at the time of the study.

There were also concerns that the question of expectations was likely to be biased toward the Form design in the metaphor experiment. Given the low experience with eBanking services that the cohort was likely to represent, their expectations would generally have been formed in the human-operated channels of Banks. As one interface clearly matched the paper

banking forms, there would be a tendency for bias toward this design to be observed in matching expectations.

The observation that you must match the expectations of users, given that they are likely to be using other designs as well as your own is not really appropriate if the design is made at an early enough point to be involved with forming those expectations. Instead, intuitive and error-free operation of the interface in first-time use was one of the focuses.

In terms of overall performance and preferences, the match to paper forms might be an important characteristic for the transactional interface. However, other data collected (including the qualitative reports and observations) would be sufficient to shed light on this fact, without introducing a potentially biased question into the usability questionnaire. Therefore, the specific statement was removed from the questionnaire for these experiments.

Clutter

The degree of clutter in the interface did not appear to be a strong contender for the eBanking version of the Web Usability Questionnaire. Examining the proposed eBanking interface designs, the overall look and feel was very similar to the original Web Portal (Design A in Pilot 1). Notably, this design was not perceived to be cluttered (5.2 on the 7-point scale, see p.76). The interface was also highly focused. Although clutter was significant in the previous study, it was not strongly related to performance or preference, it may not be a core characteristic on eBanking interfaces such as this. To reduce the length of the questionnaire, this statement was removed for the metaphor and dialogue experiments.

Link Visibility

Similarly to the observations about interface clutter, in the proposed eBanking interfaces, the page content was not information-dense. Furthermore, links within the content pane were formatted using the standard blue hyperlink colour and underline, thus making link visibility less of an issue than it had been for the Web portal designs. It was also the case for the different design alternatives that link position and style did not vary greatly. Although link visibility was a highly relevant attribute for information retrieval tasks and environments, these eBanking payment tasks were much more focused and the amount of information contained within the pages much lower, resulting in this item being removed from the questionnaire.

Link Labelling

The issue of whether link labels clearly indicated their destinations was a potentially important variable in the dialogue experiment. However, it was language generally throughout the dialogue that was of interest. Therefore this item was modified use as an interface-specific attribute in the dialogue experiment. Instead of asking only about links, the new statement asked whether words and phrases used throughout the service were understood:

- "The words and phrases used by the service were easy to understand".

Trust

The item *trust* had been very highly scored for all three Web Portal interfaces (5.21 – 5.63 on the 7-point scale, p.76 & p.99). It was concluded from these scores that the well known High Street brand of the Case Bank had ensured an appropriate degree of trust be perceived even for it's web offerings. Moreover, if even non-customers gave high trust ratings to the public accessible Web sites, Bank customers who went on to use eBanking were also likely to trust the interfaces in question. Further, this experiment did not want to mix up usability and interface design issues with the ongoing debate about how secure eBanking services were. Therefore the trust issue was not considered in this particular set of comparisons and evaluations, except as a qualitative issue when brought up by participants themselves.

Graphics and Pictures

Overall look and feel for both interface designs in each comparison was fairly similar. Indications from pilot studies were that appearance issues were of minor importance in information-retrieval on Banking Websites, not highly correlated with preferences towards designs (p.85 & p.107). However, in other Web research, attitude formation has been dependent on appearance issues (Lindgaard et al, 2006; Kim & Moon, 1998). The use of two appearance-related statements was therefore reduced to one for these experiments. Of the two items, the issue of graphics and pictures is to be less of a concern in the context of providing a transactional eBanking interface – something where serious, professional functionality was considered a key feature. Therefore the more appropriate overall attitude to appearance was retained.

Statement Modifications

The attribute relating to text size was rewritten as it was the general text size that was of interest. In the Web Portal research, the statement had been written to account for the variations in text size across pages and interfaces. In this experiment, text size was the same, although variations in text density (particularly in the dialogue experiment) meant that perceptions may be different amongst the designs. The modified statement read:

- "The text used by this service was too small".

Statement Additions

Specific issues of interest in the dialogue style experiment resulted in the question about words and phrases. In the metaphor experiment, task-specific perceptions about using different interfaces were desired. The term 'easy' was again appropriate in relation to amending transaction details in the tasks with the different metaphors. Thus an additional statement for this experiment was constructed. The statement used the term 'arrangements' which was consistently used in the priming prompts and task sheets:

- "It was easy to change existing arrangements using the service".

Resulting List

The resulting questionnaire for evaluating the usability of transactions in eBanking is shown in Figure 5.7, with attributes grouped according to the provisional categories and structure determined in the pilot tests with some modifications made to allow for the changes in the statements.

The attributes were worded to related to the 'service', the 'Internet banking service' being the term used in the introduction and other priming.

AFFECT
I found this service user friendly
I liked using this service
I did not enjoy using this service
I got flustered when using this service
I felt under stress while using this service
I felt in control when using this service

STRUCTURE
I found the organisation of this service very frustrating
Moving about this service was too complicated
I could quickly find what I wanted on this service
When using this service I didn't always know what to do next
I always knew where I was on this service

PAGE DESIGN
I found the page layout on this service confusing
I found the layout of the pages on this service very clear
The pages on this service were attractive

CONTENT
I felt that the service was helpful
Reading the pages on this service took a lot of concentration
The pages on this service were easy to understand
The text used by the service was too small
It was easy to change existing arrangements using this service (Experiment 1 only)
or
The words and phrases on this service were easy to understand (Experiment 2 only)

QUALITY
I would not use this service again
I feel that this service needs a lot of improvement
I felt this service was reliable

Figure 5.7. Perceived Usability Questionnaire Statements for the Experiments

5.4.3. Usability - Preferences

As a final quantitative measure, participants were asked to rate the two interfaces on the preference rating scale (0-30 points). This occurred after both hands-on sessions and open questions such that both experiences could be compared in terms of overall preference. The

resulting score on the scale was also considered an 'overall quality' rating because it indicated of how close to 'Best' each interface was generally perceived. The scores also offer a relative score between interfaces, and were reduced to a rank order on those terms.

5.4.4. Qualitative Data

In these two experiments, participants were not timed for efficiency; rather they were encouraged to use the 'think aloud' protocol, as previously discussed (p.58) and matching the procedure used in the pilot experiments. The qualitative observations and comments collected during this process were ideal for the early stage of development of self-service transaction interfaces.

Additional qualitative data were also collected in a debriefing interview. Participants were asked whether they noticed any differences between the interfaces. Then they were asked about what they liked, disliked and what suggestions for improvements they could offer.

As a final question, they were asked whether they would like to use the service in real life, determining the likely utility and eventual use of P3P functionality. The interview questions for each experiment are available in Appendices E (p.305) and F (p.315).

5.5. Experiment 1: Metaphor for eBanking Interfaces

5.5.1. Experiment design

The metaphor experiment investigated two alternative interface designs for eBanking services – form fill: the form metaphor and array-edit: the spreadsheet metaphor. A sample of 32 participants (customers of the Bank) took part in the experiment. The sample was balanced for age, gender and presentation order of the interfaces, the between subject factors. The dependent variables were the responses to individual items in the usability questionnaire, the quality rating and deduced preferences. Task completion and error counts were also recorded.

5.5.2. Experiment materials

Participants were asked to perform four tasks (see Appendix E, p.304). Two involved making amendments to existing payment arrangements and two involved deleting information, firstly a future payment and then an entire arrangement. These tasks allowed the users to experience and focus on several instances of the contrasting interface design metaphors. They performed the same number and scope of tasks on the second interface but details were slightly altered to engage the user in their second experience.

5.5.3. Experiment Design Summary

Metaphor for eBanking Transaction Interfaces

Design: Two cell, repeated measures, within subjects.

Independent Variables: Interface F (Form/Formal)

Interface S (Spreadsheet/Formal).

Participant Independent Variables: Age (2 levels)

Gender (2 levels)

Order of experience (2 levels)

Dependent Variables: Attitude questionnaires;

Task completion & error counts

Quality ratings & preferences

Confounding variables: Researcher Bias (randomised)

Room (randomised)

Task sheet (balanced / matched task per task)

Other data: Think aloud remarks and researcher observations

Interview questions

Intention to use

Sample size: 2 orders x 2 genders x 2 age groups x over-sampling 4 = 32

Honorarium: £20

Session Time: 1 hour

5.6. Experiment 1: Results for Interface Metaphors

5.6.1. Participants

The sample was balanced in terms of gender; 16 (50%) were under 40 years old, the other 16 were 40 years and over, with the oldest participants being in their fifties. The order of experience for the two interfaces was also balanced across the participants of differing ages and genders. Participating users made frequent use of the Internet (a recruitment criterion), accessing it mainly from home or work, with a high proportion doing both.

Some 14 (44%) participants used the eBanking service weekly or more frequently. A total of 65% of all transactions performed by eBankers were examples of passive viewing; the remainder were transactional interactions where they performed self-service tasks. Additionally, 14 (44%) of the sample described themselves as predominantly using Internet Banking to conduct their finances.

5.6.2. Performance

Most participants completed the tasks in the experiment without problems. Firstly, the overall task completion scores were computed – including any complete tasks, with or without recovered errors, the data are shown in Table 5.1. Task completion levels were very high for both metaphors, with the form achieving higher rates overall. However, a paired-samples t-test revealed no significant differences in the task completion rates, $t(31) = .895$, $p = .378$.

Metaphor	Mean Rate	St. Dev.[a]	N[b]	Lower Bound CI	Upper Bound CI
Form	3.81 (95.3%)	0.397	32	3.67 (91.7%)	3.96 (98.9%)
Spreadsheet	3.66 (91.4%)	0.827	32	3.36 (84.0%)	3.95 (98.9%)

Table 5.1. Task Completion Sum (Rate) – Alternative Metaphors

Notes:
[a] Standard deviation
[b] Number of participants in the sample

Task completion in this experiment was much higher than the pilot tests, possibly due to the focus of tasks and the interfaces – without much extraneous information in which to get lost.

Tasks completed correctly, first time, with no errors made are shown in Table 5.2. The rates were high, and the rates for the Form design slightly exceeded the Spreadsheet design, however, a paired-samples t-test revealed no significant differences in these scores, $t(31) = 1.139$, $p = .263$.

Metaphor	Mean Rate	St. Dev.	N	Lower Bound CI	Upper Bound CI
Form	3.72 (93.0%)	0.523	32	3.53 (88.3%)	3.91 (97.7%)
Spreadsheet	3.53 (88.3%)	0.915	32	3.20 (80.0%)	3.86 (96.5%)

Table 5.2. Tasks Completed 'Right First Time' – Alternative Metaphors

Examining the logs of errors made, mainly these were interactions which erred on the cautious side: e.g. where participants were asked to delete payment arrangement details for 'old' accounts that they no longer used, they occasionally chose to delete any pending payments, and keep the arrangement details. For the purposes of task completion, such an error was noted, but still counted toward task completion, rather than seen as a catastrophic error. For these tasks and interfaces, no catastrophic errors were recorded.

5.6.3. Attitude

To investigate the impact of the differing interaction metaphors, repeated-measures analysis of variance (ANOVA) procedures were carried out using the mean responses to the twenty-two usability questions completed after exposure to each of the alternative metaphors. Individual question items were also compared in the same way. Three between-subject factors included in the analyses were age group, gender and order of experience.

The mean usability attitude scores (on the 7-point response scale) for the Form and Spreadsheet designs are presented in Table 5.3.

The majority of participants responded to the simpler Form metaphor with higher usability scores. The difference in attitude scores was statistically significant, $F(1, 24) = 5.57$; $p = .027$, in favour of the Form metaphor for the mean of all the eBanking usability attributes measured.

Metaphor	Mean Score	St. Dev.	N	Lower Bound CI	Upper Bound CI
Form	5.398	0.7321	32	5.134	5.662
Spreadsheet	4.906	0.9776	32	4.554	5.259

Table 5.3. Mean usability scores (22 questions) for the contrasting metaphors

Between-subject effects showed a significant interaction between age and gender on the mean of both design variant usability scores, $F(1, 24) = 9.672$; $p = .005$. T-tests on these data indicated that both male and female younger participants score the interfaces consistently ($t = 1.749$; $df = 14$; $p = .102$). Older males score both designs significantly higher than their younger counterparts ($t = 2.270$; $df = 14$; $p = .040$). Older women score both designs significantly lower than the younger women ($t = 2.384$; $df = 14$; $p = .032$). Older women score both designs much lower than older men, a very highly significant result at $t = 3.291$; $df = 14$, $p = .005$, illustrated in Table 5.4. Younger men and older women were less positive towards these designs for eBanking services than other groups. Whether this is a consistent effect for eBanking, Web services or Internet use in general would require further research.

Age Group	Male mean	Female mean
<40	4.793	5.442
40+	5.630	4.914

Table 5.4. Mean scores for both designs by age and gender

Although participants were not balanced in terms of eBanking or Internet usage across experimental conditions in this study, the variable was also used in analysis, however, no effects were found in this sample.

The individual scores for questionnaire attributes for the alternative metaphors are illustrated in Figure 5.8. Six of the twenty-two individual usability attributes were scored significantly (at the .05 level) more favourably for the Form metaphor over the Spreadsheet metaphor, see Table 5.5.

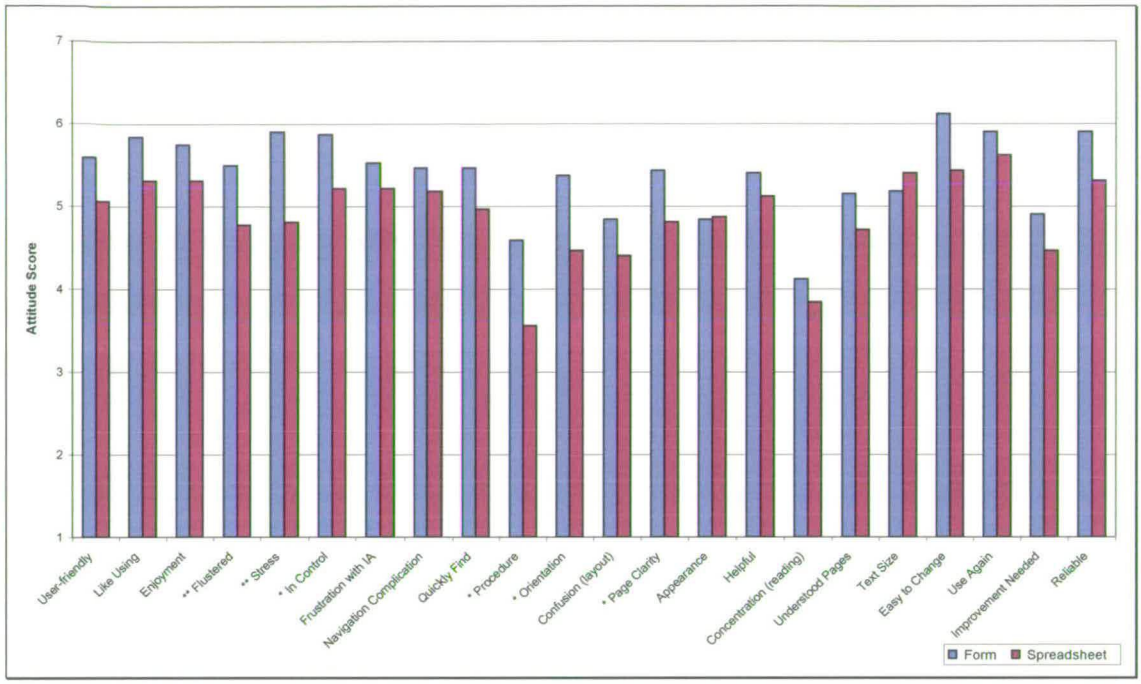


Figure 5.8. Comparison of Usability Attribute Scores for the two Metaphors

Figure 5.9 shows the interaction plot for the attribute would *use again* illustrating the age and gender effect.

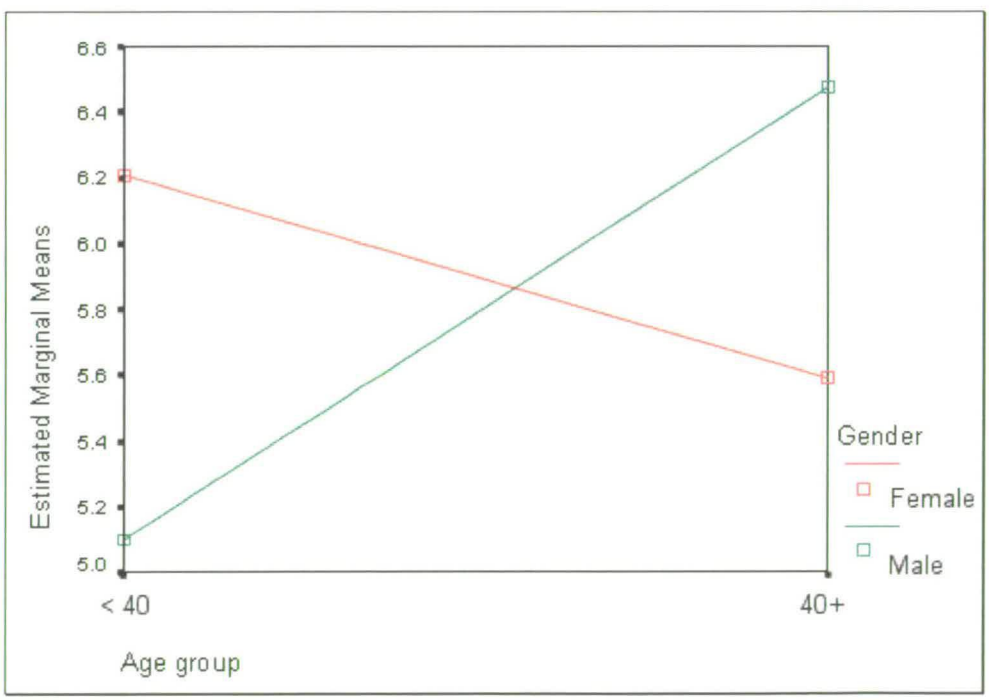


Figure 5.9. Interaction Plot for Between-subjects Age/Gender Effect: Use Again

Similar significant age/gender interactions were found for twelve of the twenty-two individual attributes, with one marginal result, as shown in Table 5.5.

Attribute	Main effect (p)	Comparison	W-S Inter[a] (p)	B-S Effects[b] (p)
Flustered	<.001	F > S	Age .047	Age/Gender .001
Stress	<.001	F > S	Age .011	-
In control	.029	F > S		Age/Gender .001
Procedure	.037	F > S	-	-
Orientation	.039	F > S	-	Age/Gender .001
Page clarity	.047	F > S	Order .047	Age .033; Age/Gender .037
Ease altering details	.055 (NS)	F > S	Order .022	Age/Gender .031
Reliable	.058 (NS)	F > S	-	-
User-friendly	.065 (NS)	F > S	Order .013	Age/Gender .024
Frustration with IA	-	-	Order .011	Age/Gender .031
Liked using	-	-	Order .017	Age/Gender .023
Enjoyment	-	-	Order .009	-
Use again	-	-	-	Age/Gender .004
Concentration (reading)	-	-	-	Age/Gender .051 (NS)
Helpful	-	-	-	Age/Gender .033
Quickly Find	-	-	-	Age/Gender .009
Understood Pages	-	-	-	Gender .032; Order .044; Age/Gender .016

Table 5.5. Interactions and Between-Subject Effects on the Individual Attributes

Notes:
a Within-subject interactions with the main effect
b Between-subject effects regardless of interface design
NS Not significant, but a marginal p value

A number of order effects were also found in attitudes towards the two different interface designs for some individual questionnaire attributes. One such significant effect was for Ease of altering details, $F(1, 24) = 5.967$, $p = .022$, the data are presented in Table 5.6. Participants who experienced the Form metaphor first scored it highly in terms of Ease of altering details, and scored the Spreadsheet metaphor considerably lower, in a paired samples t-test, this was revealed as significant, $t(15) = 3.105$, $p = .007$. However, the difference between interfaces

was not significant for those participants who used the Spreadsheet metaphor first then the Form design, t(15) = -.368, p = .718.

Order	First mean	Second mean
Form then Spreadsheet	6.19	4.69
Spreadsheet then Form	6.19	6.06

Table 5.6. Attitude towards 'Ease of altering details' by Order of Experience

Looking at the scores for the spreadsheet metaphor in terms of the ease of altering details, an independent samples t-test for the different orders of experience revealed them to be highly significantly different, t(22.936) = -2.704, p = .007 (Levene's test was significant, F(30) = 16.781, p < .001). There were no differences in the way the two groups (by order of experience) rated the ease of altering details for the Form design, t(30) = .530, p = .600.

5.6.4. Overall Quality and Preference

Participants used a linear 0-30 point scale to rate the two design metaphors they had used in terms of quality: best to worst. These ratings were collected and used to indicate preference for either interface as well as low or high quality in terms of positions on the scale. The same patterns were found in the quality ratings as were seen in the usability data. The distribution of ratings for the two interfaces were significantly different; favouring the Form metaphor, F(1, 24) = 7.957; p = .009 (see Table 5.7).

Metaphor	Preference Rating	St. Dev.	N	Lower Bound CI	Upper Bound CI
Form	22.27	5.011	32	20.46	24.07
Spreadsheet	16.92	6.391	32	14.62	19.23

Table 5.7. Mean Quality Ratings for the Alternative Metaphors

5.6.5. Preference Rankings

The quality rating were also used to determine a ranked preference for one interface or the other. The preference data indicated that the Form metaphor was strongly preferred to the more complex Spreadsheet metaphor, with 24 (75%) participants ranking the Form design

higher than the spreadsheet design. Only 8 (25%) participants preferring the Spreadsheet design; a binomial test showed a significance level $p = .007$ which is statistically highly significant.

5.6.6. Intention to Use Third Party Payments

The majority, 26 (81%), of the 32 participants expressed an interest in setting up and using third party payment arrangements using the eBanking service, with 2 (6%) undecided and 4 (13%) unlikely to use the facility. Those interested in this service felt that it would be extremely useful.

5.6.7. Qualitative analysis

Participants were interviewed about their experiences with the different eBanking interfaces. Several differences were described between the interfaces. Page layout in particular was mentioned, that the spreadsheet design (S) was more complicated being all on one page, whilst the form design (F) was more intuitive by progressing through a sequence of steps. Participants liked several aspects of the miscellaneous payment functionality: that it was quick and easy; that it increased the functionality of eBanking, and that the designs had clear instructions when needed. Participants who liked the Form design appreciated the guidance they were given with the step-by-step process. The Spreadsheet design, when preferred, was liked due to the flexibility of information being both visible and amendable in one step. These users described it as more convenient and efficient. However, this perception did not extend to the majority of the sample who described it as complicated.

Participant commentary, observations and the individual questionnaire item responses combine to produce a wealth of insight into how participants go about their banking tasks. The combination of methods allowed problems with interfaces to be identified, but also for areas where interaction proceeded intuitively to be noted. The data helped create recommendations for redesigns to be suggested based on what appeared to work well. An example of the approach was that when using the form metaphor, participants started interacting straight away, typing into form inputs. On the corresponding spreadsheet metaphor, they hesitated and stated that they were confused by the number of options and where to start, often resorting to reading some instructions. The recommendations

formulated from these observations and measures of individual attributes included using few form elements per page, presenting them in linear sequences - one per transaction - rather than as an array and choosing the input keywords well to make the interaction intuitive.

Finally, a great deal of confusion was noted concerning the different terms for transactions such as 'Payments' (external) and 'Transfers' (internal). For example, 'Standing orders' (defined as a customer's instruction to their bank to make a regular, fixed amount payment to a specified recipient automatically) were not understood across the cohort. Participants were not always aware of the differences – tending to confuse 'Standing orders' and 'Direct debits' in particular. Typical comments from interview questions are presented in Appendix E, p.305.

5.7. Analysis of the eBanking Usability Questionnaire

5.7.1. Questionnaire Reliability

Individual items in the questionnaire were examined to ensure reliability. Cronbach's alpha (p.37) technique was used to ensure that the questionnaire had high internal consistency. Cronbach's alpha was computed to be .925 for the form metaphor and .949 for the spreadsheet. These values are above the threshold of .8 indicating good inter-item reliability.

Examining the values of α for each item if it were deleted reveals higher α's for the Form design for *concentration (reading)* and *text size*. Both indicated that α when deleted would increase only by .001. Since this is a very small change in α, there is no strong evidence to remove either item.

For the Spreadsheet design, two items also indicated that they would raise α scores if removed, *confusion (layout)* and *text size*. Again, both would only produce a relatively small increase in α to .951 and .954 respectively, offering only weak evidence to remove them from the statement list.

The usability questionnaire has shown high inter-item reliability in the format created for evaluation eBanking interfaces. Almost all the individual items positively contribute to the overall questionnaire sum and mean. The only item which could be considered for deletion would be the attitude to the size of text and only weak evidence that this would be beneficial.

5.7.2. Analysis of Neutral Responses

Generally, participants did not select the neutral response in expressing their attitudes to the eBanking usability characteristics posed in the questionnaire. Out of a total of 704 responses to questions for each interface (22 questions, 32 participants), the Form design was given a neutral score 12.2% of the time, 13.9% for the Spreadsheet metaphor.

For the Form metaphor, the highest frequency of neutral scores (12) was for the attribute *Appearance*. This was scored neutral by 37.5% of the participants. The second highest frequency was much lower for the attribute *Confusion (layout)* which scored neutral only 7 times (by 21.9% of participants).

Similarly, for the Spreadsheet metaphor, the highest frequency of neutral scores (9) was found for the attributes *Appearance* and *Improvement needed*. They were both scored neutral by 28.1% of participants. The second highest frequency (8) was associated with *Frustration with IA*.

This analysis supports the finding that appearance issues were perhaps perceived less important in terms of satisfaction and usability for Banking services, both in terms of the public Web portal and for the authenticated zone of eBanking.

The appearance attribute overall was being scored neutral just less than a third of the time; this is not strong evidence to remove it from the statement set.

5.8. Relationships between Metrics

5.8.1. Usability – Attitudes and Preferences

There was a strong correlation between usability attitudes and preference (quality) ratings for the differences between the two metaphor designs (F-S). The scatterplot in Figure 5.10 shows a positive association between the two metrics, $r = .789$ ($p < .001$). This is a reasonably large effect and a high positive association.

For the Form metaphor, the results were similar with a highly significant positive correlation $r = .635$, $p < .001$, $R^2 = .415$. For the Spreadsheet metaphor, the correlation was also highly significant and positive $r = .470$, $p = .007$, $R^2 = .221$.

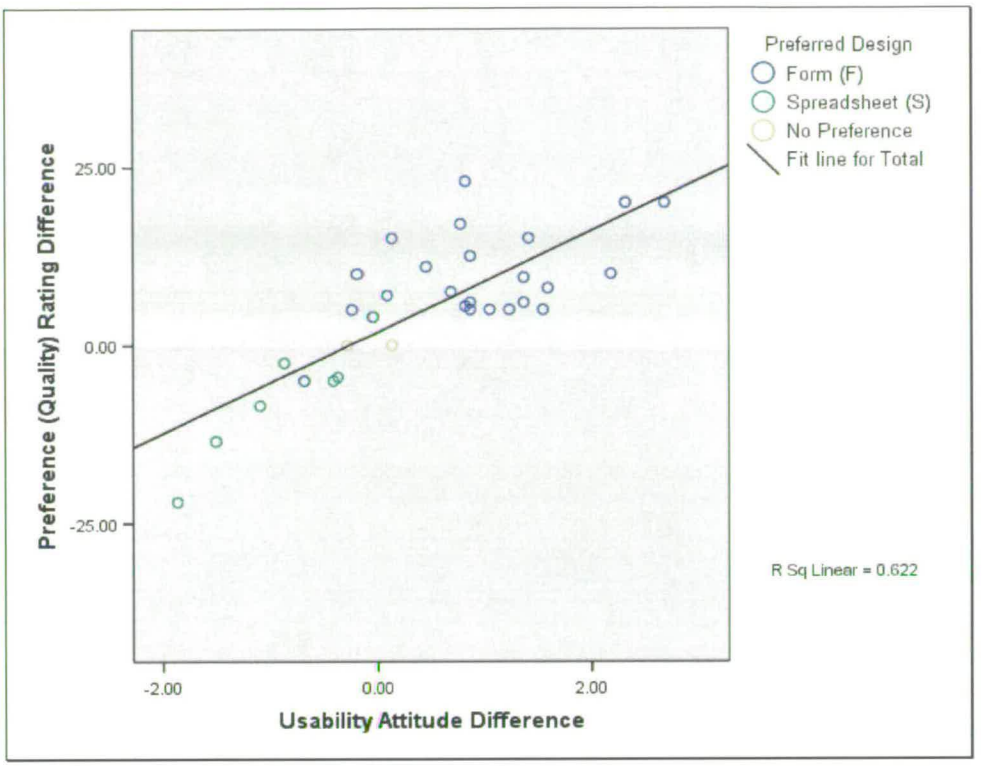


Figure 5.10. Usability Attitude – Preference Relationship (Metaphor Differences)

Correlations were within the ranges already computed for the two pilot experiments. They were all highly significant at the $p < .05$ level. The mean of the usability questionnaire is positively related to the preference score each interface for eBanking payments.

Individual Attributes in the eBanking Usability Questionnaire

In order to determine which characteristics in an eBanking interface were most highly correlated with overall preferences, the correlations were computed for each attribute in the questionnaire, first in terms of differences, then in absolute scores for each individual interface metaphor, see Figure 5.11 and Table 5.8.

The tasks and interfaces evaluated in terms of using an eBanking service considered much more specific needs in a fairly focused information space than was the case of the pilot experiments on the Web portal designs. Due to the interface variables being studied, the tasks were restricted to interactions in the payments section of the eBanking service. Therefore, the salient characteristics of the interface were likely to be somewhat different from information retrieval tasks in a comprehensive, broad and generic Web site.

From the differences in scores between interfaces, it was apparent that navigation, speed finding and page clarity were still important, alongside user-friendliness, like using and enjoyment. Understanding the page content, the ease of changing payments and whether the site would be used again also correlated highly. For the eBanking tasks, it was text size, concentration reading and reliability which were less highly correlated. All items were positively associated with preferences.

For the form design, attributes in the category *affect* were also salient in their high association with preference scores. Text size, appearance and concentration reading were less important with hardly any association.

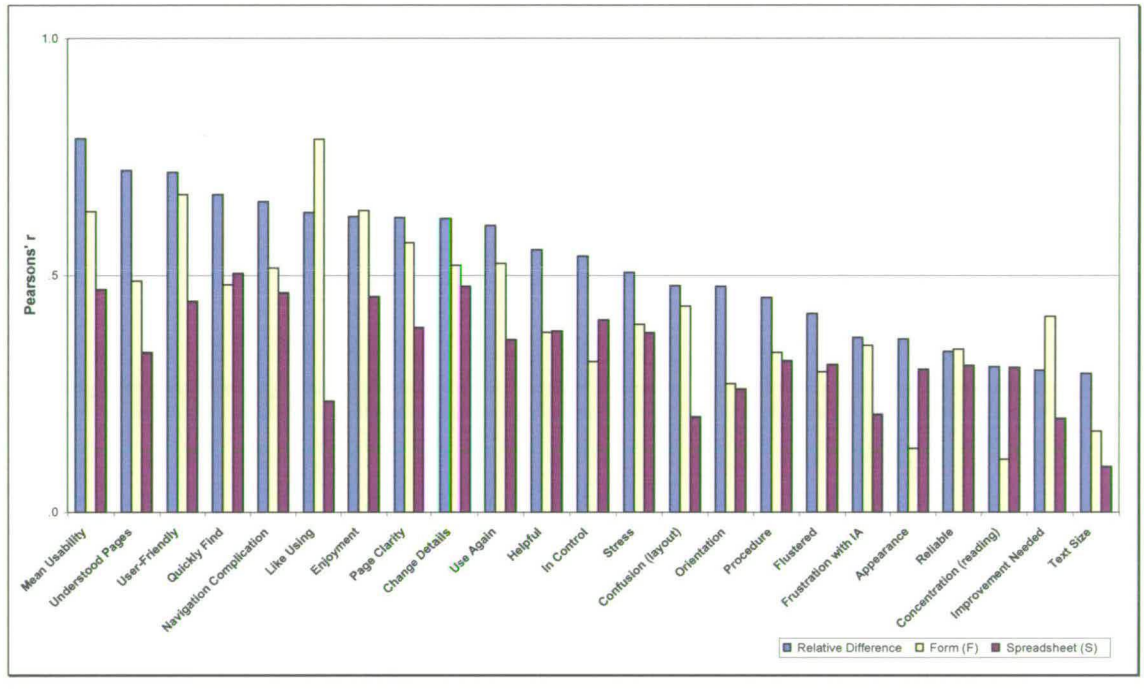


Figure 5.11. Correlations for Usability Attributes and Preference (Quality) Ratings

For the spreadsheet metaphor, the correlations were not as high, with only speed finding breaking $r > .5$ (at $r = .504$). The ease of changing details was noticed as being highly associated with preference.

Attribute	Difference F-S		Form (F)		Spreadsheet (S)	
	r	*p*	*r*	*p*	*r*	*p*
Understood Pages	.722	<.001	.488	.005	.337	.059
User-Friendly	**.718**	**<.001**	**.671**	**<.001**	**.445**	**.011**
Quickly Find	**.671**	**<.001**	**.480**	**.005**	**.504**	**.003**
Navigation Complication	**.656**	**<.001**	**.516**	**.002**	**.463**	**.008**
Like Using	.633	<.001	.788	<.001	.234	.197
Enjoyment	**.624**	**<.001**	**.637**	**<.001**	**.455**	**.009**
Page Clarity	.622	<.001	.569	.001	.390	.027
Change Details	**.620**	**<.001**	**.521**	**.002**	**.477**	**.006**
Use Again	.605	<.001	.525	.002	.364	.040
Helpful	.554	.001	.380	.032	.382	.031
In Control	.541	.001	.318	.077	.406	.021
Stress	.506	.003	.397	.025	.379	.033
Confusion (layout)	.478	.006	.435	.013	.201	.270
Orientation	.477	.006	.271	.133	.260	.151
Procedure	.453	.009	.337	.059	.320	.074
Flustered	.419	.017	.296	.100	.312	.082
Frustration with IA	.369	.038	.352	.048	.207	.257
Appearance	.366	.039	.135	.462	.301	.094
Reliable	.339	.058	.344	.054	.309	.085
Concentration (reading)	.307	.087	.112	.543	.306	.089
Improvement Needed	.300	.096	.413	.019	.198	.278
Text Size	.293	.104	.171	.350	.096	.601
Mean Usability Score	**.789**	**<.001**	**.635**	**<.001**	**.470**	**.007**

Table 5.8. Significant Correlations between Usability Attributes and Preference Scores

Where correlations are all highly significant for all three computations, and above $r = .4$ (lowered from .5 to account for the relatively lower scores obtained for the Spreadsheet metaphor), this indicates the characteristics of the eBanking interface which associate well with final preference scores. User-friendliness, speed finding, navigation, enjoyment and the ease of making payment arrangement changes were highest associated with preferences.

Again, these correlations are generally higher than previous publications would suggest, but with a wider range than in the pilot experiments where highly significant correlations were in the range $.387 < r < .741$ (see p.108).

The mean scores from the usability questionnaire appear to offer a good summary of the characteristics, with 28 highly significant attributes and 14 significant aspects of the interface having high relationships with final preference scores (see Table 5.8). For example, looking at the score differences for usability questionnaire items and preference scores, there were 15 highly significant correlations (attributes: understood pages, user-friendly, quickly find, navigation complication, liked, enjoyment, page clarity, change details, use again, helpfulness, control, stress, layout confusion, orientation and procedure) and 3 significant correlations (attributes: flustered, frustration with IA and appearance).

5.8.2. Usability – Attitudes and Performance

Usability attributes, including the mean scores, may be related to measures of task completion (performance), although data from the first pilot was not conclusive. The correlation matrix was computed for the differences between the two interfaces. Like Pilot experiment 1, this data set showed a significant correlation between task performance differences and usability mean score differences, $r = .497$ ($p = .004$), a positive direction and a relatively large effect, just short of .5. This is within the range found by Hornbæk & Law (Hornbæk & Law, 2007) and again corroborates the notion that effective performance has a positive association with satisfaction.

There were also some individual aspects of usability which significantly correlated with the task performance metric for the differences between the two metaphor designs: *easy to change arrangements*, $r = .667, p < .001$; *reliable*, $r = .620, p < .001$; *in control*, $r = .512, p = .003$; *understood pages*, $r = .508, p = .003$; *user-friendly*, $r = .502, p = .003$; *quickly find*, $r = .468, p = .007$; *page clarity*, $r = .464, p = .007$; *like using*, $r = .449, p = .010$; *enjoyment*, $r = .437, p = .012$; *navigation complication*, $r = .419, p = .017$; *would use again*, $r = .403, p = .022$; *flustered*, $r = .367, p = .039$.

For the individual interfaces, the results were very different. The Form metaphor was the preferred design overall, with no significant correlation between task performance and mean usability scores, $r = .164, p = .369$ (although the relationship was in the positive direction). In addition, only one of the twenty-two attributes in the questionnaire had a significant correlation with the performance measure: *page clarity*, $r = .383, p = .030$.

However, for the less preferred interface – the spreadsheet metaphor, there was a significant correlation overall between mean usability and performance, $r = .572, p = .001$. Similarly,

for individual attributes: *easy to change arrangements*, $r = .675, p < .001$; *reliable*, $r = .669$, $p < .001$; *concentration (reading)*, $r = .500, p = .004$; *in control*, $r = .529, p = .002$; *understood pages*, $r = .444, p = .011$; *user-friendly*, $r = .589, p < .001$; *quickly find*, $r = .462$, $p = .008$; *page clarity*, $r = .416, p = .018$; *enjoyment*, $r = .579, p = .001$; *navigation complication*, $r = .431, p = .014$; *would use again*, $r = .420, p = .017$; *flustered*, $r = .477, p = .006$; *appearance*, $r = .377, p = .034$; *helpful*, $r = .461, p = .008$; *improvement needed*, $r = .389, p = .028$.

5.8.3. Usability - Performance and Preference

The relationship between task performance and final preferences was also explored. Table 5.9 displays the correlations for the differences, then each interface individually.

Again, similarly to the performance/attitude relationship, there was not a significant association for the Form interface alone – regardless of task performance, this interface was scored highly preferred and 'best' in terms of quality. On the other hand, the difference in task performance on the two interfaces did relate significantly to differences in preference ratings. Similarly, task performance on the Spreadsheet interface was highly correlated with performance.

	Differences (F-S)	Form (F)	Spreadsheet (S)
r	.500	.269	.486
p	.004	.136	.005

Table 5.9. Correlations between Performance (Task Completion) and Preference

Compared to the correlations in the Web Portal experiments (p.88 and 110), performance in the eBanking tasks had a more consistent positive association with preference ratings, but in terms of strength and significance, no clear pattern was established.

5.8.4. Comparison of Metrics

The mean usability and quality ratings are displayed comparatively using normalised scales (as a fraction of the total score possible) in Figure 5.12. The error bars indicate the 95% CI of the means.

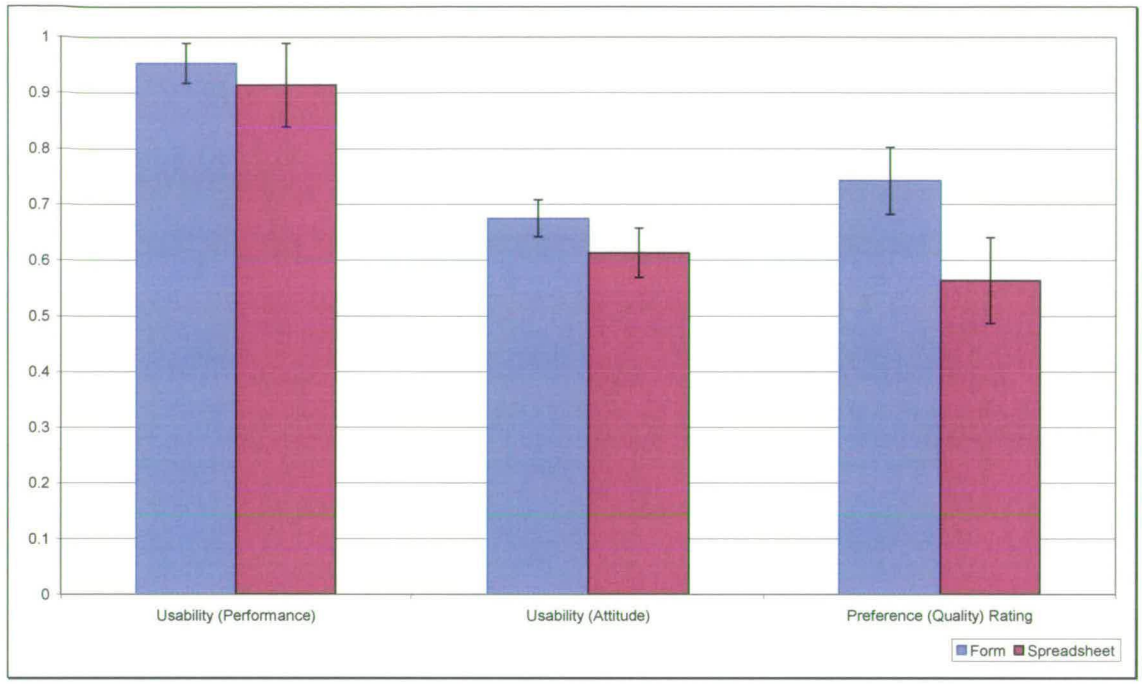


Figure 5.12. Comparison of Normalised Usability Metrics for the Alternative Metaphors

They illustrate the wider range of performance, usability and preference scores for the spreadsheet design compared to the form metaphor. The range displayed by the preference ratings shows no overlap in CIs, clearly illustrating the perceived superiority of the form design in terms of participants' overall preferences.

5.9. Experiment 1: Interface Metaphors – Discussion

eBanking services provide a low-cost channel desired by banks to be used by a wide range of people. The Form metaphor was the appropriate choice for account transactions on the interface. The usability metrics of overall attitudes and preferences were significantly different, higher than the alternative interface and generally good in terms of a wide range of usability and interface characteristics. The Form metaphor was identified as the right design to progress to full development. This finding potentially extends to other online financial transactions. The eBanking user base is likely to be highly diverse and the linear approach was highly usable across the participant sample. The Form design is also straightforward to develop.

The editable table of the Spreadsheet design was unfamiliar to users. Despite this, a sizeable minority preferred the Spreadsheet. It is speculated that they may have financial savvy: familiarity with numerical work, finance or wide experience with online transactions. These are possible characteristics which may divide the population. Future work towards the design of a payment interface for high-end users might benefit from such comparisons, should reliable methods to divide the population in these terms be found.

Between-subject effects and interactions on overall usability due to age or gender variables were not significant in this experiment. However, the older female group had lower attitudes to some specific navigation and clarity aspects of the designs. This group also reported somewhat lower Internet use which might account for their lower attitudes. This group also showed some distinctive attitudes towards interface characteristics in the pilot experiments, suggesting that these between-participants factors are important to consider when evaluating the Web and eBanking services.

Attitude toward a broad range of usability and specific interface characteristics was measured by the questionnaire. Overall usability evaluated by this metric as experienced during hands-on use was highly positively correlated with the final preference ratings for the two metaphors. Correlations ranged between .470 and .789 (Table 5.8, p.152), higher than coefficients previously published (Hornbæk 2006). The strong positive and highly significant correlation coefficients computed for these interfaces (individually and in comparison) confirms that the usability questionnaire provided data which was highly related to overall quality ratings and preference tendencies.

5.9.1. Limitations

Experiment 1 on eBanking interface metaphors has provided some of the first hands-on usage data concerning usability and interface design characteristics for eBanking. However, the evaluation involved a fairly small sample, not all of whom were active users of the eBanking service. There were some difficulties recruiting large groups of experienced users of the specific eBanking service, resulting in very limited data on how expert users would perform on the various interfaces. However eBanking sessions are typically short, sporadic (Sarel & Marmorstein, 2003) and use is discretionary. The collection of first-time use data seems appropriate given this framework. However, combining more experienced users into an evaluation would be a future goal of eBanking usability research, once a critical mass of

regular, experienced and 'power' users are available. The perceived usability and preferences of experienced users could possibly be represented by participants given a full training session (to model learned usage). Such data may be of interest to further support the development of the linear or spreadsheet approach, a longitudinal study may also collect interesting data in this regard. Neither option was explored further in this thesis, which concentrates only on the large-scale, early evaluation of interface design ideas to indicate the most potentially successful, and in the metrics proposed to determine this.

It is possible that the scope of just four banking tasks to perform on each interface provided too short an experience to explore the designs. This aspect of the experiment design may have been particularly disadvantageous in evaluating the Spreadsheet design. The spreadsheet metaphor may become more effective and efficient after a longer period of use.

It may be the case that a spreadsheet metaphor would be more appropriate for banking staff in an environment where efficiency dominates and training can be provided. Evaluations in this context may involve task timing, and would also need to consider the learning or training time required. These extensions are not really appropriate in the context of self-service eBanking applications.

Finally, the use of task sheets may have proved too unrealistic in this context as no participants were observed to try the multiple task completion afforded by the Spreadsheet design. It is possible that a set of realistic bills and documents, arranged as a bundle rather than a list, may be more appropriate for investigating this feature. However, it may also be true that eBankers do not typically perform such multiple tasks. This question may benefit from field research to complement controlled study.

5.9.2. Outcome and Follow up

One of the major outcomes of the interface evaluations for eBanking transactions was the confusion noted in the use of banking jargon for transaction terms and the separation of different types of transaction into sections based on these terms. Having determined a successful interaction metaphor on which to base self-service transaction functionality, the research went on to explore a more generic, plain language version of the service. This study was also used as a platform to investigate possible relationships between the usability metrics and determine the salient characteristics of usable eBanking interface designs.

5.10. Experiment 2: Dialogue Styles

5.10.1. Experiment design

The dialogue style experiment investigated two alternative wording styles within the Form metaphor for payments in eBanking. Typical banking jargon was used in instructions and labels in one interface, referred to as the Formal dialogue style. Simple, 'plain English' language was substituted in the second experimental interface, referred to as the Plain Language dialogue style. A total of 29 participants took part in the experiment; none had been involved in the metaphor experiment. The sample was balanced for age, gender and presentation order of the interfaces, the between-subject factors. The dependent variables were the responses to individual items in the usability questionnaire and the quality rating and preferences. Task completion and error counts were also recorded.

5.10.2. Experiment materials

Participants were asked to perform three tasks (see Appendix F, p.314). Tasks were a transfer, a bill payment (from set-up to paying the bill) and similarly to set-up and pay a miscellaneous payment; tasks focused on the contrasting interface dialogues. They performed the same tasks with the second interface, but specific details were changed to ensure engagement with the second design.

5.10.3. Experiment Design Summary

Design: Two cell, repeated measures, within subjects.

Independent Variables: Interface F (Form/Formal)

Interface PL (Form/Plain Language).

Participant Independent Variables: Age (2 levels)

Gender (2 levels)

Order of experience (2 levels)

Dependent Variables: Attitude questionnaires

Task completion & error counts

Quality ratings & preferences.

Confounding variables: Researcher Bias (randomised)

Room (randomised)

Task sheet (balanced / matched task per task).

Other data: Think aloud remarks and researcher observations

Interview questions

Intention to use

Sample size: 2 orders x 2 genders x 2 age groups x over-sampling 4 = 32

Honorarium: £20

Session Time: 1 hour

5.11. Experiment 2: Results for Dialogue Style

5.11.1. Participants

Recruitment difficulties resulted in a sample of only 29 participants completing this experiment, this was a few less than the 32 desired for full factorial statistical comparison.

Participants in the 40+ age group were slightly more difficult to recruit with the Internet-savvy requirement. The resulting participant sample consisted of 14 (48%) females and 15 (52%) males. It was composed of just over 60% in the under 40s age group. Again, most participants made frequent use of the Internet, accessing it mainly from home or work, with a high proportion, 39 (89%) doing both.

Some 15 (52%) of the participants used the Bank's eBanking service weekly or more frequently. Some 95% of all transactions performed consisted passive viewing; the other transactions were examples of transactional interactions and a slightly reduced figure of 11 (38%) of this sample described themselves as predominantly using Internet Banking to conduct their finances.

5.11.2. Performance

Most participants completed the tasks in the experiment without problems, see Table 5.10. There were fairly few errors made, and none were observed as catastrophic (not corrected).

Dialogue	Performance	St. Dev.	N	Lower Bound CI	Upper Bound CI
Formal	2.793 (93.1%)	0.4913	29	2.606 (86.9%)	2.980 (99.3%)
Plain Language	2.897 (96.6%)	0.3099	29	2.779 (92.6%)	*

Table 5.10. Task Completions Sum (Rate) for Alternative Dialogue Styles

*When completion rates are very high and close to 100%, particularly for small sample sizes, the upper CI computed by the Wald method results in an upper bound above 100%. Therefore the Adjusted Wald CIs were also computed for this data using the LaPlace point estimate as the midpoint of the interval (Lewis & Sauro, 2006; Sauro & Lewis, 2005).

By this method, the completion rates were:

- Formal Design Completion Rate 92.1% (Lower bound 85.5%, Upper bound 97.1%)
- Plain Language Design Completion Rate 95.5% (Lower bound 89.9%, Upper bound 99.2%)

There were no statistically significant differences between the two interface dialogue styles, as computed in a paired samples t-test, $t(28) = -1.14$, $p = .264$. In a repeated-measures ANOVA there were no age or gender differences in terms of the rates of task completion.

There was also no difference between interfaces for the tasks completed correctly first time with no errors requiring correction. The proportion of errors was very small, 22 (76%) participants completed all their tasks correctly first time on the Formal interface, 23 (79%) on the Plain Language interface.

5.11.3. Attitudes

To investigate the impact of the differing dialogue styles, the mean responses to the usability questionnaire completed after each exposure were subjected to repeated-measures ANOVAs. Individual items on the usability questionnaire were also compared in the same way. Three between-subject factors were included in the analyses as follows: age group, gender and order of experience.

The mean usability scores (7-point response scale) for the Formal and Plain English styles are presented in Table 5.11.

Dialogue	Usability	St. Dev.	N	Lower Bound CI	Upper Bound CI
Formal	5.2790	0.7930	29	4.9774	5.5806
Plain Language	5.3025	0.7138	29	5.0310	5.5740

Table 5.11. Mean Usability Attitudes for the Alternative Dialogue Styles

There was no significant difference between the overall attitudes to these two dialogue styles for the 29-participant sample $F(1, 21) = 0.008$; $p = .931$; $N = 29$. There were also no significant differences in the ways that men and women, or participants of different age groups, responded to the two dialogue styles. There were also no main effects due to the order of presentation of the interfaces.

The ANOVAs calculated on individual usability attributes showed few differences, interactions and between-subject effects. All other comparisons were not statistically significant. The comparisons where effects were found are shown in Table 5.12 and the chart in Figure 5.13.

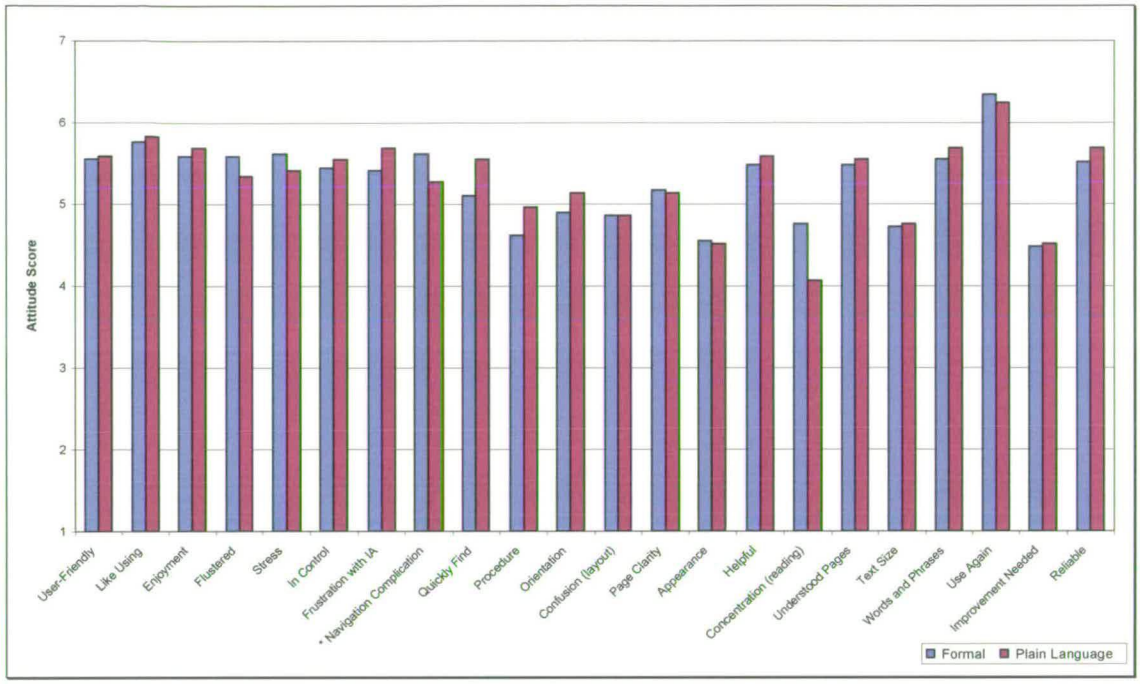


Figure 5.13. Comparison of Usability Attribute Scores for the two Dialogue Styles

Only two attributes showed any trend toward significance: navigation complication was just significant, with the formal design being scored higher than the Plain language dialogue. Marginally significant was the degree of concentration required to read pages, again the trend here was favouring the Formal dialogue.

There was a highly significant within-subject order effect, $F(1, 21) = 10.520$, $p = .004$, for the attribute: degree of confusion with layout, showing that the second interface experienced caused less confusion than the first. The effect was similar for both sequences (Formal then Plain Language and vice-versa). The similar interaction modes across the two interface designs (menu driven, hyperlinked text, with simple form-fill dialogues) may have resulted in a learning effect for these attributes.

Attribute	Effect (*p*)	Direction	W-S Inter (*p*)	B-S Effects (*p*)
Navigation Complication	.049	F > PL	-	Age/Gender .038 (F1, M2 > M1, F2)
Concentration (reading)	.073*	F > PL	-	-
Confusion (layout)	-	-	Order .004 (F-PL: F<PL; PL-F: F>PL)	Age/Gender .053* (F2 < M1, F1, M2)
In Control	-	-	Gender .048 (F: F>PL; M: PL>F)	Order .014 (F-PL < PL-F)
Enjoyment	-	-	-	Age/Gender .029 (M2, F1 > M1, F2)
Helpful	-	-	-	Age .028 (Y > O) Age/Gender .039 (F1 > M1, M2 > F2)
Procedure	-	-	-	Age .004 (Y >> O) Gender .017 (M > F)
Stress	-	-	-	Age/Gender .055* Order .055*
Appearance	-	-	-	Order .042 (F-PL > PL-F)

Table 5.12. Interactions and Between-Subject Effects on the Individual Attributes

Between-subjects effects were mainly age/gender interactions, typically older females were giving strongly lower attitudes towards attributes than other groups. Again, this might be attributed to the relative lack of computing experience in this group generally, but this was not a controlled factor in this experiment. There were some tendencies for older participants to give lower scores than their younger counterparts

No significant differences or order effects were found when comparing the results to the usability question pertaining to the understanding of the words and phrases used on the two interfaces, even though this question was posed specifically to address these interface changes. Participants did not really notice the language changes as the difference between the interfaces (from qualitative data collection) therefore this result is consistent with the general observations resulting from the experiment sessions.

5.11.4. Preference (Quality) Ratings

Participants used a linear 0-30 point rating scale to rate the two interfaces from best to worst. These scores were collected and used to indicate both a quality rating, and a ranked preference for either interface. The same patterns were found in the quality ratings as were seen in the usability data. That is, no significant difference between the two interfaces. The data are shown in Table 5.13.

Dialogue	Preference	St. Dev.	N	Lower Bound CI	Upper Bound CI
Formal	20.39	5.651	29	18.24	22.54
Plain Language	20.13	5.102	29	18.19	22.07

Table 5.13. Mean quality ratings for the contrasting dialogues

5.11.5. Preference Rankings

The quality scores were also used to determine a ranked preference for one interface or the other. The preference data indicates that the Formal dialogue style was preferred by 17 (59%) of the 29 participants, with the other 12 (41%) preferring the Plain Language dialogue style. A binomial test showed that the groups were not significantly different. There were no undecided votes despite participants typically failing to notice any explicit differences between the interfaces they used.

5.11.6. Intention to Use Third Party Payments

All 29 (100%) of the participants expressed an interest in performing third party payments on eBanking. They agreed with the participants in the metaphor experiment, feeling that this service would be extremely useful - allowing eBanking to perform similar functions to automated telephone-banking services and human-operated channels. This was an extremely high rate of interest, or usage intention for making payments to companies and other individuals using the online service. It seemed that the actual experience of filling in new payment details may have been related to this increased interest. In fact, many participants in this group expressed surprise at how easy it was to set up new payments themselves on eBanking, avoiding the hassle of calling or visiting the Bank.

5.11.7. Qualitative Data

Participants were interviewed about their experiences with the different dialogue styles. Of the 29 participants, just under one third (9, 31%) mentioned the language changes, commenting mainly on wording changes in the dialogue boxes. Although many participants said they noticed slight differences, only few were able to describe the differences accurately. The characteristics that they identified did not typically include the language tone (formal or plain English), or the removal of banking terms such as standing order from the site. In fact those who did often suggested that the plain English version was more formal, or that they preferred the formal tone of that version, the opposite of the desired variable change.

As detailed comments about the individual interfaces were desired, reminder screen pictures were given to participants before asking them what they liked, disliked and could suggest to improve. They liked the interfaces generally, their range of functions and found the linear steps clear and easy to use. They disliked a range of interface characteristics from the similarity of the buttons for making, changing or deleting payments, the pop-up dialogues to menus and the colour scheme.

Participants frequently mentioned that when they read instructions on Web pages, they tended to look over them very quickly as they want to get on with using a service, not reading about it. In general, they were very happy with the amount of information they had to read, stating that it was clear, concise and simply written.

The major finding was that participants who used eBanking had not explored the provision to set up any payments online. The idea of allowing users to send money to friends and other specific bank accounts was highly praised for convenience and speed.

Again, some confusion was observed regarding the different terms for banking transactions, particularly the difference between Direct Debits other transactions. Participants were also unclear as to whether 'Transfers' could be made to external accounts. However, almost all participants expressed the need for financial terms to be consistent across branch, telephone and Internet channels. Many suggested adding a glossary to educate them in the correct usage of transactions and terms. Although some participants were very positive about the use of plain English terms like one-off and regular to distinguish payment types, a very slight majority (14, 48% c.f. 12, 41%) preferred the traditional terms and the clear distinction between them (even if they didn't fully understand these distinctions). Many participants expressed a desire to understand banking terms to give them the confidence to use self-

service channels. A single page containing details of all outgoing monies was often requested. Typical comments from interview questions are presented in Appendix F.

5.12. Analysis: The eBanking Usability Questionnaire

5.12.1. Questionnaire Reliability

The reliability of the questionnaire items was computed using Cronbach's alpha (see section 2.3.2, p.37). The alphas were .859 for the differences between the two dialogue styles, .923 for the formal language interface, and .913 for the plain language version. All three values are above the threshold expected, suggesting that individual items all positively contribute to the overall questionnaire sum and mean.

By examining the alphas when items are deleted, there were several candidates which, if removed, could increase the questionnaire reliability. For the differences between the two interface dialogues, deleting the item *appearance* would result in a higher alpha of .867; similarly for *text size* with an alpha of .865. Deleting *reliable* would result in a very slight increase to $\alpha = .861$; and *concentration reading* to $\alpha = .860$. The first two interface characteristics show more substantial variation from the questionnaire alpha, indicating possible candidates for removal.

Looking at the individual designs, the Formal language version also indicates the possible removal of the *text size* attribute, with $\alpha = .935$ – again, a substantial difference to the alpha for the full set (.923). Two attributes showed minor increases, deleting *appearance* results in an alpha of .928, removing *reliable* $\alpha = .924$. This indicates more evidence toward text size as an interface characteristic behaving differently to the other attributes.

Finally, for the Plain Language version, only one item resulted in a higher alpha for the scale when deleted. Again the attribute was *text size*, removing the attribute brought $\alpha = .927$, again substantially different to the original alpha (.913).

This combines to indicate some evidence that *text size* in this experiment was not related as strongly to the other usability attributes and interface characteristics. In fact, outside the laboratory, text size in a browser is under user-control through their own preferences. Some research on text size on the Web has already been undertaken (Ling & van Schaik, 2006 and 2002; Bernard et al., 2003).

5.12.2. Analysis of Neutral Responses

Analysis of the attributes which scored the neutral response of 4 on the 7-point Likert scale show around one third of participants scored *appearance* neutrally for both interfaces: 9 participants (31.0%) for the Formal language version, 10 participants (34.5%) for the Plain Language version. Another high number of neutral responses was seen for the item *stress* for the plain language interface, with 7 participants (24.1%) having no strong attitudes to this attribute.

Appearance aspects of eBanking design and transaction interactions appear to have less relevance to customers and users in this context than in other environments on and off the Web.

However, the strength of the evidence to delete appearance from the metric is not very strong, as almost two thirds of the cohort generally responded non-neutrally to this statement.

In fact, for both designs, participants did not generally select the neutral response in reaction to items in the questionnaire. Out of a total of 638 responses to questions for each interface (22 questions, 29 participants), the Formal version was only given a neutral score to questions 9.9% of the time, 11.6% for the Plain Language version. This is in fact less than the values seen in the other experiments.

5.13. Relationships between Metrics

5.13.1. Usability – Attitudes and Preference (Quality)

There was a strong correlation between perceived usability and preference - a Pearson correlation analysis showed a highly significant positive correlation, $p < .001$. The correlation was similar to that seen in the metaphor experiment, and the data are illustrated in Figure 5.14, Pearson's $r = .616$, $R^2 = .379$. This is a reasonably large effect and a high positive association.

The possibility of outliers in a small group is difficult to determine as they may just represent opinions which were under-represented in the gathering of the random sample. In these experiments, participants were not removed unless there were procedural errors in the session or they did not match the target participant sample (e.g. not Internet users).

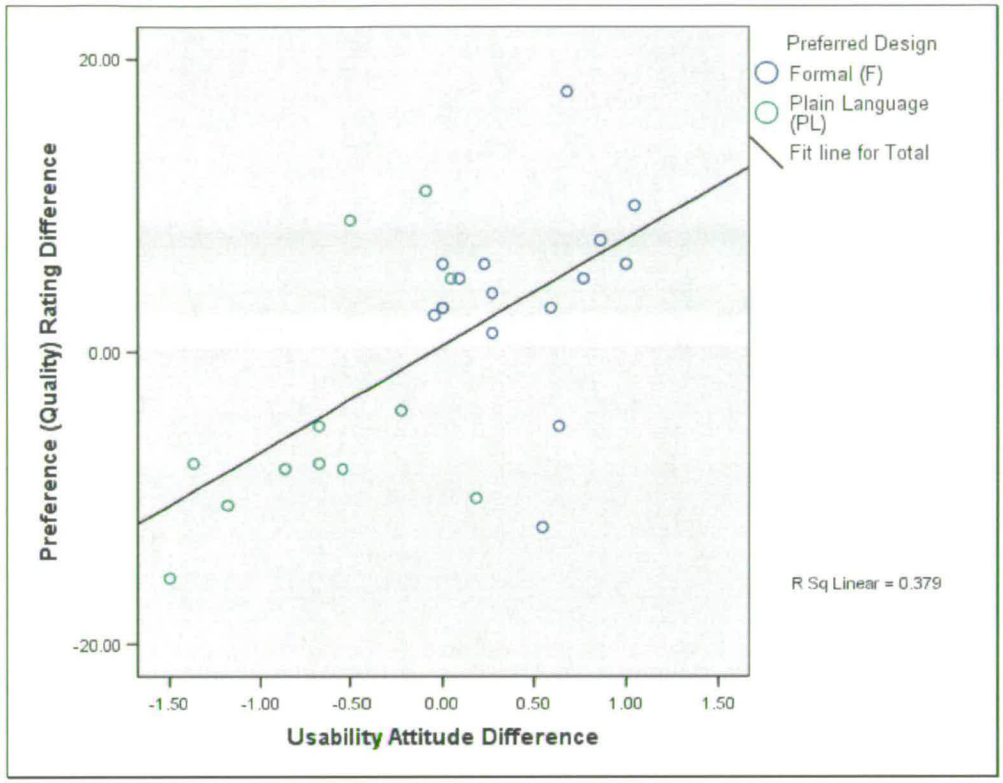


Figure 5.14. Usability Attitude – Preference Relationship (Dialogue Differences)

For the Formal dialogue, the results were similar with a highly significant positive correlation $r = .630$, $p < .001$, $R^2 = .397$. For the Plain Language dialogue, the correlation was also highly significant and positive $r = .519$, $p = .004$, $R^2 = .269$.

Correlations in this experiment were highly significant and positive, as for the metaphor experiment. The mean of the usability metric is positively related to the preference score each interface when evaluated in terms of making eBanking payments.

Individual Attributes in the eBanking Usability Questionnaire

In order to determine which characteristics in a transactional eBanking interface were most highly correlated with overall preferences, the correlations were computed for each attribute in the questionnaire, first in terms of differences, then in absolute scores for each individual interface dialogue design, see Figure 5.15. and Table 5.14 (ordered by magnitude of correlations for the Formal design).

In this evaluation tasks and interfaces were specifically examined in the context of setting up and making new payments to third party accounts. It was of interest to see which aspects of the interface designs would have high relationships with participants' final preferences in this context. The tasks were slightly different to those in the metaphor experiment, thus adding to the range of uses for which these relationships are being explored.

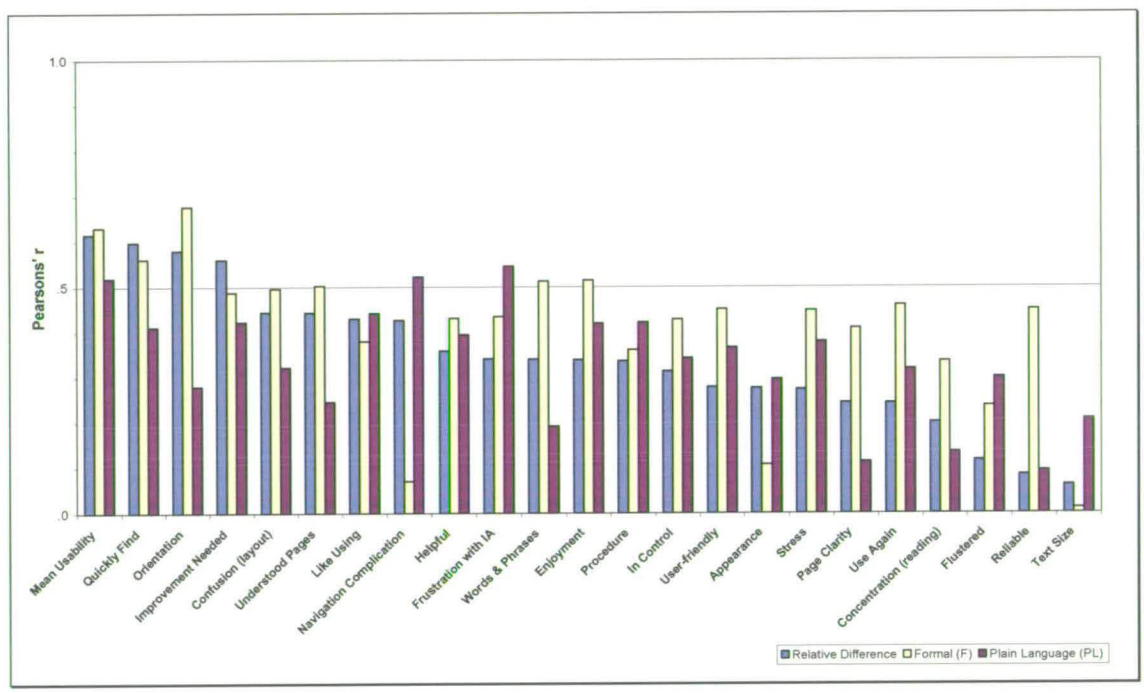


Figure 5.15. Comparison of Pearson Correlation Coefficients for Individual Attributes & Preference Rating.

From the differences in scores between interfaces, it was apparent that the speed it took to find appropriate pages, orientation and the lack of improvement required were the most important attributes related to preferences. Also highly significantly related were layout, comprehension of content, navigation and attitude towards using the service. For the

eBanking payment tasks, it was text size, reliability and flustered which were less highly correlated. Similarly, concentration reading was not highly correlated with preferences, which is corroborated by the qualitative data observing that pages were quickly scanned and instructions rarely needed when interacting with the form metaphor and indicating the lesser impact of dialogue in a simple form-fill exercise.

Attribute	Difference F-PL		Formal (F)		Plain Language (PL)	
	r	*p*	*r*	*p*	*r*	*p*
Orientation	.580	.001	.677	<.001	.279	.142
Quickly Find	***.597***	***.001***	***.561***	***.002***	***.410***	***.027***
Enjoyment	**.339**	**.072**	**.514**	**.004**	**.419**	**.024**
Words & Phrases	.340	.071	.512	.005	.192	.319
Understood Pages	.442	.016	.502	.006	.245	.200
Confusion (layout)	.443	.016	.496	.006	.322	.089
Improvement Needed	***.560***	***.002***	***.487***	***.007***	***.422***	***.023***
Use Again	.243	.204	.459	.012	.319	.092
User-friendly	.278	.144	.450	.014	.366	.051
Reliable	.084	.664	.449	.015	.093	.631
Stress	.273	.151	.447	.015	.379	.043
Frustration with IA	**.340**	**.071**	**.434**	**.019**	**.545**	**.002**
Helpful	**.358**	**.057**	**.430**	**.020**	**.394**	**.035**
In Control	.313	.098	.427	.021	.342	.069
Page Clarity	.243	.203	.408	.028	.113	.559
Like Using	**.429**	**.020**	**.380**	**.042**	**.441**	**.017**
Procedure	.336	.075	.360	.055	.421	.023
Concentration (reading)	.201	.297	.335	.076	.135	.486
Flustered	.117	.547	.237	.217	.300	.114
Appearance	.276	.147	.107	.580	.296	.119
Navigation Complication	.426	.021	.070	.718	.522	.004
Text Size	.062	.751	.011	.953	.207	.282
Mean Usability Score	**.616**	**<.001**	**.630**	**<.001**	**.519**	**.004**

Table 5.14. Significant Correlations between Usability Attributes and Preference Scores

For the formal dialogue design alone, attributes related to orientation, speed finding the information, enjoyment and understanding of words and phrases were also highly associated

with preference scores ($r > .5$). There were also many other attributes significantly associated with preference at $r > .4$. Text size, appearance and navigation complication were considered less important in this evaluation, with hardly any association with the final preference rating.

For the Plain language design, the correlations with preference were not as high, but IA and navigation complication were highly associated at $r > .5$. Similarly, attitude to using and the degree of improvement needed were also highly related to preferences. Page clarity, reliability and concentration reading were barely associated with preferences.

Where correlations are all highly significant for all three computations, and above $r = .4$ (lowered from .5 to account for the relatively lower scores obtained for the Plain Language design), this indicates the characteristics of the eBanking interface which associate well with final preference scores. As no difference was found between the two designs, scores for the individual interfaces, not the comparison were concentrated on. *Speed finding*, *enjoyment* and *improvement needed* were highest associated with preferences across the tasks and interfaces.

The mean scores from the usability questionnaire appear to offer a good summary of the characteristics.

5.13.2. Usability - Attitudes and Performance

In the dialogue experiment, data showed no significant correlation between task performance differences and usability attitude differences, $r = .299$ ($p = .115$). Although the relationship was in a positive direction, the effect was very small. However, it is within the range found by Hornbæk & Law (Hornbæk & Law, 2007) and again corroborates the notion that effective performance has a positive association with satisfaction.

There were some individual aspects of usability which significantly correlated with the task performance metric for the differences between the two dialogue designs. Firstly there was a negative association between *appearance* and task performance, $r = -.534$, $p = .003$. The other significant relationships were positive, *orientation*, $r = .430$, $p = .020$; *procedure*, $r = .408$, $p = .028$; there was also a marginally significant relationship between task performance and *stress*, $r = .361$, $p = .055$.

For the individual interfaces, the results were slightly different. The Formal dialogue was the preferred design overall, although not significantly so. There was a marginally significant

correlation between task performance and mean usability attitude scores for this dialogue design, $r = .328, p = .082$, and the relationship was in the positive direction. In addition, four of the twenty-two attributes in the questionnaire had a significant correlation with the performance measure: *confusion (layout)*, $r = .588, p = .001$; *navigation complication*, $r = .537, p = .003$; *words and phrases*, $r = .511, p = .005$; *improvement needed*, $r = .395, p = .034$.

For the slightly less preferred interface – the Plain Language dialogue, there was also a marginally significant correlation between attitude mean scores and performance, $r = .315, p = .096$. Similarly, for individual attributes: *frustration with IA*, $r = .504, p = .005$; *stress*, $r = .437, p = .018$; *procedure*, $r = .421, p = .023$; *use again*, $r = .399, p = .032$.

5.13.3. Usability – Performance and Preference

The relationship between task performance and final preferences was also explored. Table 5.15 displays the correlations for the differences, then each interface individually.

	Differences (F-PL)	Formal (F)	Plain Language (PL)
r	.315	.379	.437
p	.097	.043	.018

Table 5.15. Correlations between Performance and Preference Scores

The trend in this data is to see increasing relationship between performance and preference, such that although in terms of the differences between preference ratings toward different interfaces (where differences were not obvious or highly influential in terms of usability) the association is positive, a medium strength but not significant at $p < .05$. For the formal dialogue design, the relationship between task performance and preference is significant, positive at a fair effect size. Finally for the Plain Language dialogue design, the relationship is strongly positive and significant.

In the eBanking tasks, performance did appear to be significantly associated with preference ratings.

5.13.4. Comparison of Metrics

The various usability measures were compared using normalised scales, see Figure 5.16 illustrating the similarity in responses for the two interfaces. The error bars indicate the 95% CI of the means. (For the task completion rates, these are the asymmetric Adjusted Wald CIs due to the high rates obtained).

The error bars illustrate the slightly wider range of performance, attitude and preference scores for the formal dialogue design compared to the Plain language. The means and CIs displayed by all the usability metrics show the overlaps, clearly illustrating the lack of perceived differences observed by participants in eBanking tasks with alternative dialogues.

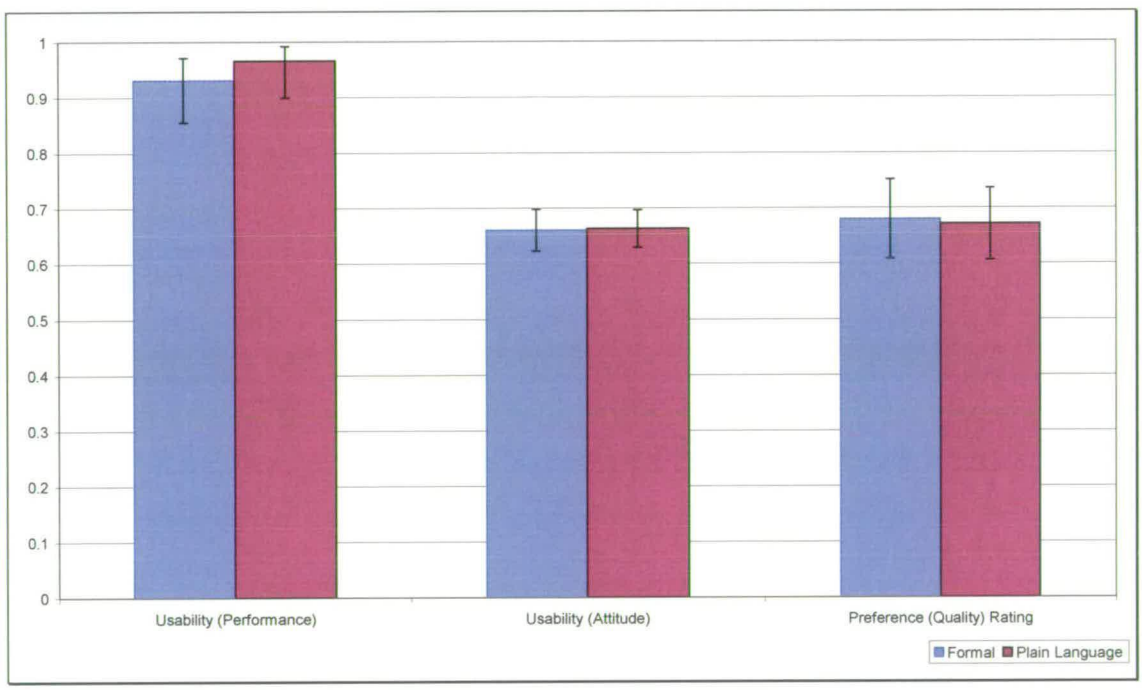


Figure 5.16. Comparison of Normalised Usability Metrics for the Alternative Dialogues

5.14. Experiment 2: Dialogue Style – Discussion

Different dialogue styles had no significant effects on users' perceptions of usability or quality (preference). It is difficult to make any comments about dialogue style changes because it was observed that many did not read the instructions, particularly on their second experience. Consequentially, some failed to notice the differences between the designs.

However, participant comments offered some ideas of how future experiments of this kind could be focused. For example, where tasks involve straightforward data entry, people tended to start data input without reading instructions. Intuitive design is essential and form-fill works well. The labelling of input fields may be more important than tonal changes or formality of instruction text. This appears to be due to participants' general lack of motivation or interest in reading on-screen text. This finding corresponds with other studies of reading on the Web (Morkes and Nielsen, 1998; Spool et al, 1997), suggesting that when users can simply scan the page and begin interacting, they will not be likely to read sentences fully. In fact, they tend to view any instructions as a potential barrier to interaction, and focus on areas on the page where they can begin entering their data.

The extension of eBanking into applications for loans or mortgages requires the presentation of important information to customers, who are unlikely to read details carefully online. The eBanking environment must pay close attention to these issues and design to comply with Financial Services Authority (FSA) regulations and procedures. Where tasks are much more complex, instructions may also be more critical to success. Such tasks may require further study of dialogue style and jargon in creating appropriate interfaces.

However, in the context of straightforward eBanking tasks consisting mainly of data entry, navigation to the form is the main challenge. Difficulties in navigation appeared to stem from confusion in selecting the appropriate terms, such as Payments & Transfers, Standing Orders and Direct Debits. This is a potential problem for self-service eBanking. Some of the jargon used in the site was broadly recognised and expected, but this was not true across the cohort. It became clear that users wanted consistency across channels, despite misunderstanding some jargon terms. Many suggested a glossary would be helpful.

Other evidence from participant comments suggests that language tone was not consistently interpreted in terms of the jargon vs. plain English alternatives proposed. However, participants did comment on wording changes when they appeared in pop-up dialogue boxes

requiring user action. This observation suggests some helpful guidance to the design of future dialogue style studies: certain visual presentations might focus more attention on dialogues. Generally, visual characteristics may have more effect than language tone in an online environment. Variables such as alternative layout, sequence and confirmation techniques could be investigated in the context of more information-rich and complex banking tasks and applications.

Finally, as in Experiment 1 (metaphors), attitude toward interface usability was highly positively correlated with preference ratings. This establishes a consistent relationship between usability perceptions and comparative preference ratings in two different sets of transaction tasks in an eBanking context.

5.15. Comparison of Participant Samples

As twenty-one of the usability attitude questions were identical across the two experiments, and both included use of the Form design (Formal dialogue style) interface, there was opportunity to compare a portion of the two samples who experienced the Form design in the first instance before comparing it to either the Spreadsheet metaphor or Plain Language dialogue designs. It is important to note however, that the tasks were not identical for the metaphor and dialogue style experiments. The usability attitude questions were taken after direct experience and before exposure to the competing interface. Preference ratings are not comparable in this way due to being taken after experience of two alternative interfaces.

The univariate analysis of variance (ANOVA) was used to compare the usability scores for the first instance of the Form/Formal design for each experiment. The data are shown in Table 5.16. and illustrated in Figure 5.17.

Experiment	Mean Usability	St. Dev.	N	Lower Bound CI	Upper Bound CI
Metaphor (1)	5.339	0.5924	16	5.0236	5.6550
Dialogue (2)	5.073	0.8630	15	4.5951	5.5509

Table 5.16. Mean Attitude Scores (21 matched questions) for the Form/Formal Design, Seen First, in Both Experiments

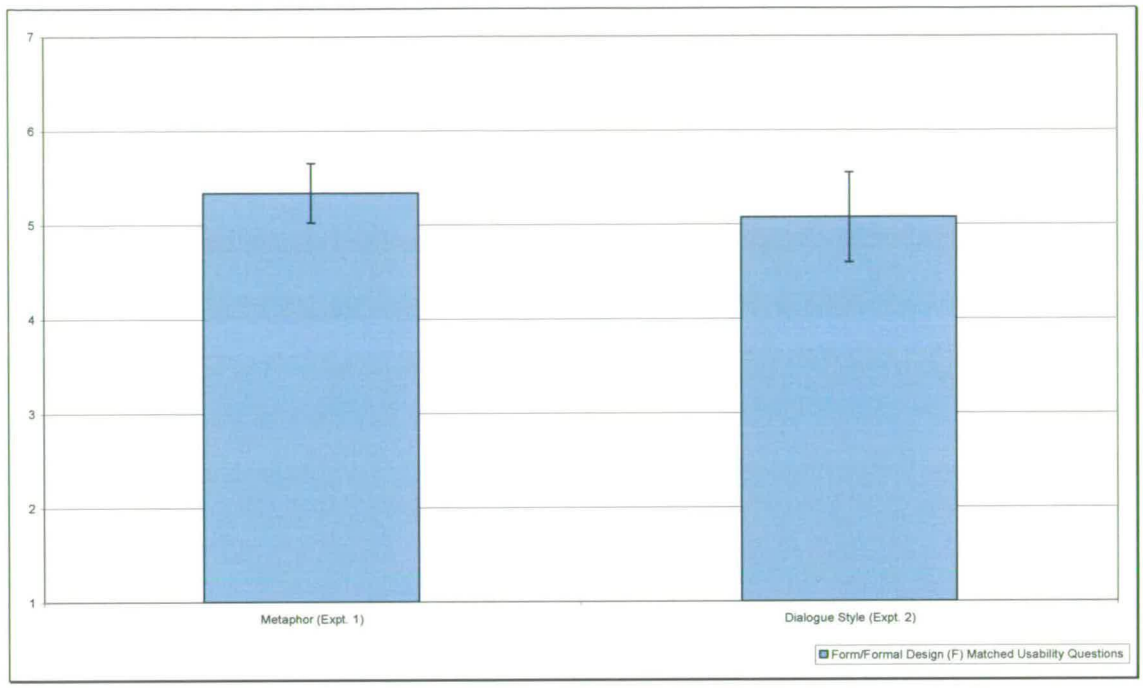


Figure 5.17. Mean Attitude Scores (21 matched questions) – Form/Formal Design

The data indicated that the two randomly selected samples did indeed respond similarly to the same interfaces (despite their experience being mediated through different tasks). There was no significant difference in the two usability scores: $F(1, 23) = 1.652$; $p = .212$; $N = 31$, although the range of scores captured in the second experiment was larger than the first.

Comparing scores for the Form/Formal interface for all three usability metrics (task performance, usability and preferences) for the two experiments for all participants and tasks, Figure 5.17 shows the consistent trend of lower scores for the dialogue experiment. Given that one of the main differences between the two experiments was the scope of tasks explored, it can be concluded that the creation of payment details online may be more demanding than changing and deleting payments on an interface. The effect of experiment in the results for task completion rates were marginally significantly higher for experiment 1 than experiment 2 (amendments/deletions vs. set up and pay). For the other comparisons, experiment number was not significant:

- Task Completion Rates, $F(1, 57) = 3.971$ $p = .051$; $N = 61$.
- Matched Attitude Questions (mean), $F(1, 57) = 1.826$; $p = .182$; $N = 61$.
- Mean Attitude Scores, $F(1, 57) = .209$; $p = .649$; $N = 61$.
- Preference Ratings, $F(1, 57) = 1.826$; $p = .182$; $N = 61$.

Cronbach's alpha (p.37) computed for the 31 participants who had first exposure to the Form/Formal interface in both experiments indicated high inter-item reliability, $\alpha = .905$ (N= 21 matched questions). Items which increase alphas when deleted were: text size (.911) and would use again (.906). Only text size appears to be a genuine candidate for deletion.

When computed for all 61 participants matched usability responses toward the form/formal interface experienced, $\alpha = .918$, again indicating high inter-item reliability.

Items which when deleted increase the reliability of the scale included: text size (.925), appearance (.921) and reliability (.919). Indicating that text size might be unrelated to the other attributes. However, none of the alpha scores improve greatly – and the original alpha is very high, therefore there is satisfactory evidence that each individual statement in the usability questionnaire is making a positive contribution to the overall usability attitude scale.

Given the consistent findings on the attribute text size, there may be reason to consider it for removal from the usability inventory. However, as the actual text size was not a variant of the designs tested in the experiments, it was considered an important attribute to collect and check that the default text size experienced with a design was widely appropriate to the target audience for their tasks. In this respect, the scores for this attribute were positive for all designs in both experiments, however the attitudes were also higher for the simpler tasks of amendment and deletion (experiment 1) than for set up and payment (experiment 2).

5.16. Summary of Hypotheses and Evidence

Hypothesis and results for Experiment 1 – User Interface Metaphors

Hypothesis H_0 E1a: The different interface metaphors will not result in different usability attitude and performance scores.

Partially Rejected: The different interface metaphors resulted in significantly different user attitudes toward usability. However, they did not result in significantly different performance scores.

Hypothesis H_0 E1b: The different interface metaphors will not result in different user perceptions of preference.

Rejected: The different interface metaphors resulted in significantly different user perceptions of preferences or overall quality in designs.

Hypothesis H_0 E1c: There will be no relationship between measures of usability performance, attitude or preference.

Partially Rejected: There was a *consistently significant, positive and strong* correlation between measures of usability attitudes and preference.

There was a consistent positive association between measures of usability attitudes and performance, but not always strong or significant.

There was a consistent positive association between performance and preference, but not always strong or significant.

Hypothesis and results for Experiment 2 – Interface Dialogue Style

Hypothesis H_0 E2a: The different dialogue styles will not result in different usability attitude and performance scores.

Not Rejected: The different dialogue styles did not result in significant different user attitudes toward usability or performance scores.

Hypothesis H_0 E2b: The different dialogue styles will not result in different user perceptions of preference.

Not Rejected: The different dialogue styles did not result in significant different user perceptions in terms of preferences.

Hypothesis H_0E2c: There will be no relationship between measures of usability performance, attitude or preference.

Partially Rejected: There was a *consistently significant, positive and strong* correlation between measures of usability attitude and preference.

There was an inconsistent association between measures of usability attitude and performance which were either not significant or only marginally so.

For both individual designs there was a significant, positive relationship between performance and preference.

5.17. Discussion of eBanking Interface Usability

A wide range of users react similarly to the form metaphor in eBanking payment transactions. There were no general gender or age differences. This probably explains the ubiquity of forms on the Web. Further, this experiment has empirically validated their use in data entry tasks online in an eBanking context.

Despite heuristics suggesting that plain, simple language would be more usable, dialogue-styles were barely noticed in the highly effective linear form fill process. Users stated a desire for consistency across banking channels. In eBanking services, this suggests that consistency overrides language considerations when employing the form metaphor.

Web users dislike reading on-screen and more dramatic changes in visual presentations may have more impact (positive or negative). The usability engineering methodology including participant commentary, observation helped gather this result. Detailed usability questionnaires combine with such qualitative data and help formulate redesigns in an iterative development process.

The linear rating of preference allowed fine detail of strength and direction of user preferences to be analysed. This was also used to infer the overall quality of the interface as subjectively perceived. In addition, the relationship between usability attitudes and preferences could be examined. The relationship between perceived usability and preference rating was positive and significant for both experiments over a range of eBanking tasks.

In the metaphor experiment (1), correlations show a high relationship between usability attitudes and preference (quality) scores, $p = .001$. The attribute *use again* also highly correlates with preference, for the Form this correlation was $r = .525, p = .002$ and for the Spreadsheet, $r = .364, p = .040$ (Table 5.8, p.152); and with mean usability questionnaire, for the Form this correlation was $r = .676, p < .001$; and for the spreadsheet, $r = .804, p < .001$ (see Appendix E, p.306). Both designs were scored positively for *use again*, the Form scored 5.91 out of 7 on this attribute, whilst the Spreadsheet scored 5.63 (Figure 5.8, p.144). Finally, intention to use (measured in the interview) was high for the service; there were also high perceptions of service utility. There were also qualitative indications that intention was related to the users preferred interface more often than the functionality itself. In this experiment, people also exhibited clear preferences for one or other design.

In the dialogue experiment (2), mean attitude differences between the two designs were highly correlated with preference score differences, $p = .001$ (p.167), matching the results of the first experiment. Similarly, the score on the attribute *use again* highly correlated with the preference quality rating for the Formal interface $r = .459, p = .012$ (p.170). However, this correlation was only marginal for the Plain Language version ($r = .319, p = .092$, see p.170). Attitude towards whether the interface would be used again correlated highly (positively) with the overall mean usability score for both interfaces. For the formal style this correlation was $r = .483, p = .008$; and for the plain language version, $r = .461, p = .012$ (see Appendix F, p.316). Both dialogues also scored very highly in terms of the attitude towards reuse, with 6.34 for the formal style and 6.24 for the plain language version (Figure 5.13, p162). These scores are very high and positive towards reuse. A possible explanation for the increases was that participants actually set up and paid a new recipient with the service, possibly illustrating better utility in the functionality provided. This corresponded to a behavioural intent registering at 100%. Given that this statistic is obviously subject to a ceiling effect that cannot be attributed to a real population, confidence intervals in this data were computed to further establish probable boundaries for the intent score. Similarly to methods appraised in terms of task completion scores with small samples (Lewis and Sauro, 2006), the Laplace method of estimation was computed ((x + 1)/(N + 2)), giving an estimated usage intention score of 96.77%. Compare this to a similar estimate of 79.41% for the first experiment. Over both experiments, the usage intention is estimated at 88.89%.

Intention to use is high for the service in terms of its utility – with an indication that this related to the users preferred interface more often than the functionality itself. Intentions were higher in the second experiment perhaps due to tasks exposing participants to a higher degree of utility in the interface. This has parallels to the idea that trialability can drive adoption from innovation diffusion theory (Rogers, 1993).

Generally, usability seems to be highly positively related to preferences and subjective comparisons in terms of overall quality.

In these two experiments, the usage intentions (gathered qualitatively) were very high and positive towards using such designs in real life. This corresponds with reasonably high usability attitude, performance and preference (or overall quality) ratings. This suggests that certain levels of quality, subjective and objective usability must be reached to drive usage intention. Usage intentions rose considerably compared to the intention to use financial portals. The levels of performance were higher than in the Web Portal experiments, raised

from around 70% effective to 90+%. Corresponding quality ratings were around 65% for eBanking compared to below 50% for the Web Portal designs. Similarly usability attitudes were also higher towards the eBanking interfaces, although generally more varied.

This data highlights two possible avenues for further research. Firstly, whether usability, preference and usage intention for a specific interface design would generally correlate and show a high relationship. This would offer evidence to support usability engineering assumptions that high usability can predict end product success and select the most appropriate early design to develop. In order to examine this relationship more detailed measure of usage intention will need to be developed, specifically related to individual designs themselves. The attribute *use again*, the preference rating and overall usability attitude scores are all potential guides to user intentions.

5.17.1. Limitations

The main limitation of these studies was the relatively small samples of users who participated. This was partly to do with cost and time constraints, but also the requirement of participants being customers of the Bank and potential eBankers. Using these specific recruitment criteria focuses the evaluations, but it does result in a smaller pool of potential participants. Yet this focus avoids extraneous issues of brand and reputation from participants not already associated with the Case Bank. In some Web usability work a much more diverse set of participants would be ideally required, but in the eBanking domain it is more appropriate to focus on customers and Web users. This is because, in practice, switching banks is still rare (BBC News, 2004) and unlikely to be driven by access to the eChannel alone.

Another limitation was that participants were restricted to a short exposure to each different interface. In this way the experiment was able to focus specifically on the areas of the interfaces where differences between them were most apparent. The disadvantage was that participants only experienced three or four tasks using each interface. They were not given any free exploration. Again this was due to control reasons in the experimental comparison, but particularly the Spreadsheet design may have suffered if it had a steeper learning curve than the Form design.

In testing new user interfaces, practice effects can be important. Testing for practice effects in the laboratory is difficult. By incorporating groups of novice and experienced users into

experiments it may be possible to examine effects of practice and learning, although this was beyond the scope of these experiments which assume sporadic use of the banking service. This assumption should hold in terms of performing transactions, for example, people don't usually need to pay bills every day. It may be that business use of eBanking would differ in this regard and this is a possible avenue for research.

A limitation of the interview intention question was the inability to distinguish whether one or other design was influencing this choice more. It became clear from some comments that many participants answered this question from the point of view of their preferred choice, indicating scope for an intention metric that is interface-specific, similar to the usability questionnaire attribute *use again*.

The numbers of participants involved in the study also restricted the analysis of the questionnaire structure to a preliminary examination of the groupings. Again this was performed using semantic considerations, usability and preference results and by examining the correlation matrices (Appendix E, p.307 & F, p.317) and reliabilities for proposed groups.

5.17.2. Guidelines for eBanking Interfaces

In the design of an eBanking payment interface, the linear sequential form metaphor provided a very usable interaction platform for set up and making payments. It is unlikely that users would read through any onscreen instructions, although what use they would make of these in error is unclear from the scope of these experiments. Using appropriate input field keywords would provide general cues; in addition, users can gather signals simply from the size and shape of dialogue boxes.

To facilitate intuitive interaction with form filling, it would be advisable to keep the number of inputs in a form to a minimum. Using specific forms for each payment offers a way to reduce form elements and increase the clarity of the information required for each transaction.

Employing consistent language between service channels is appropriate in the eBanking environment. Banking jargon is not always understood, but it is recognised and generally familiar. Explanations of transaction types, brief and available on request, e.g. a glossary, in an accessible format, would be of benefit to many users. Additionally, offering cross-linking

between transaction types enhances navigation and exploration of the various transaction types.

The design offering an overview table of arrangements, offering specific forms for each transaction via a hyperlink is a metaphor that may benefit from further extension. For example, the overview table could be extended to include all transaction arrangements, essentially allowing users to ignore the categorisation of transaction types and instead focus on recipient names in navigation. Payment set up might be offered using simple questions to guide users to the right transaction type. For the minority who are competent with banking terms, the transaction keywords and menu navigation should be retained. The menu could offer function-centric (rather than specific to one account) access to transactions, with the 'from' field also being dynamic between a customers' accounts. Such operation would speed up navigation and the performance of multiple, routine transactions for users who frequently access the service.

This empirical evidence and the corresponding observations provide information for researchers and designers in creating guidelines and heuristics for eBanking service design. Transactions (payments and transfers) are the first, most basic foundation of any eBanking service (Jayawardhena & Foley, 2000). eBanking is generally perceived to be difficult to use (Hudson, 2002), by creating additional functionality without addressing usability and human factors issues, this perspective is not likely to be improved. Instead, providing easy and intuitive basic functionality for a wide range of potential users, in a secure environment (Mattilla et al, 2003) creates a solid service which can be built upon.

In terms of transactional interface design, use consistent language with other Banking channels. Simple, linear stepwise processes were suitable for a wide range of potential users in a Web context.

5.17.3. Outcomes and Further Work

In the experiments conducted thus far, IA & navigation consistently have high associations with preference. Further, content attributes were relatively more important in eBanking than on Web in general information searching. Additionally, affect attributes were important in driving people to use a new service online.

The structure of the usability attributes covered in the questionnaire is proposed as shown in Figure 5.18.

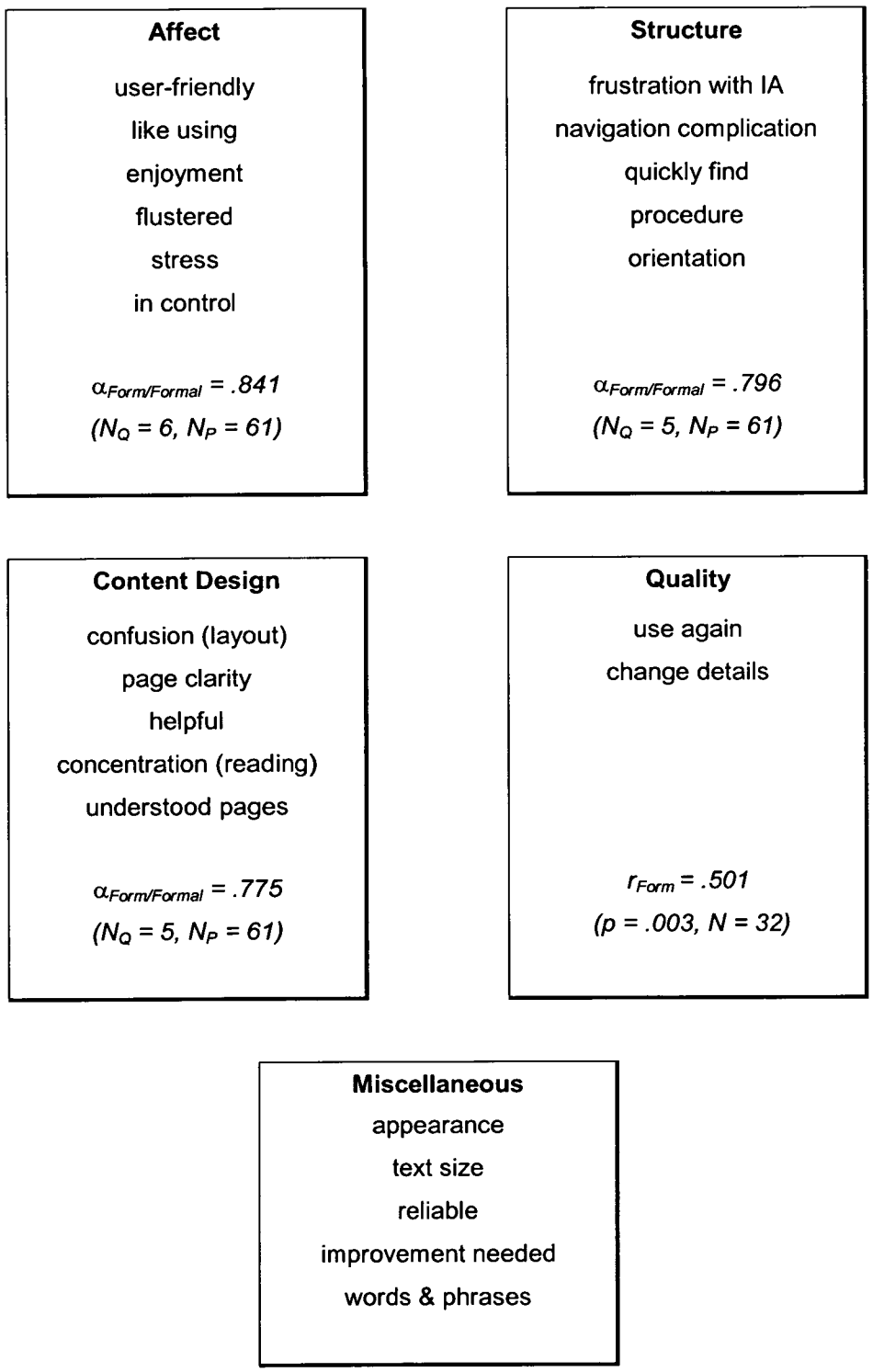


Figure 5.18. The Elements of eBanking Usability – Subjective Evaluation Components

These eBanking usability attributes relate to transaction set up, payment, alteration and deletion within an authenticated, personal eBanking service. The reliability scores relate to the full N=61 cohort using the Form/Formal design in both experiments. The correlation scores for the attributes in the quality group only refer to the first experiment where the task-specific question (ease of changing details) was used.

Attitudes, performance and preference were all above the midpoint of their scales, considered therefore generally positive. Similarly, usage intentions toward performing payments on the eBanking services were also highly positive. This brings up the question of how well comparative preferences and overall quality can be considered related to usage intentions. Using the midpoint as a guide to whether quality was generally good or bad provides an arbitrary divide, but based on the experiments conducted thus far it can't be concluded that preference is akin to usage intent. So in order to explore the usability-intention relationship further, the preference metric needs to be extended to measure the adoption construct of intention for each interface option. The metric would be attributed to each interface, rated after direct exposure in line with the usability questionnaire.

By undertaking a further experiment using such a modified preference metric, the usability and salient interface characteristics which explain variances in usage intention can be identified. The relationships can then be compared to previous results using the comparative preferences. Thus the use of preferences as proxy for intention to use can be discussed. An appropriate platform for a usability-intent experiment must be selected.

As well as offering the ability to perform financial transactions, eBanking also contains highly personal, account information. An online banking statement is typically presented as a vast, linear, date ordered collection of banking information. By combining the original information-retrieval theme from the Web Usability Questionnaire, and including the eBanking-specific attributes constructed for the metaphor and dialogue experiments, a further questionnaire will be constructed which is focused on finding and using statement data in eBanking. The next experiment will study various designs for interfaces to eStatement information. One of the key aims will be to find a design which will prove better than traditional paper statements and therefore drive adoption of the eStatement service. The adoption question becomes whether bank customers who are Internet users will be prepared to make the change from paper-based statements to eBanking with any of the proposed statement designs. For this reason, different levels of utility in designs were also examined in relation to intentions and usability.

Chapter 6. Statement Search Design for eBanking User-Interfaces

This chapter begins by introducing electronic statements (eStatements) and the heuristics and principles in their design. It then considers experimentally how the usability engineering methodology can be adapted and used to explore how the design of a service affects potential uptake, relating to the proposed usability-usage model described in Chapter 2. Potential rewards for the service provider in choosing the most appropriate design could assist in managing the development budget. Usability and usage intention are investigated in designing eStatement facilities to enhance eBanking.

6.1. Introduction

6.1.1. eStatements

eStatements are electronic versions of transaction records traditionally sent as paper statements. These can be displayed literally replicating each individual paper page as separate screens. In eBanking, users can be offered the ability to access their data online in a more efficient or personalised manner. Using searches, eBanking users can display lists of relevant results, and jump directly to the relevant transaction in any particular statement. This follows the same principle as general searches on the Web (Nielsen, 1993a). Search metaphors could be considered as simple (e.g. single field) or advanced (e.g. multiple fields). The advanced search metaphor may be viable in eBanking due to the vast amount of detailed transaction data held. The data is also highly categorised (e.g. by transaction type, amount etc.)

Statement searching could provide a better alternative to the literal statement metaphor of paper sheets. Considerations, from a service providers' point of view, dictate that any user-interface or functionality development is balanced by cost-savings. In the case of eStatements, this depends upon a sufficient proportion of users opting for electronic over paper statements. Therefore identifying what degree of search functionality will be required to drive this adoption is of interest and this research explores these factors.

Usability engineering methods have been successful in comparing interface design options (i.e. working prototypes) and selecting the most promising to develop and roll-out (Weir et al, 2006; Shneiderman and Plaisant, 2005). However, usability engineering does not usually attempt to predict usage of an application in the real world. To better address the case for the service provider, usability measures will have to be combined with other data to investigate whether functionality and usability influence switching behaviour.

Forrester Research (Graeber et al, 2003) reported on the early adopters of eStatement facilities in Canada. They found that despite no incentives to switch, older, wealthier users, online bankers and online bill payers were interested in adopting these services. They also reported wide interest from younger consumers. Their research clearly indicates that there is a potential market for eStatements, but that age factors may be a factor in adoption intentions.

Paper statements are a costly matter for banks. On a monthly basis, printing, postage and valuable man-hours are spent supplying customers with their paper statements. People often

throw away or lose these statements, and in a worse case scenario are reprinted, usually at the customers' expense. All this is wasteful for the bank, the customer and the environment.

In the literature review concerning adoption, acceptance or usage of Information Technology, there is much to suggest that utility and usability are important drivers of adoption. Usability as measured by perception and performance measures in experimental settings are expected to be important factors in predicting intention to use a system – in this case a switch from paper to online statement delivery. This experiment aimed to select the most appropriate level of functionality in design for eStatements services. It also aimed to test the validity of the assumption that usability (as measured by a range of utility, performance and perception metrics) is related to successful design in terms of real-world use, or usage intent.

6.2. Interface Designs Considered in the Research

The simplest implementation of eStatements was the control. Two levels of search functionality were investigated in comparison to the basic data presentation design. The use of metaphor has already been exposed as a valid technique to assist in learning to use a system. Appropriate selection and design of the metaphor and resulting user interface will help the user to transfer knowledge from one situation to the next. Thus it was important to design a search engine and results pages which were consistent with expectations (Baeker et al, 1995).

The design of search facilities on the Internet are typically keyword driven. However, database search engines adopt a raft of searchable options. There are also resource considerations in terms of the number of hits to the database and the time taken to extract the data required to populate search results. These concepts were beyond the scope of the experiment but they were influential in the choice of two levels of search - a constrained set of search options from which only one strategy could be picked compared to an extensive advanced search query allowing multiple search criteria to be specified at once.

Each eStatement design variant was implemented based on a fully functional mirror of the eBanking service as was being used at the time of the experiment. The pages appeared as functionally faithful to the real-world site as possible. The search engine interfaces were both designed to adhere to general principles of usability for form-fill interaction: to display meaningful field titles, logical groupings, sequences, and familiar labelling (Shneiderman,

1995). The search designs were expected to allow effective keyboard data entry and good user satisfaction.

Variant	Search Criteria	Search Type
Data only (DO)	None Linear paging only	N/A
Simple search (SS)	Date - single Amount - exact Statement - sheet number	'OR'
Advanced search (AS)	Date - range Amount - exact or range Transaction type (from list) Transaction description	'AND'

Table 6.1. Interface Design Variants Considered in Experiment 3.

The three different eStatement designs are described in Table 6.1 and illustrated in Figures 6.1, 6.2 and 6.3 respectively. In the participant sessions, the interfaces were referred to by unrelated letters (i.e. A, B and C) to avoid biasing participants with potentially leading expressions (such as basic or advanced search).

6.2.1. Data Only

The Data Only design (Figure 6.1) features a user interface for the account statement function which is essentially identical to that which was currently being seen by the bank's eBanking users. The data is displayed in a coloured table.

The key difference was that the amount of account transaction historical data available per account was increased from 2 months in the live service to 12 months in the mirror site for the experiment. The metaphor was that of the traditional paper sheets. A user can navigate linearly between these pages via 'Previous Page' and 'Next Page' buttons, which are paginated to match exactly with the corresponding paper statement sheets; and had then to be scanned line by line.

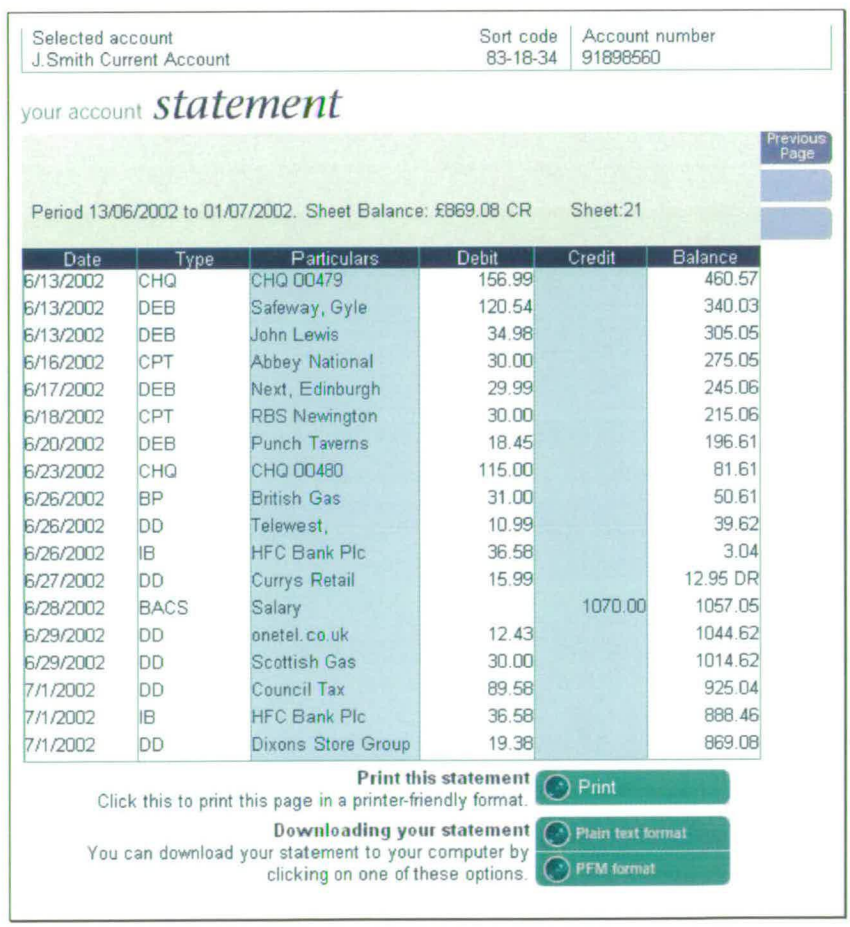


Figure 6.1. Data Only (DO)

6.2.2. Simple Search

The Simple Search design (Figure 6.2) extends the Data Only design with a set of search definition fields whereby the user can select one of three search criteria - choosing to search for either a date (on or prior to a single specified date), an exact amount or a statement sheet number. The user then clicks the 'Find' button on this search page and the search results appear. The search fields remain at the head of the results page. The user can click a pertinent search result to jump to the full relevant statement sheet containing that item – thus borrowing from the metaphor of standard search engine behaviour for displaying results. The search is accessed from the main statement page as shown in Figure 6.4.

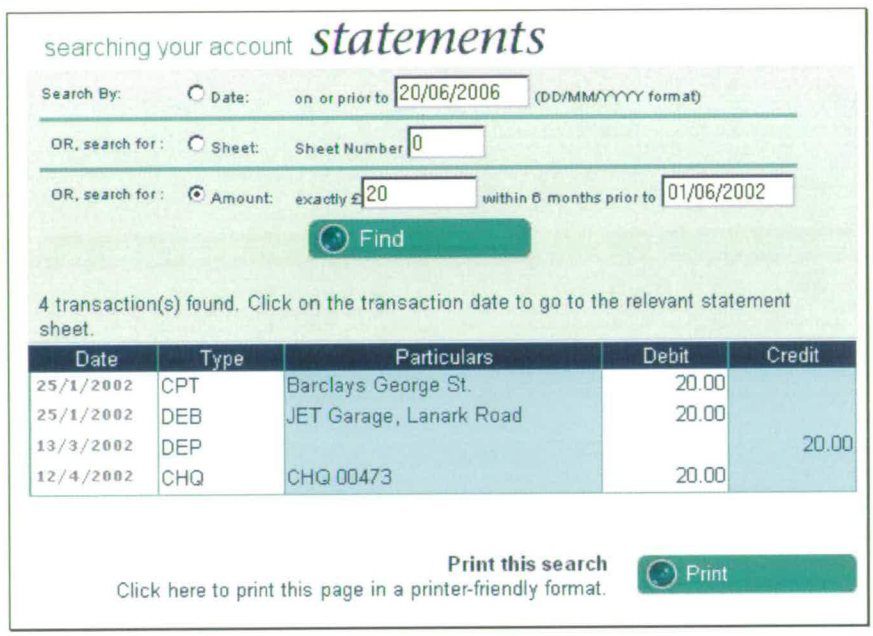


Figure 6.2. Simple Search (SS)

6.2.3. Advanced Search

In the Advanced Search design (Figure 6.3), the search facility allows the user to select any combination of search criteria from four options - searching for a date range, an exact amount or range of amounts, by transaction type (i.e. debit card or standing order) and by a text search on transaction description, the details column of the statement page (displayed as 'Particulars' both as column heading and search field label). Again the user can click a pertinent search result to launch the relevant full statement sheet containing that item. Also, when the user clicks the 'Find' button on the search page, the search results are split into sub-pages on the search results page (ten transactions per group), where the familiar search engine metaphor of the 'Previous', 'Next' and 'Most recent' buttons are used to help the user navigate through the search results.

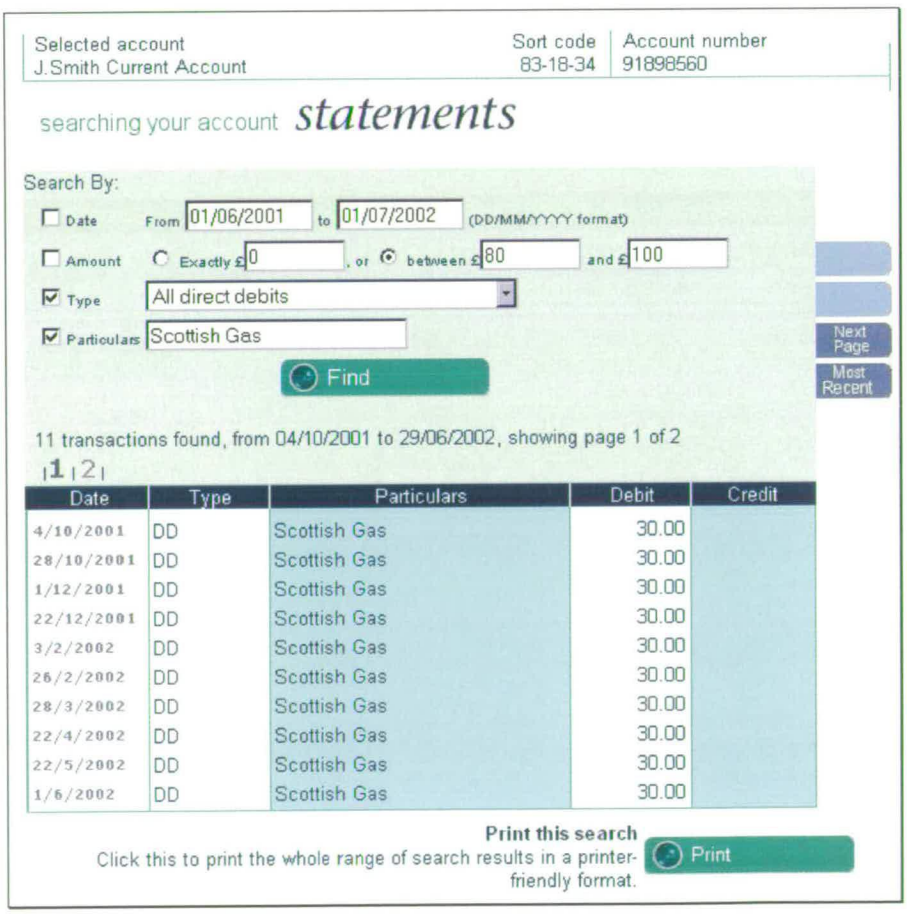
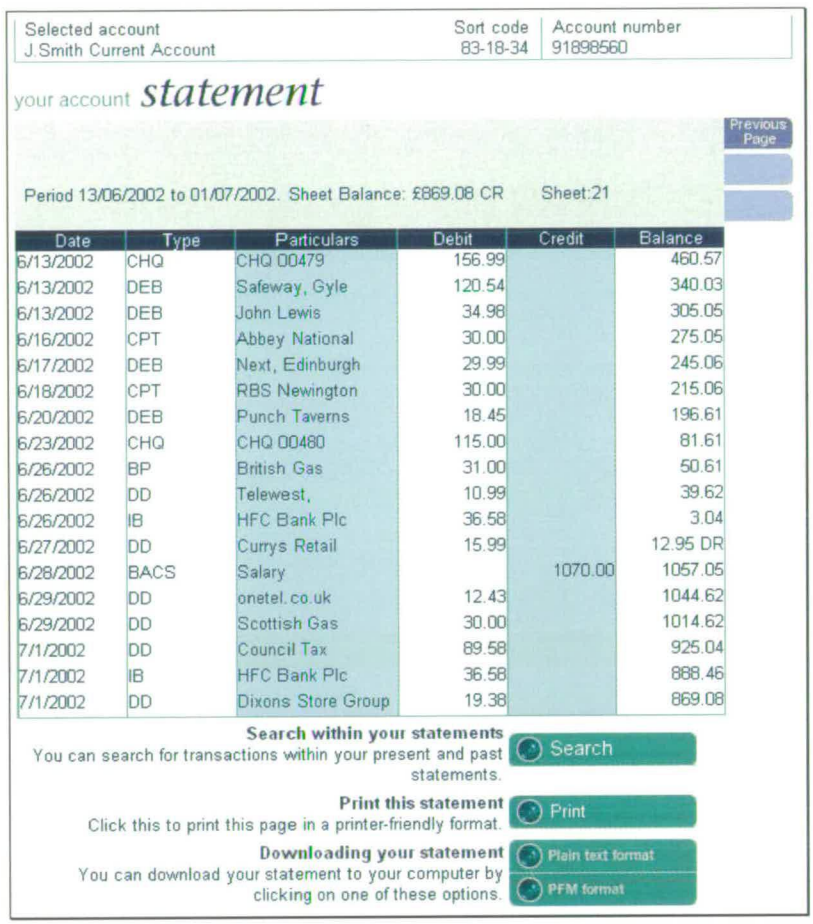


Figure 6.3. Advanced Search (AS)

In the case of two search interface designs (Simple and Advanced), the search is accessed via a button below the data on the main statement page as shown in Figure 6.4.



Selected account: J.Smith Current Account | Sort code: 83-18-34 | Account number: 91898560

your account *statement*

Previous Page

Period 13/06/2002 to 01/07/2002. Sheet Balance: £869.08 CR Sheet:21

Date	Type	Particulars	Debit	Credit	Balance
6/13/2002	CHQ	CHQ 00479	156.99		460.57
6/13/2002	DEB	Safeway, Gyle	120.54		340.03
6/13/2002	DEB	John Lewis	34.98		305.05
6/16/2002	CPT	Abbey National	30.00		275.05
6/17/2002	DEB	Next, Edinburgh	29.99		245.06
6/18/2002	CPT	RBS Newington	30.00		215.06
6/20/2002	DEB	Punch Taverns	18.45		196.61
6/23/2002	CHQ	CHQ 00480	115.00		81.61
6/26/2002	BP	British Gas	31.00		50.61
6/26/2002	DD	Telewest,	10.99		39.62
6/26/2002	IB	HFC Bank Plc	36.58		3.04
6/27/2002	DD	Currys Retail	15.99		12.95 DR
6/28/2002	BACS	Salary		1070.00	1057.05
6/29/2002	DD	onetel.co.uk	12.43		1044.62
6/29/2002	DD	Scottish Gas	30.00		1014.62
7/1/2002	DD	Council Tax	89.58		925.04
7/1/2002	IB	HFC Bank Plc	36.58		888.46
7/1/2002	DD	Dixons Store Group	19.38		869.08

Search within your statements
You can search for transactions within your present and past statements. Search

Print this statement
Click this to print this page in a printer-friendly format. Print

Downloading your statement
You can download your statement to your computer by clicking on one of these options. Plain text format | PFM format

Figure 6.4. Location of the Search Button

6.3. Research Questions and Hypotheses

This research aimed to determine whether alternative functionality (utility) in eStatement services will result in significantly different attitudes toward usability of the interface. In addition, it was of interest in this research to investigate whether utility and usability in design would influence user intentions to switch from paper statements if offered access to an eStatements service.

6.3.1. Hypotheses

Experiment 3: Level of Utility in eStatement Interface Designs

The null hypotheses tested were:

Hypothesis H_0 E3a: The different eStatement interfaces will not result in different usability attitude and performance scores.

Hypothesis H_0 E3b: The different eStatement interfaces will not result in different usage intentions.

Hypothesis H_0 E3c: There will be no relationship between the usability measures of performance, attitude or usage intention.

It is predicted that a measurable usability difference will be achieved through the different functionality available in the eStatements interface. Similarly, it is predicted that functionality will influence users' intentions to switch to online statement delivery. Participant variables such as age and gender were balanced in the experimental design along with the order of experience for the three interfaces. Each interface was substantially different from the next and altered in one design feature only (as per Figures 6.1 – 6.3). Data were collected to test the null hypotheses and indicate whether they were accepted or rejected at the .05 level in a two-tailed test, and if rejected the direction of the difference is indicated.

6.3.2. Participants

The participants in the experiment were recruited as being representative of current and potential members of the target market for the sponsoring Bank's eBanking service. They were all 'Internet-savvy' (must have used the Internet in the last month) customers of the Bank. They had wide ranges of experience with the Internet and eBanking services. A representative sample of 182 participants took part in the research which was conducted at two locations in the UK - Edinburgh and Cheltenham. A three-level split on age groups was planned for the experiment, investigating younger, middle and older age groups separately.

The participants each spent approximately 75 minutes answering demographic questionnaires, using all three eStatement design variants (randomised order), responding to usability questionnaires, assessing their stance on paper statements and finally completing an exit interview. Volunteer participants were contacted by letter and telephoned to be booked into a timetabled slot.

6.3.3. Tasks

The tasks involved looking up items in a statement, e.g. finding out what date a cheque was paid out of the account. Tasks were chosen carefully to ensure that they could be successfully completed with each of the eStatement design variants and that they represented common banking tasks. The tasks focused on the statement pages exclusively to gather usability data specific to the concept of eStatements rather than eBanking in general. This also allowed specific items about the statement process to be included in the usability questionnaire.

To compare strategies for completing the various types of task, five types of task were posed for each interface. The details were varied slightly to engage the user with the three different versions of the service. Each type of task could be completed using an optimum path (specific to each design variant), but participants were free to pursue different filtering methods using the search criteria available.

The full task lists are included in Appendix G, their types consist:

- Type 1: Finding Direct Debits to a named company in two different (specified) months
- Type 2: Finding a Debit Card transaction to a named company for a certain amount (specified)
- Type 3: Printing a missing Statement sheet (with a specified number and associated date range)
- Type 4: Locating a Cheque with a specified number and approximate amount
- Type 5: Printing a record of Direct Debits to a named company from three specified (sequential) months

6.4. Dependent Variables

6.4.1. Usability - Performance

The usability of each interface was firstly measured using objective measures of task performance. In this experiment these included task completion, use of various search criteria and print options. In addition, some qualitative information about task completion and any noticed errors and their correction were recorded. Where task completion rates are poor, or error rates are high, the interface effectiveness is called into question.

6.4.2. Usability - Attitude

The task of performing transaction searches can be thought of as a specialised form of information-retrieval, as was studied in the pilot experiments. Further however, eStatement tasks are within the eBanking specific context of private account details. The tasks considered in the evaluation suited some measures from the eBanking and Web Usability Questionnaires used in the previous experiments.

In conducting this experiment, the relationships between functionality, usability and usage intention were also of interest. Referring back to key adoption principles, the concepts of functionality, utility and relative advantage were introduced alongside the previous usability questions.

Using the attitude statements from the Web Portal experiments and the eBanking transactions experiments, a summary table was drawn up and noting which attributes consistently appear either as aspects which score high attitude scores, are significant differentiators between designs, and/or have correlation coefficients of $r > .5$ in relation to preference scores (relative differences and/or the individual interface). For the first pilot experiment, this involved collecting positive-scoring attributes for the winning design (A), noting those which were significantly different between the two portals, and noting high (>.5) relationships with preferences for the relative differences, and for Design A in isolation. Where no difference was found, such as in the dialogue experiment (2) both sets of absolute correlations, and usability attribute scores were collected. The resulting table is shown in Appendix G.

The attributes of trust, clutter and expectations – removed from the questionnaire for the eBanking transaction evaluations (Experiments 1 & 2) were replaced.

Trust

The issue of trust although rated highly for Web Portal designs, was not a significant differentiator between alternative portal designs, neither was it related to preferences. However, when it comes to asking people about switching preferences, a degree of trust in the service may well be an important factor. Research does suggest that trust issues can be a potential barrier to adoption (Gefen et al., 2003). Therefore the issue was included in this experiment and the relationship between usage intention and trust perception could thus be investigated.

Clutter

Clutter had played an important role both in terms of being a highly-scoring attribute of the preferred Web portal designs in the pilot test and in differentiating between portals and correlating highly with preference ratings. Clutter was originally introduced as a factor which may affect information retrieval and interaction due to an overload of screen elements being considered unnecessarily complex and reducing the visual salience of important content, navigation and orientation cues. Although the attribute was dropped for the fairly clutter-free designs evaluated for transactions in eBanking, the statement pages contain dense information, and the tasks are retrieval in nature. Therefore, the attribute was again considered important in this context.

Matching Expectations

Again, this was an attribute removed from the eBanking transaction questionnaire due to the desire to avoid comparisons with a counter-service expectation. Paper statements are a self-service activity, mainly to review and store. Transaction searches however, are generally performed by counter or telephone (staff-based) service. However, with statements being a well-known format, and information search in general being a very well-known and used Web service, the issue of expectations when it came to eStatements was of interest again. Therefore the attribute statement relating to expectations was reinstated in the questionnaire list.

Removed Items

'Words and phrases' and 'Change details', the experiment-specific items from the previous eBanking questionnaire were not included as they were not relevant to the eStatement search designs and variants, or the tasks included in this evaluation.

Additional Items

Some additional items were included in the questionnaire to begin to interpret characteristics relating to the potential usage or switching behaviour. Usefulness or relative advantage:

- REPLACE: The online Statement Service is a poor replacement for paper statements
- CONVENIENT: Using the online Statement Service was more convenient than paper statements

Utility and integrity concerns about the match of the service to statement task requirements:

- SUITABLE: The format of the statement printed out from the online Statement Service was suitable for my needs
- AUTHENTIC: The printout from the online Statement Service did not look authentic

The resulting attributes which were retained for the eStatements experiment are shown in Figure 6.5.

AFFECT

1. I liked using the online Statement Service
2. I felt in control when using the online Statement Service
3. I felt flustered when using the online Statement Service
4. I found the online Statement Service user-friendly
5. I did not enjoy using the online Statement Service
6. I felt under stress when using the online Statement Service

STRUCTURE

7. Moving through the pages of the online Statement Service was too complicated
8. I always knew where I was on the online Statement Service
9. I found the organisation of the online Statement Service very frustrating
10. I could find what I wanted quickly with the online Statement Service
11. When using the online Statement Service I didn't always know what to do next
12. The options available on the online Statement Service matched my expectations

CONTENT DESIGN

13. I felt that the online Statement Service was helpful
14. I found the layout of the pages on the online Statement Service very clear
15. The pages on the online Statement Service were easy to understand
16. The layout of the online Statement Service was confusing
17. Using the online Statement Service took a lot of concentration
18. The pages on the online Statement Service were very cluttered

QUALITY

19. I would be happy to use the online Statement Service again
20. I feel that the online Statement Service needs a lot of improvement
21. I felt the online Statement Service was reliable
22. I don't trust the information from the online Statement Service
23. I liked the appearance of the online Statement Service
24. Some of the text used by the online Statement Service was too small

UTILITY

25. The online Statement Service is a poor replacement for paper statements
26. Using the online Statement Service was more convenient than paper statements
27. The format of the statement printed out from the online Statement Service was suitable for my needs
28. The printout from the online Statement Service did not look authentic

Figure 6.5. Web Usability Attributes used in Experiment 3 – Grouped by Semantics and Previous Structural Statistics.

6.4.3. Usability - Other Measurements

As for previous experiments, participants were not timed for efficiency during their first use of the eStatements interfaces. Instead the 'think aloud' protocol was employed to gather qualitative comments and increase understanding of statement tasks and search strategies.

Final qualitative data were also collected in a debriefing interview. Participants were asked about what they liked, disliked and what suggestions for improvements they could offer about various aspects of statement design, search criteria, printouts and presentations. They were also asked to rank the interfaces in order of preference. This simple ranking was collected because in this experiment design the rating question measured the potential to switch to eStatements, described next. The interview questions are available in Appendix G.

6.4.4. Usage Intention Measurement (or BI)

An important feature of this experiment was the measurement of participant's intention to suspend (switch off) paper statements. Many adoption measures rely on actual or self-reported usage data, but such information cannot be provided for a system in the initial prototype phase. Therefore participants were asked about their intentions to use eStatements. A linear response scale of 0 to 12 was used for this measure (similar to the 0-30 scale described in Figure 3.3, p.56). It offered participants a forced choice between definitely wanting to keep paper statements and wanting to stop paper statements, indicating the relative strength of their opinion. This question was incorporated into the experimental measures set to model eBanker switching intentions from paper to online. Note that participants were asked to choose between the two delivery methods as a forced choice, this provided clarity for the switching intention measure in terms of the question being asked in an experimental setting. The resulting intention metric is limited, in that it assumes a trade from paper to online statement usage, something which might not be required or occur in practice.

The relationship between usability and intent to use a service is of primary interest, therefore a scalar measure of usage intention was desired to use in correlation analysis. The method previously used to gather preference information offered a potential hands-on measure. This method was selected due to its success in previous studies as an attitude gathering tool.

The usage intention measure designed was a self-report attitude rating. The measurement was first taken at the start of the experiment session (baseline) and again following each

exposure to the three eStatement interface designs. After participants' direct use with each prototype interface, their intent to switch to that version of the service was measured using the 0 to12 point scale (actually measured in inches on a ruler, equivalent to the 0-30 cm scale for preference gathering) on which participants placed a marker between two extremes labelled "Keep sending paper statements each month" and "Stop sending paper statements each month". Measurements were made from the 'keep sending' end of the scale such that higher scores indicate greater intention to suspend paper statements.

Absolute intention scores indicate the potential for each interface, including the initial size of the cohort ready to adopt such a service with no incentive of added functionality. Additionally, differences between the initial and post-experience values offer a measure of the change in intention to suspend statements under the influence of the eStatement facility experienced (and its corresponding functionality and usability). Salient user-interface characteristics and user attitudes highly related to usage intention can also be explored individually.

Despite the limitations of this measure, it was thought to be a simple and understandable method for participants to express their subjective intentions towards adoption.

6.5. Experiment 3: Usability of Alternative Interface Designs for eStatements

6.5.1. Experiment Design

The research reported here used a three-cell, repeated-measures within subjects experiment design. A sample of 182 participants (customers of the Case Bank) took part in the research. The cohort was roughly balanced for age, gender and presentation order of the interfaces, the between subject factors. Location and eBanking frequency of use were not balanced, but randomised in the sample and investigated as between-subject factors. The dependent variables were the responses to individual items in the usability questionnaire and the usage intention rating. Task completion and use of search and other features was also recorded along with their preferred option from the three designs.

6.5.2. Experiment Materials

Participants used one of three persona details sheets with five corresponding statement tasks on each interface (see Appendix G). The task and persona details were balanced among the cohort. The narrow focus of the tasks emphasised the varying eStatement functionality. They performed the same number and scope of tasks on the second and third interface but actual details were slightly altered to engage the user in their subsequent experiences. The alterations consisted name, date and amount changes whilst the tasks themselves were matched from interface to interface, see full details in Appendix G (p.320). Participants were asked to print statement items in one of the tasks and a desktop colour printer was provided for this purpose - set up and ready to print at the press of the print button on the interface or using the browser menu option.

6.5.3. Experiment Design Summary

Experiment 3: Level of Utility in eStatement Interface Designs

Design: Three cell, repeated measures, within subjects.

Independent Variables: Interface DO (Data Only)
Interface SS (Simple Search)
Interface AS (Advanced Search).

Participant Independent Variables: Age (3 levels)
Gender (2 levels)
Order of experience (6 levels)
Location (2 levels)

Dependent Variables: Attitude questionnaires;
Task completion & error counts
Usage intention ratings & preferences

Confounding variables: Researcher Bias (randomised)
Room (randomised)
Task sheet (balanced / matched task per task)

Other data: Think aloud remarks and researcher observations
Interview questions
Intention to use
Search logs.

Sample: 6 orders x 2 genders x 3 ages x 2 locations x over-sampling 4 = 192.

Honorarium: £20.

Session Time: 75 minutes.

6.6. Experiment 3: Results for eStatements Interfaces

6.6.1. Participants

The participant sample consisted of 182 (computer users, most of whom were Internet-savvy). They were also customers of the Case Bank. Four sets of data were removed due to incomplete data collection and technical problems. The cohort was taken from two cities in the UK - Edinburgh (125, 70%) and Cheltenham (53, 30%) as current and potential future users of eBanking services. Difficulties in recruiting participants in Cheltenham resulted in slightly fewer than the planned sample size, but sufficient numbers for statistical analysis were collected.

The resulting sample (N=178) consisted of 97 (54.5%) male and 81 (45.5%) female participants. They were recruited to approximate the age distribution of the Case Bank's customer base and 65 (37%) were under age 30; 45 (25%) were aged between 30 and 49; 68 (38%) were 50 years old or more. Experienced Internet users made up over 90% of the sample, 108 (61%) used the Internet on a daily basis and a further 53 (30%) used it at least once a week. A total of 108 (61%) used the Case Bank's eBanking service, 8 (7% of eBankers) using it on a daily basis and 52 (47% of eBankers) using it on a weekly basis - mainly for balance enquiries (88, 80%), transaction enquiries (86, 78%), funds transfers (57, 52%) and paying bills (50, 46%).

6.6.2. Performance

Most participants completed the tasks in the experiment without problems. The overall task completion scores were computed and are shown in Table 6.2.

Task completion levels were very high for all three eStatement interfaces. The advanced search interface achieved the highest rates overall. A repeated-measures ANOVA was run on the completion data for the three interfaces, with age, gender and order as the between-subject factors. Mauchly's test indicated that the assumption of sphericity had been violated, χ^2 (2) = 10.013, p = .007, therefore the degrees of freedom were corrected using the Greenhouse-Geisser estimates of sphericity (ε = .936). The results show that the completion scores differed significantly, F (1.873, 267.769) = 5.811, p = .004 (see Example 3 in Appendix A, p.283). Post hoc pairwise comparisons using the Bonferroni adjustment for multiple comparisons revealed that the significant difference lay between the Data Only

interface and the Advanced Search version, $p = .006$, with the search interface having higher performance scores. The difference between the Data Only and the Simple Search interfaces was marginally significant ($p = .083$).

eStatements by...	Mean Rate	St. Dev.	N	Lower Bound CI	Upper Bound CI
Data Only	4.72 (94.4%)	0.655	178	4.62 (92.4%)	4.82 (96.3%)
Simple Search	4.80 (96.0%)	0.514	178	4.72 (94.4%)	4.87 (97.5%)
Advanced Search	4.89 (97.8%)	0.381	178	4.83 (96.6%)	4.94 (98.9%)

Table 6.2. Task Completion Sum (Rate) for the Alternative eStatement Interfaces

There were no within or between-subject effects on the task performance. When location was included as a between-subject factor, this also had no effect on the result. Internet use and Internet banking status were not balanced across the cohort and therefore were not considered as between-subject factors.

Task completion rates were high for all three interfaces, indicating the utility of all three designs for performing statement lookup tasks. However, the significantly higher rates of performance for the Advanced Search reflected the assistance a search engine provided compared to manual trawling through transaction data sequences on multiple pages.

6.6.3. Attitude

The mean usability attitude scores for the three eStatement interface designs are shown in Table 6.3.

eStatements by...	Usability	St. Dev.	N	Lower Bound CI	Upper Bound CI
Data Only	4.346	0.980	178	4.202	4.492
Simple Search	4.481	1.090	178	4.320	4.642
Advanced Search	5.221	0.868	178	5.092	5.349

Table 6.3. Mean Usability Attitudes Toward the Alternative eStatement Interfaces

A repeated-measures ANOVA on the usability attitude responses with the mean responses to the three usability questionnaires as the within-subject variable and age group, gender and order of experience as between-subject variables, showed a significant difference between

the three designs, $F(2, 286) = 49.456$, $p < .001$. There was also a significant interaction between interface and order of experience $F(10, 286) = 2.452$, $p = .008$. There was also a significant between-subject order effect, $F(5, 143) = 4.663$, $p = .001$. There were no significant within or between-subject effects for age group or gender.

Pairwise comparisons in the repeated-measures ANOVA (using the Bonferroni adjustment for multiple comparisons) showed that there was no significant difference in usability between the Data Only and Simple Search interface designs, but that the Advanced Search interface was judged to have significantly higher usability than either of the other two (in both cases, $p < .001$). These results are summarised in Table 6.4. It can be seen then that the mean usability of the three search interfaces increased significantly from the Data Only and the Simple Search to the Advanced Search.

Pair		Mean difference	p
Data only	Simple search	-0.117	.635
Data only	Advanced search	-0.814	<.001
Simple search	Advanced search	-0.697	<.001

Table 6.4. Pairwise Comparisons of the eStatements Interfaces

The within-subject order effect is illustrated in Figure 6.6. It is clear from the graph that the Advanced Search (AS) interface obtained the higher scores within each order sequence. Also, it is clear that when this interface was experienced last in sequence, it obtained even higher scores. Conversely, when the Data only (DO) interface was experienced last in sequence, it received very low scores (around the neutral point of the scale). Finally, when the Simple Search (SS) was experienced directly after the Advanced Search, the scores for the simpler version were also very low (close to and below neutral on the scale). Generally, it can be summarised that these effects are derived from the interaction of two user responses: (a) the first interface experienced (whichever that was) received a moderately positive usability score (around 4.4 -5.2 on the 7-point scale) and, (b) the interface experienced immediately after the Advanced Search interface received a significantly reduced score. The order interactions therefore further serve to confirm that participants rated the Advanced Search interface as significantly better in terms of usability than either of the other two.

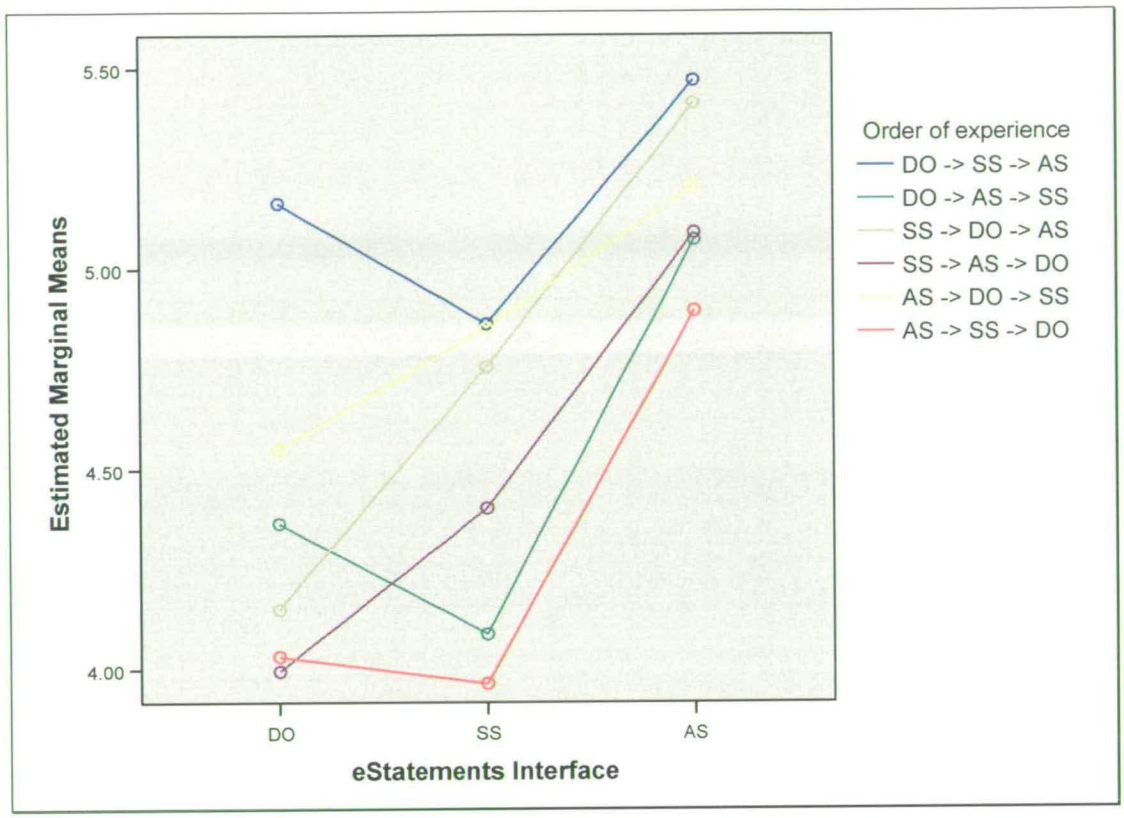


Figure 6.6. Order Effect on the Usability Scores for the eStatement designs

The between-subject order effects were examined using pairwise comparisons (Bonferroni). There were significant differences in the overall combined mean of the three usability scores due to the different orders of experience, as shown in Figure 6.7.

The progressive order (DO – SS – AS) obtained significantly higher usability scores for all three interfaces than the opposite order, AS – SS – DO, $p = .001$; mean overall usability scores were also higher for this progressive ordering than the orders DO – AS – SS, SS – AS – DO ($p = .018$ and .039 respectively). The figure indicates that when the Advanced Search was followed by another interface, scores for all three designs were significantly lower, with the exception of the order AS – DO – SS where all three scores were fairly high.

The analysis was also run using location as a between-subject effect, but this variable had no effect on the usability scores. Finally, inter-researcher differences were included in the ANOVA, and no researcher bias was found.

The individual scores for questionnaire attributes for the alternative eStatement interface designs are shown in Figure 6.8.

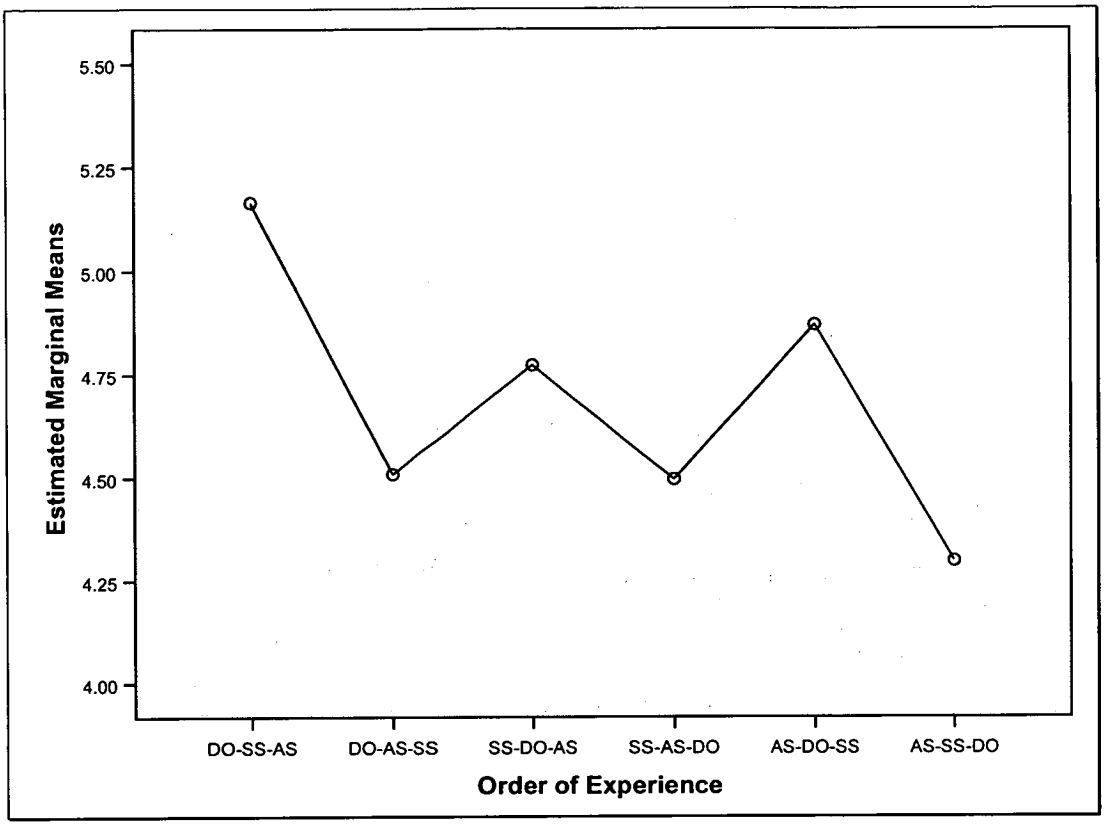


Figure 6.7. Between-Subject Order Effects on the Usability Scores

To examine the specific usability attributes responsible for the effects described above, separate repeated measures ANOVA analyses were carried out for each of the 28 attributes measured by the usability attitude questionnaire, with age, gender and order of experience as the between-subject factors. The main results and order effects are presented in Table 6.5. Stars in Figure 6.8 indicate that the score for the Advanced Search was significantly higher than either both other designs (**) or just one of the two other designs (*).

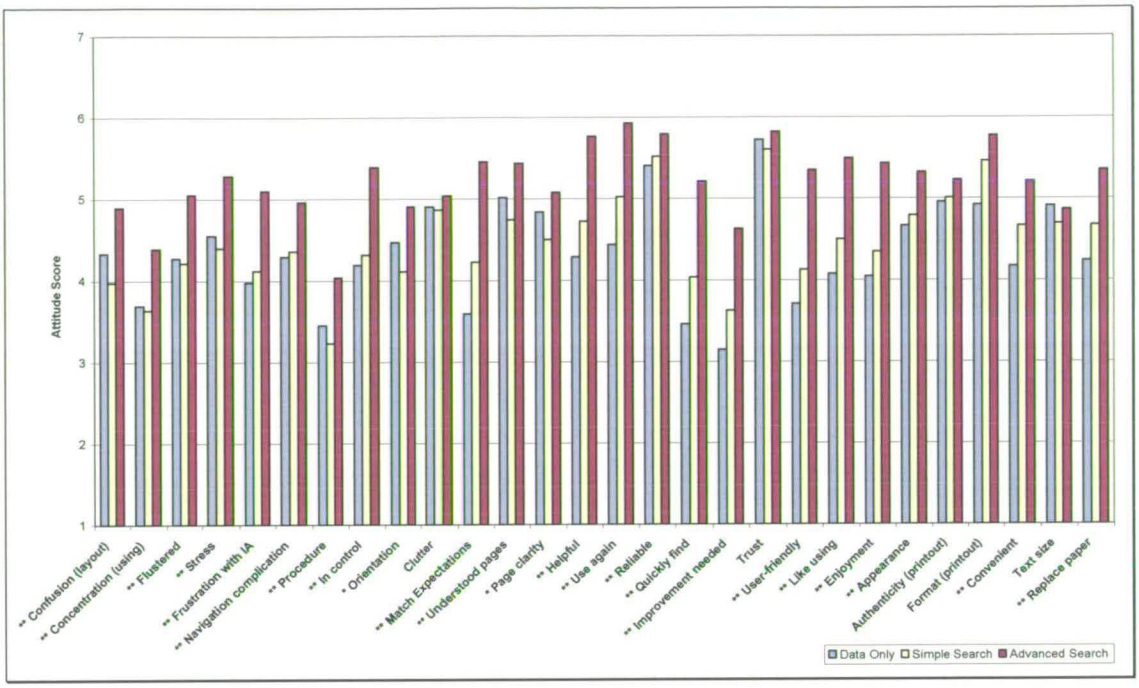


Figure 6.8. Comparisons of Usability Attitude Scores for the eStatement designs

In almost all cases where significant differences were found, the Advanced Search interface scored better than either of the other two designs. Over a wide range of usability issues covering affect, structure, content, quality and utility concerns, the Advanced Search interface was judged superior to the other two interfaces.

Attributes where no significant difference was found between the Data Only and Simple Search interfaces (again, where the Advanced Search interface was judged to be the best) mainly related to the affect group, navigation, IA, procedure, layout and appearance issues. In contrast, the Simple Search interface was judged to be better than the Data Only interface for attributes more concerned with statement task functionality – matching expectations, speed finding, like using, use again, improvement needed, convenience, suitability and as a replacement for paper. However, the Data Only design scored higher than the Simple Search for attributes relating helpfulness and understanding the page content, it was also marginally significantly better in terms of page layout.

Attribute	*p*	Pairwise Comparisons	W-S Order (*p*)	B-S Order (*p*)
Liked using	<.001	AS >> SS >> DO	<.001	.003
In control	<.001	DO = SS; AS >> DO, SS	.025	.002
Flustered	<.001	DO = SS; AS >> DO, SS		.006
Enjoyment	<.001	DO = SS; AS >> DO, SS	<.001	.003
Stress	<.001	DO = SS; AS >> DO, SS	-	.028
Navigation complication	<.001	DO = SS; AS >> DO, SS	-	.030
Orientation	<.001	AS = DO; DO = SS; AS >> SS	.044	.007
Frustration with IA	<.001	DO = SS; AS >> DO, SS	-	-
Quickly find	<.001	AS >> SS, DO; SS > DO	-	.002
Procedure	<.001	DO = SS; AS > DO; AS >> SS	-	-
Matched expectations	<.001	AS >> SS >> DO	<.001	.047
Confusion (layout)	<.001	DO = SS; AS >> SS; AS > DO	.043	.019
Concentration (using)	<.001	DO = SS; AS >> DO, SS	-	-
Improvement needed	<.001	AS >> SS, DO; SS > DO	<.001	.018
Reliable	<.001	DO = SS; AS >> DO, SS	-	.007
Liked appearance	<.001	DO = SS; AS >> DO, SS	.045	.025
Convenience	<.001	AS >> SS >> DO	-	-
Replace paper	<.001	AS >> SS, DO; SS > DO	<.001	-
User-friendly*	<.001[a]	DO = SS; AS >> DO, SS	.015	.008
Helpful*	<.001[b]	AS >> DO >> SS	<.001	.001
Understood pages*	<.001[c]	AS >> DO, SS; DO > SS	-	.028
Would use again*	<.001[d]	AS >> SS >> DO	<.001	.003
Suitability (printout)*	<.001[e]	AS >> SS >> DO	<.001	.045
Page clarity	.001	DO = SS, AS; AS >> SS	-	.001
Text Size*	.037[f]	DO = SS = AS[#]	-	-

Table 6.5. ANOVA results, Pairwise Comparisons and Order effects

Notes:
X >> Y – denotes a highly significant difference between pairs of interfaces in the Bonferroni pairwise tests, $p < .01$
X > Y indicates a significant difference at $p < .05$ (Bonferroni pairwise test)

[#] Although the main effect indicated significant differences between interfaces for this attribute, the pairwise comparisons with Bonferroni adjustment were not significant (or marginal) at $p < .05$.

* Where Mauchly's test indicated that sphericity had been violated, the df were corrected using the Greenhouse Geisser estimate, in the cases indicated in the table the statistics were:
[a] $\chi^2(2) = 7.410$, $p = .025$, $\varepsilon = .952$; [b] $\chi^2(2) = 9.092$, $p = .011$, $\varepsilon = .942$; [c] $\chi^2(2) = 11.646$, $p = .003$, $\varepsilon = .927$; [d] $\chi^2(2) = 8.365$, $p = .015$, $\varepsilon = .946$; [f] $\chi^2(2) = 13.833$, $p = .001$, $\varepsilon = .915$

There were no significant differences between the three interfaces on some of the attributes, relating to trust, authenticity of the printout and clutter. There were also no significant pairwise differences toward the text size attribute. To a degree, these non-significant results are as expected because many of the general interface design characteristics (text size, general clutter issues, branding etc.) and the format of the printout were matched (the same) across the three versions of the service.

In summary, the Data Only and Simple Search interfaces were judged to be similar with respect to many affect, structure, quality and content indicators. The Simple Search interface was judged to be better than the Data Only interface with respect to some task and search related-aspects. With respect to issues relating to trust, printout authenticity and clutter, all three interfaces scored similarly.

The Advanced Search produced the most positive attitudes to attributes in all categories of usability – affect, structure, content, quality and utility.

Analysis of the individual questionnaire attributes also showed significant interactions of order with the main effect (interface design) for the nearly half the individual attributes, as shown in Table 6.5. Plotting the results revealed the same patterns as was seen in the interactions between interfaces and the mean usability score. An illustration of the results is shown in Figure 6.9 depicting the interaction graph for the attribute, *would use again*.

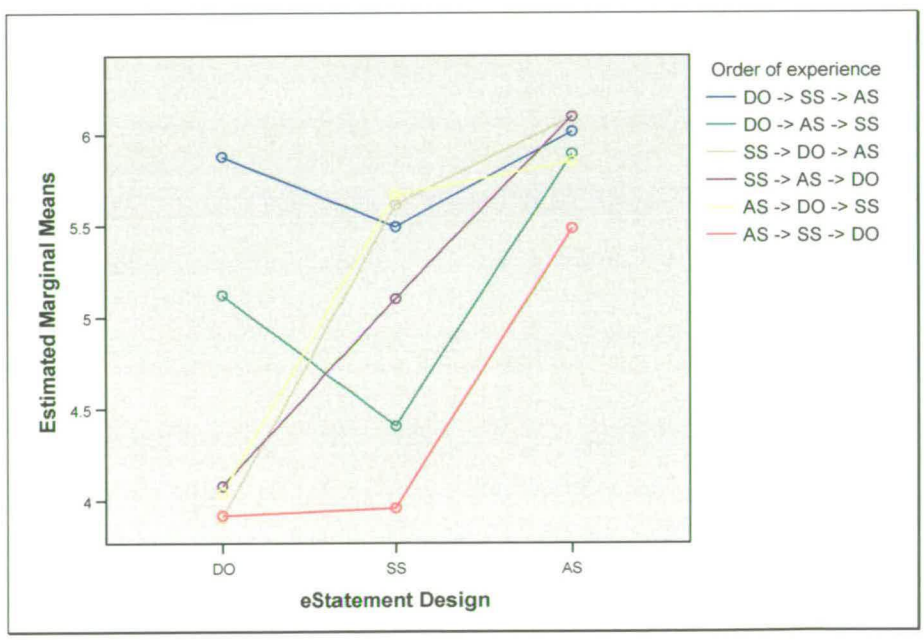


Figure 6.9. Illustration of the Order*Interface Interaction: Would Use Again

Similarly, between-subject order effects were also apparent for many individual questionnaire items, as shown in Table 6.5. Again, the individual results revealed the same patterns as seen for mean usability scores. An illustration of the results is shown in Figure 6.10 depicting the order effect for the attribute, *would use again*.

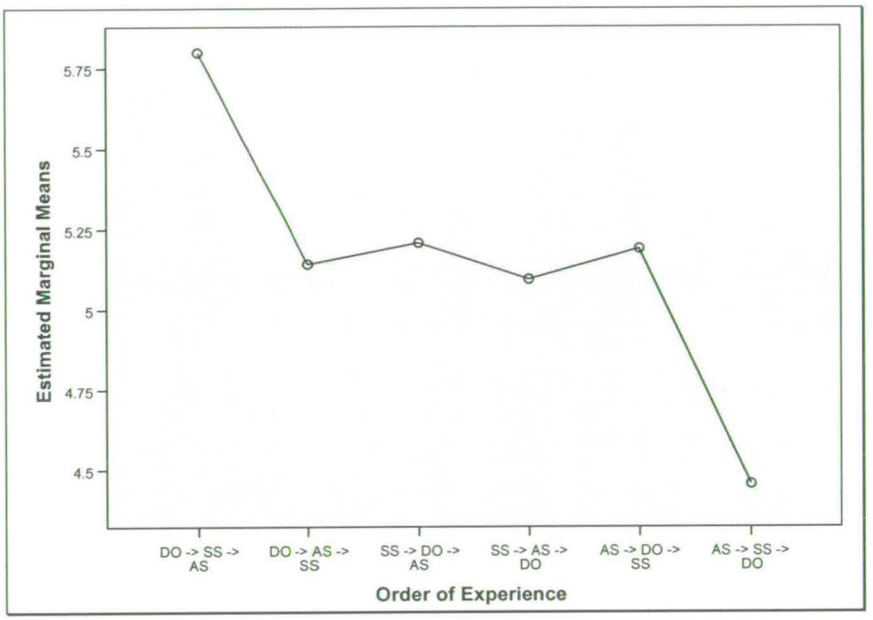


Figure 6.10. Illustration of the Between-Subjects Order Effect: Would Use Again

Again the progressive increase in functionality results in the highest overall scores, with the reverse producing the lowest overall scores. In this illustration, these two extremes exaggerate the effect, whilst other orders of experience do not cause such overall changes in scores.

Finally, there were some age and gender interactions and between-subject effects, Table 6.6. Figure 6.11 shows an illustration of the pattern seen for the age interaction with eStatements interface for matching expectations, user-friendliness and improvement needed (plotted). In this interaction, older customers gave higher scores to the data only and simple search variants than younger and middle age range customers, conversely, they also gave slightly lower scores to the advanced search than other customers. However, generally scores were comparable between age groups more for the data only and the advanced search designs.

Attribute	W-S Interactions (*p*)	B-S Effects (*p*)
Improvement needed	Age (.011) DO:O>Y,M; SS:O>Y>M; AS:Y>M>O	-
Matched expectations	Age (.018) DO:O>Y,M; SS:O,Y>M; AS:Y=M=O	-
Stress	Age (.030) DO:M>Y>O; SS:Y>O,M; AS:Y>M>O	-
Procedure	Age (.033) DO:Y=M=O; SS:Y>O>M; AS:Y>M,O	Age (.017) Y>M,O
Concentration	Age (.036) DO:Y=M=O; SS:Y>O>M; AS:Y>M>O	-
User-friendly	Age (.042) DO:O>Y,M; SS:O,Y>M; AS:Y=M=O	
Authenticity (printout)	-	Gender (.033) F>M
Helpful	-	Gender (.035) M>F
Orientation	-	Age (.002) Y>>O

Table 6.6. Age and Gender Within- and Between-Subject Effects

Notes:
Y = Young (<35 years), M = Middle age group (35 to 49 years) and O = Older (50+ years)

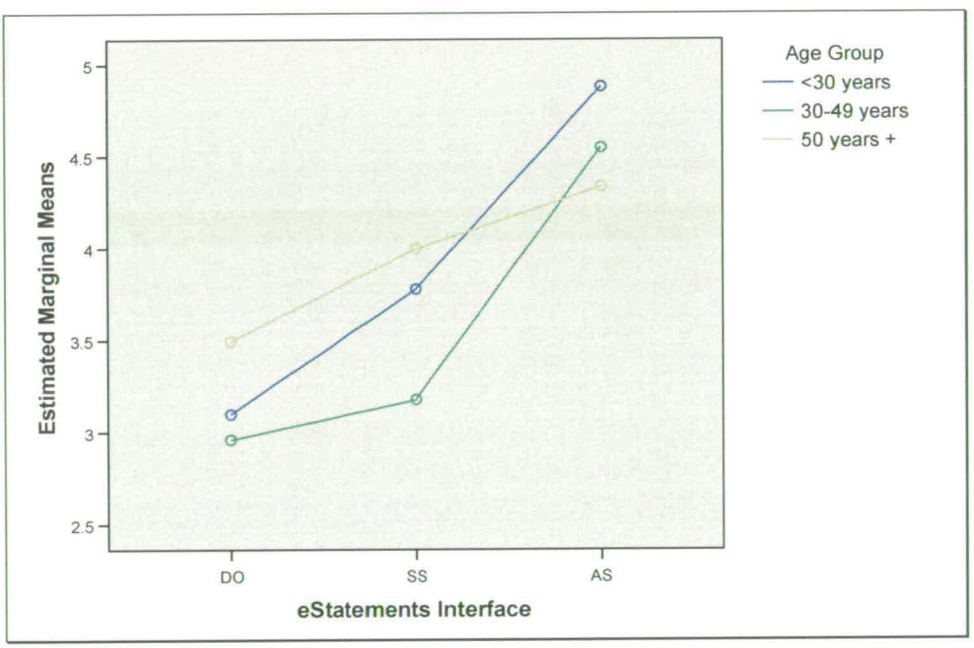


Figure 6.11. Illustration of the Age*Interface Interaction: Improvement needed

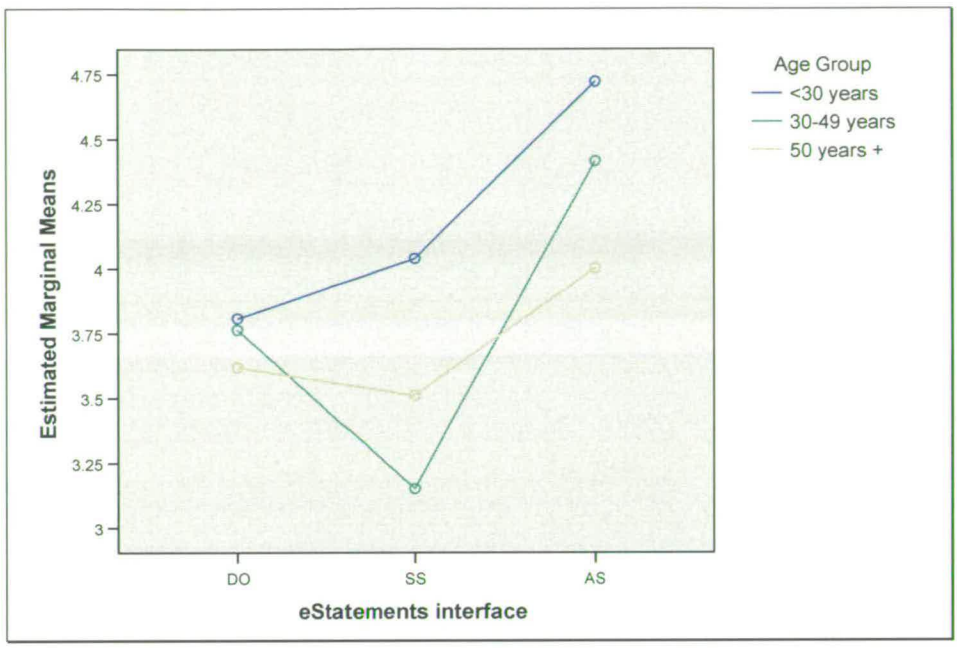


Figure 6.12. Illustration of the Age*Interface Interaction: Concentration (using)

Figure 6.12 illustrates the age interaction for concentration using (plotted), stress and procedure. This interaction shows a different picture, with younger customers having higher perceptions of these attributes for all three interfaces, with increasingly positive attitudes for the increasing functionality provided. Other age groups however saw a distinct dip in terms of these attributes when relating to the simple search, particularly prominent for those customers the middle age range (30-49 years). On these attributes, only the data only design was comparably scored between age groups.

Finally, some of the individual attributes obtained scores below 4.0 (7-point scale) and therefore indicated potential usability problems with the design. Poor scores for the Data Only design included: matching expectations and how quickly they could find items. For Data Only and Simple Search designs, participants felt they had to concentrate hard, struggled with procedures and felt improvements were needed. For the Advanced Search design, the lowest score was 4.0 (7-point scale) for procedure.

6.6.4. Intention to Switch to Paper Statements

Participant's scores on the intention to switch scale (as a surrogate for an adoption BI measure) for each interface, their initial scores and the relative changes in these scores (ΔBI)

are shown in Table 6.7. The initial intention to switch was based on the current live proposition of access to 2 months past transaction history, with no incentive of a comprehensive eStatement service. The sample mean initial intention to switch was measured at 3.59 on a 0 to 12 point scale.

The descriptive statistics and confidence intervals for the relative changes in intentions are displayed in Table 6.8.

eStatements by...	BI	St. Dev.	N	CI (Lower)	CI (Upper)	ΔBI
Initial	3.587	3.590	178	3.056	4.118	-
Data Only	4.194	3.687	178	3.648	4.739	+.607
Simple Search	5.286	4.006	178	4.693	5.879	+1.699
Advanced Search	6.607	4.194	178	5.987	7.228	+3.020

Table 6.7. Mean Usage Intention for the Alternative eStatement Designs

eStatements by...	ΔBI	St. Dev.	N	CI (Lower)	CI (Upper)
Data Only	.607	3.900	178	.0293	1.183
Simple Search	1.699	4.262	178	1.068	2.329
Advanced Search	3.020	4.264	178	2.389	3.651

Table 6.8. Relative Usage Intention Scores for the Designs

A repeated-measures ANOVA was run on the relative change in intention (ΔBI) for the three interfaces, with age, gender and order as the between-subject factors. Mauchly's test indicated that the assumption of sphericity had been violated, χ^2 (2) = 8.630, p = .013, therefore the degrees of freedom were corrected using the Greenhouse-Geisser estimates of sphericity (ε = .944). The results show that the relative intention scores differed significantly, F (1.889, 270.074) = 54.335, $p < .001$. Post hoc pairwise comparisons using the Bonferroni adjustment for multiple comparisons revealed that significant differences lay between each pair of interfaces, $p < .001$ in each case, with the Advanced search interface having the higher intention scores, see Table 6.9.

Note that all three changes in intention were positive (indicating an increased intention to switch to the electronic service for each of the three eStatement design interfaces used in this experiment) and that the change was progressively greater from the Data Only to the Simple

Search and to the Advanced Search interfaces despite the randomised order of presentation across the group.

Pair		Mean difference	*p*
Data only	Simple search	1.114	< .001
Data only	Advanced search	2.434	< .001
Simple search	Advanced search	1.320	< .001

Table 6.9. Pairwise comparisons of eStatements design variants

There were no significant within or between-subject effects on the intention data for the relative scores on each design. There was a marginally significant order interaction, F (9.443, 270.074) = 1.752, $p = .074$ (with the Greenhouse Geisser correction of the df).

It can be concluded from this analysis that each progressively more powerful interface significantly increases the expressed intentions of participants to switch from the use of paper statements to the eStatements service in eBanking, with the effect being very strong indeed for the Advanced Search design.

Analysis of the marginally significant interaction with order showed a similar effect to the usability analysis - the intention to switch was strongest for the Advanced Search design and was reduced for either of the other two interfaces when immediately following experience of the Advanced Search interface.

6.6.5. Alternative Usage Intention Estimations

The usage intention measure set out to estimate the strength of tendencies toward switching from paper statements to the eStatements service design just experienced in the experiment. A further examination of the intention scores is required to give additional indications of potential adoption rates for the service provider, that is: how many participants would initially be willing to give up paper delivery, and how much this might increase given the eStatement variants examined in the research sessions.

Adoption Estimates Based on Extreme Scores

The most cautious analysis of the intention measure would assume that those indicating extreme scores, at either end of the sliding scale (12 inches) would be considered entirely positive or negative with regard to switching from paper statements to online delivery. Those who scored zero would be indicating their total reluctance to switch from paper; those scoring twelve indicating a strong desire to switch to eStatements. These figures are shown in Table 6.10.

Status	Initial	Data Only	Simple search	Advanced search
Will switch	5 (3%)	9 (5%)	15 (8%)	23 (13%)
Will not switch	22 (13%)	21 (12%)	16 (9%)	12 (7%)
Total	27 (15%)	30 (16%)	31 (17%)	35 (20%)

Table 6.10. Participants with Extreme Scores on the Intention Scale

When these two groups of users are compared, it is again clear that the numbers prepared to switch increase with the addition of more historical statement data, and correspondingly for each increasingly powerful search facility. Again, the Advanced Search achieved a much higher level of predicted statement suppression than the other two options.

Adoption Estimates Based on Positive or Negative Intention Indications

Taking a less cautious approach, another analysis of the intention measure could assume that those indicating either positive or negative scores, either side of the scale midpoint of 6 inches, would be indicating positive or negative switching intentions. Those who scored exactly 6 on the scale would be indicating their indecision on the matter. These figures are shown in Table 6.11.

Status	Initial	Data Only	Simple search	Advanced search
Positive Intentions	38 (21%)	44 (25%)	66 (37%)	96 (54%)
Negative Intentions	131 (74%)	122 (69%)	99 (56%)	74 (42%)
Total	169 (95%)	166 (93%)	165 (93%)	170 (96%)

Table 6.11. Switching Intention Indications from the Neutral Point

The number of undecided (neutral) scores on the intention scale remained fairly consistent at between 8 and 13 participants at the initial stage, and for each design in turn. For the groups indicating positive or negative intentions, it is clear that the numbers considering the switch increase with the addition of more historical statement data, and correspondingly for each increasingly powerful search facility. Again, the Advanced Search resulted in many more positive intentions than the other two options, with this design over half the cohort indicated positive switching intentions on the scale.

Summary of Usage Intentions

Only the Advanced Search received mean usage intention ratings above the midpoint of the scale, with an estimate of between 13 and 54% of the cohort having positive intentions to switch with this design of eStatements in eBanking. Initial scores show that between 3 and 21% of the cohort were already showing positive intent to switch from paper to online delivery for their statements, without the incentive of a specific service or application.

6.6.6. Preferences for eStatements Interfaces

Asked for an explicit preference between the three eStatement options (see p.321), 143 (80.3%) participants expressed a preference for the Advanced Search facility, 7 (3.9%) for the Simple Search facility and 4 (2.2%) for the extended Data Only feature, with 24 (13.5%) not answering this question due to sessions running over time (and being cut short). This results in a proportion of 92.9% (of those who gave a preference) favouring the Advanced Search. This is a strong bias toward this design, with participants describing it as the kind of interface they would expect for a search facility.

6.6.7. Search Logs

Search usage was logged automatically in the experiment and a summary of the findings is presented in Appendix G (p.323). Task and search log analysis showed the benefits of the Advanced search, encouraging more searching and being more likely to include the target item in the results. Some potential problems with the search interfaces were highlighted to focus on in redesign:

- The statement 'sheet' metaphor – matching paper sheets – was not widely understood or considered helpful, monthly chunks were instead considered more natural.
- Mistakes occurring due to the restriction of using of only one from three criteria (corresponding to a Boolean 'OR') in the Simple search suggested that participants did not notice these limitations and a combination of fields ('AND') was more appropriate.
- The 'Particulars' option in the Advanced Search was not used as frequently as expected.
- The default 'Type' search, 'All transactions' was redundant in the Advanced Search, and several other options did not provide sufficient narrowing of results being too generic, e.g. 'All debits' and 'All credits'.

6.6.8. Qualitative Data

The qualitative data included researcher observations and participant commentary and interview responses. Some detailed comments are available in Appendix G. In summary, most participants considered that the amount range would be of most use. Suggestions regarding the addition of a keyword search provided further evidence of misunderstanding the particulars field and its usage. Some recommended this field be labelled 'keyword search'.

Many participants stated that (in real life) they wouldn't know which sheet number they were looking for. The sheet metaphor was not widely valued, and not all participants realised that their online statements were organised into sheets to match the paper equivalents.

When considering printouts from eStatement services, half the participants wanted the printout to look exactly like the (postal) paper statements from the Bank. Some people were concerned about the authenticity of the printout and whether it could be used to prove identity, address or apply for accounts. Others were only concerned with the data being accurate and easy to read.

Paper statements are often used as a form of address confirmation and for identity and financial-related purposes. As these issues were not the focus of the experiment design, no specific data were collected about the use of printed statements. What is required from eStatements services in terms of printouts is a topic which still needs work. For example,

Banks could provide official statements on request to eBankers, or allow them to print out identical pages to paper statement users for official purposes. For many people, however, being able to print data conveniently and accurately was the only requirement and a simple content pane print was suitable for this purpose.

Observations about each individual interface noted that many participants were not impressed by the data only version (DO). Particularly if they experienced it after one of the search interfaces, they expressed dismay, suggesting it was more difficult than using traditional paper versions. Paging back and forward through the online 'sheets' was time-consuming and tedious and accompanied by outward signs of frustration. Some even suggested that by now they would have given up and phoned the bank instead. It was also easy for participants to miss the target transactions when manually scanning multiple pages. Only providing the raw data was perceived to be functionally deficient.

In terms of the simple search (SS), some participants found it tricky to use and others even resorted to the next and previous paging buttons instead. The 'OR' nature of the search and the use of radio buttons posed the most problems: some participants were deleting items in previous fields *after* specifying new criteria in another field - this resulted in the blank field being selected when the search button was activated. Others automatically tried to fill multiple criteria, and this did not depend on whether they had previously used the advanced search. One participant commented that "It's more frustrating to have a service that is not up to standard than to have no search facility at all".

The advanced search (AS) was preferred, seen as the most useful and usable. There were however some hesitations observed during its use: the criteria were generally intuitively employed, but occasionally resulted in very large numbers of search results, and the paging metaphor used to group and display search results was seen to cause confusion (even though it is standard in many Web searches).

6.7. Analysis of the eStatements Usability Questions

6.7.1. Questionnaire Reliability

In order to assess the reliability of the usability attitude questionnaire, the coefficient of reliability (Cronbach's alpha, p.37)) was measured. The reliabilities of the questionnaires

applied to the individual designs were: Data Only design, $\alpha = .939$; Simple Search $\alpha = .961$; and Advanced Search, $\alpha = .951$. All alphas indicated good inter-item reliability.

Examining the resulting alpha if any item in the questionnaire was omitted revealed any attitude statements which could potentially be removed from the question set.

For the Data Only design, there were four items which resulted in very small increases in alpha if they were deleted. Removing *authenticity (printout)* or *text size* would result in $\alpha = .942$ (a very small increase) and removing *suitability (printout)* or *trust* would increase α by .001 ($\alpha = .940$). These are therefore not strong candidates for deletion.

Similarly for the Simple Search, only two attributes were identified: removing *authenticity (printout)* or *text size* would result in $\alpha = .962$ (again a very small increase). These are therefore not strong candidates for deletion.

For the Advanced Search, these two candidates again were identified: *authenticity (printout)* and *text size*. Removing either would result in $\alpha = .953$ (a very small increase).

Although consistently appearing, these increases are very small, indicating no strong evidence that they should be removed from the questionnaire.

The same two items are found when the questionnaires from the different interfaces are pooled, $\alpha = .954$ with a small increase of .002 when either *authenticity* or *text size* are removed. Overall, this is not strong evidence for removing any items.

6.7.2. Analysis of Neutral Responses

Generally, participants did not select the neutral response in expressing their attitudes to the eBanking usability characteristics posed in the questionnaire, see Table 6.12. Out of a total of 4984 responses to questions for each interface (28 questions, 178 participants), the designs were given a neutral score 8.3-10.1% of the time, with the proportion decreasing for each increasingly sophisticated statement interface.

For the Data Only design, the highest frequency of neutral scores (28) was for the attributes *appearance* and *replace paper*, scored neutral by just 15.7% of the participants. The second highest frequency was for *Stress* which scored neutral 27 times (15.2% of participants).

Similarly, for the Simple Search design, the highest frequency of neutral scores (37) was found for the attribute *appearance* (20.8% of participants). The second highest frequency

(26) was associated with *matched expectations* (14.6% of participants) followed closely by 25 participants (14.0%) scoring neutrally for *fluster*, *stress* and *authenticity (printout)*.

For the Advanced Search design the highest frequency (27) was for *stress*, but again this was only 15.2% of the cohort, the remainder scored neutral by less than 15% of the cohort, 24 participants (13.5%) scoring *replace paper* neutrally.

This analysis shows clearly that more than 80% of participants could respond positively or negatively towards all the attributes in the questionnaires for the Data Only and the Advanced Search designs. Only the Appearance attribute for the Simple Search design came close to being an attribute which was not as strongly associated with interface use and real-world potential, but with less than 21% of people feeling neutral towards this interface characteristic, this is not a strong contender for removal from the questionnaire attribute list.

Therefore, all the items were retained in the questionnaire, with each considered a valid contribution to overall interface perceptions.

Attribute	Data Only	Simple Search	Advanced Search
Would use again	7 (3.9%)	4 (2.2%)	2 (1.1%)
Suitability (printout)	10 (5.6%)	12 (6.7%)	4 (2.2%)
Understood pages	16 (9.0%)	17 (9.6%)	5 (2.8%)
Quickly find	9 (5.1%)	10 (5.6%)	6 (3.4%)
Helpful	13 (7.3%)	14 (7.9%)	6 (3.4%)
In control	20 (11.2%)	15 (8.4%)	7 (3.9%)
Orientation	16 (9.0%)	9 (5.1%)	10 (5.6%)
Trust	18 (10.1%)	20 (11.2%)	10 (5.6%)
Concentration (using)	16 (9.0%)	12 (6.7%)	11 (6.2%)
Procedure	16 (9.0%)	7 (3.9%)	12 (6.7%)
Reliable	24 (13.5%)	20 (11.2%)	13 (7.3%)
Appearance	**28 (15.7%)**	**37 (20.8%)**	13 (7.3%)
User-friendly	16 (9.0%)	14 (7.9%)	14 (7.9%)
Confusion (layout)	17 (9.6%)	14 (7.9%)	15 (8.4%)
Liked using	22 (12.4%)	19 (10.7%)	17 (9.6%)
Enjoyment	20 (11.2%)	16 (9.0%)	18 (10.1%)
Frustration with IA	15 (8.4%)	18 (10.1%)	18 (10.1%)
Matched expectations	22 (12.4%)	**26 (14.6%)**	18 (10.1%)
Authenticity (printout)	16 (9.0%)	**25 (14.0%)**	18 (10.1%)
Page clarity	16 (9.0%)	17 (9.6%)	19 (10.7%)
Improvement needed	13 (7.3%)	16 (9.0%)	20 (11.2%)
Convenience	22 (12.4%)	21 (11.8%)	20 (11.2%)
Flustered	24 (13.5%)	**25 (14.0%)**	22 (12.4%)
Navigation complication	12 (6.7%)	14 (7.9%)	22 (12.4%)
Text Size	22 (12.4%)	21 (11.8%)	22 (12.4%)
Clutter	19 (10.7%)	22 (12.4%)	23 (12.9%)
Replace paper	**28 (15.7%)**	21 (11.8%)	**24 (13.5%)**
Stress	**27 (15.2%)**	**25 (14.0%)**	**27 (15.2%)**
Total	*504 (10.1%)*	*491 (9.8%)*	*416 (8.3%)*

Table 6.12. Count and Percentage of Neutral Scores in the Questionnaires

6.8. Relationships between Metrics

6.8.1. Usability Attitudes and Usage Intentions

It was also of interest to this study to investigate possible correlations between participants' expressed intentions to switch to paper statements and their usability attitudes toward each eStatement search design. There was a strong correlation in all three cases between usability attitude and the relative score for intention to switch (ΔBI) - a Pearson correlation analysis showed in each case a highly significant positive correlation as shown in Table 6.13.

Design	Correlation (r)	Significance (*p*)
Data only	.399	<.001
Simple search	.495	<.001
Advanced search	.417	<.001

Table 6.13. Correlations between Usability and Relative Intent to Switch

Similarly, there was a strong correlation between perceived usability and the absolute score for intention to switch, as shown in Table 6.14.

Design	Correlation (r)	Significance (*p*)
Data only	.367	<.001
Simple search	.441	<.001
Advanced search	.419	<.001

Table 6.14. Correlations between Usability and Intentions to Switch

Usability perceptions and usage intention indications were very highly significantly correlated. The highest relationship was seen for those participants scoring the simple search highly for usability also scored high intentions to use the service. Similarly, the correlation was also strongly positive for the Advanced search. The scatter plots and results for relative differences between pairs of eStatement interfaces can be found in Appendix G, p.331-2. In summary, the mean usability to intention correlations for the interface pairs were: AS-SS, $r = .569, p < .001$; AS-DO, $r = .413, p < .001$ and SS-DO, $r = .592, p < .001$.

It is possible that the extensive use of paging for task completion when using the simple search (see Appendix G, analysis of search logs, p.323 for more details) influenced the

usability and intention measures resulting in the higher correlations of usability and intention for this search design, such that those who scored it more usable also offered more highly positive intention scores and vice versa. In contrast, for the advanced search, intentions were higher generally, despite some usability concerns.

Individual Attributes in the eStatement Usability Questionnaire

In order to determine which characteristics of the eStatement interface were most highly correlated with intentions to use, the correlations were computed for each attribute in the questionnaire; the absolute scores for each eStatement design were computed and compared to the absolute intentions to switch scores. Further analysis on the relative scores and relative differences in interfaces can be found in Appendix G. The correlations are shown in Figure 6.13 and Table 6.15.

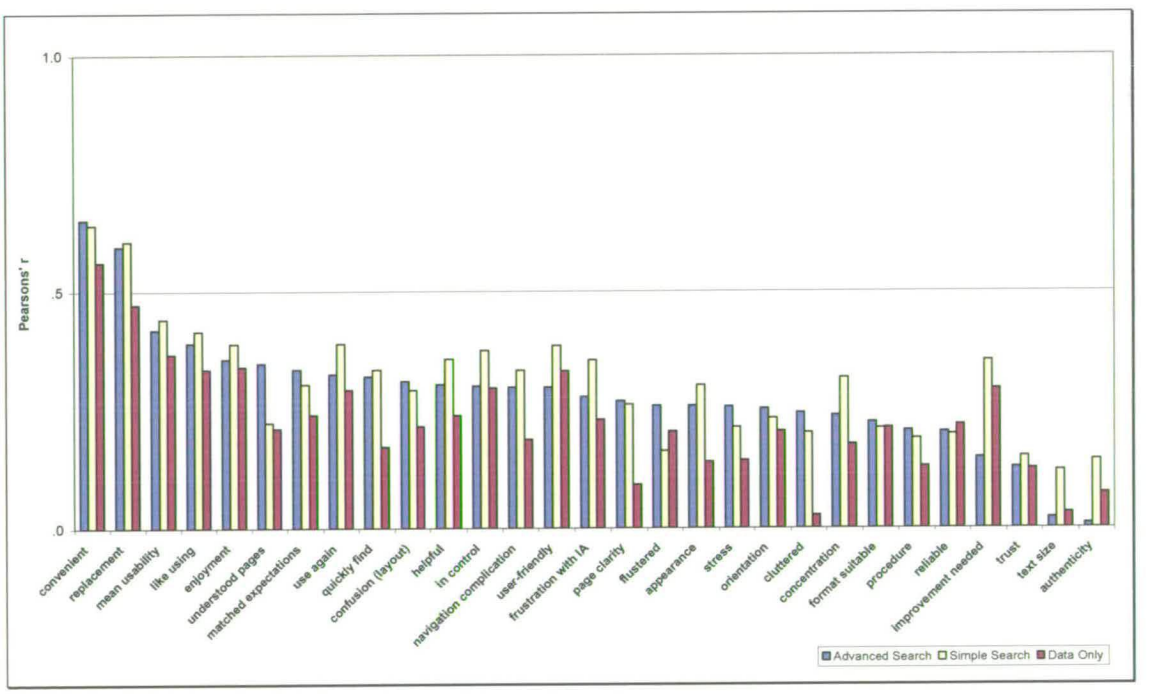


Figure 6.13. Correlation Coefficients between Usability and Intention Scores

The statement search tasks and eStatement interfaces and search designs evaluated considered a fairly specific banking need within a focused information space. This was even more focused than the previous eBanking tasks, and much more personal and account

specific than the information searches on the Web portal designs. It was anticipated that the salient characteristics of the interfaces were likely to have similarities to both the general information retrieval tasks and other eBanking tasks. Moreover it was expected that the utility-focused additional attributes would also be very highly correlated with intentions.

The correlations between usability perceptions and intention to use were very high with respect to convenience and replacing paper for both search interfaces, $r > .5$.

For other correlations, the values were not as highly correlated as previous experiments looking at usability – preference relationships. However, many of the correlations were still highly significant. For all three eStatement interface designs, it was apparent that the two utility concepts of convenience and replacing current paper-based Statements had the highest and very significant relationships with intentions. This shows that it was the relative advantage items which were dominant. For the Advanced search, convenience scores explained 42.5% of the variance in usage intention scores.

Additionally, salient qualities relating to usage intention scores for the Advanced Search design included many of the attributes from the *affect* (like, enjoy, user-friendly etc.), *structure* (IA, navigation, quickly find etc.) and *content* design (understood, helpful, layout confusion etc.) groups. Also highly significantly correlated with usage intention scores were the *quality* attributes of use again and appearance.

As most of the attributes were highly correlated with usage intention scores, noting those which did not appear to be related to usage intentions indicated that the *authenticity* of the printout was not a salient factor influencing eStatements usage intention; similarly, *trust* and *text size* were not significantly correlated with intentions and *improvement needed* was only marginally significant.

For the Simple search design, the correlations were also high and very similar to the pattern seen for the Advanced search. For this design, only *text size* was not significantly related to usage intention, with *trust* and *authenticity* marginally significant only.

The correlations for the Data Only design were not as high as the two search interfaces, only convenience breaking $r > .5$ (at $r = .562$, $p < .001$). Again, however, many of the affect, structure and content design attributes were highly related to usage intention. Those attributes which were not significantly related to usage intention included: *clutter*, *text size*, *authenticity*, *page clarity*, *trust*, *procedure*, *appearance* and *stress*.

Attribute	Advanced Search		Simple Search		Data Only	
	r	*p*	*r*	*p*	*r*	*p*
convenient	***.652***	***<.001***	***.641***	***<.001***	***.562***	***<.001***
replacement	***.595***	***<.001***	***.606***	***<.001***	***.472***	***<.001***
like using	**.390**	**<.001**	**.415**	**<.001**	**.335**	**<.001**
enjoyment	**.357**	**<.001**	**.389**	**<.001**	**.340**	**<.001**
Understood pages	.348	<.001	.222	.003	.210	.005
matched expectations	.335	<.001	.303	<.001	.239	.001
use again	.325	<.001	.389	<.001	.292	<.001
quickly find	.320	<.001	.334	<.001	.172	.021
confusion (layout)	.310	<.001	.291	<.001	.215	.004
helpful	.303	<.001	.357	<.001	.238	.001
In control	.300	<.001	.375	<.001	.296	<.001
navigation complication	.297	<.001	.333	<.001	.187	.012
user-friendly	.297	<.001	.385	<.001	.332	<.001
Frustration with IA	.277	<.001	.355	<.001	.229	.002
page clarity	.268	<.001	.261	<.001	.092	.221
flustered	.258	.001	.163	.029	.204	.006
appearance	.258	.001	.301	<.001	.140	.063
stress	.256	.001	.213	.004	.143	.057
orientation	.252	.001	.232	.002	.205	.006
cluttered	.243	.001	.201	.007	.027	.718
concentration	.238	.001	.317	<.001	.177	.018
format suitable	.224	.003	.211	.005	.212	.005
procedure	.206	.006	.189	.012	.131	.082
reliable	.203	.007	.198	.008	.218	.003
improvement needed	.149	.047	.353	<.001	.294	<.001
trust	.129	.087	.151	.045	.126	.095
text size	.022	.773	.122	.106	.033	.666
authenticity	.009	.901	.144	.055	.074	.328
mean usability	***.419***	***<.001***	***.441***	***<.001***	***.367***	***<.001***

Table 6.15. Significant Correlations between Usability Attributes and Usage Intentions

Again, some of the correlations are higher than previous publications would suggest, but in general, they are lower than for pilot studies or for the eBanking transaction interface tasks.

The mean scores across the full set of eStatement usability attributes offer a good indicator of intention to use: with 64 highly significant attributes and 7 significant aspects of all three interfaces having high relationships with intention scores.

Where correlations are all highly significant for all three computations, and above $r = .3$ (lowered from .5 to account for the significance of a correlation of .3 at this sample size (Bryman & Cramer, 2001), see p.37), this indicates the characteristics of the eStatement interface which associate well with final intention to use scores. For the eStatement tasks it was convenience, replacement, like using and enjoyment (as well as the mean usability scores) which most highly correlated with intentions across the three designs. All items were positively associated with intentions.

Data for the differences between two pairs of interfaces is presented in Appendix G, p.333.

From the differences in scores between interfaces, the most differentiating characteristics were convenience, like using, match expectations, speed finding, user-friendliness, replacement for paper, use again, control, improvement, enjoyment and helpfulness. Again, printout authenticity, text size, trust, clutter and appearance were the least important differentiators of usage intentions.

6.8.2. Usability – Attitudes and Performance

Examining the relationship between usability attitude scores (means) and task performance on each interface, the scores were not correlated for the Data Only design, $r = .033$ ($p > .05$). However, for the search designs there were significant positive correlations between attitude and performance:

- Simple Search: $r = .265$ ($p < .001$)
- Advanced Search: $r = .158$ ($p = .035$)

In terms of the individual usability attributes and their association with performance, *reliable* ($r = .168, p = .025$) was the only item significantly correlated for the Data Only design.

For the Simple search design many items were correlated:

- Concentration (using) $r = .203$ ($p = .006$)
- Flustered $r = .175$ ($p = .020$)

- Stress $r = .224$ ($p = .003$)
- Frustration with IA $r = .202$ ($p = .007$)
- Navigation Complication $r = .308$ ($p < .001$)
- Procedure $r = .180$ ($p = .016$)
- Control $r = .270$ ($p < .001$)
- Orientation $r = .189$ ($p = .012$)
- Understood pages $r = .184$ ($p = .014$)
- Helpful $r = .248$ ($p = .001$)
- Use again $r = .206$ ($p = .006$)
- Reliable $r = .216$ ($p = .004$)
- Quickly find $r = .161$ ($p = .032$)
- Improvement needed $r = .206$ ($p = .006$)
- User-friendly $r = .226$ ($p = .002$)
- Like $r = .218$ ($p = .004$)
- Enjoy $r = .212$ ($p = .005$)
- Appearance $r = .199$ ($p = .008$)
- Convenient $r = .270$ ($p < .001$)
- Replacement $r = .286$ ($p < .001$)

Although many correlations were significant, they did not represent strong effects, with less than 10% of the variance in individual items explained by performance differences (or vice versa).

For the Advanced search, there were some significant positive correlations, but again the strengths were weak:

- Confusion (layout) $r = .156$ ($p = .037$)
- Concentration (using) $r = .167$ ($p = .026$)
- Stress $r = .175$ ($p = .020$)
- Control $r = .171$ ($p = .022$)
- Helpful $r = .198$ ($p = .008$)

- Use again $r = .194$ ($p = .010$)
- Improvement needed $r = .163$ ($p = .030$)

6.8.3. Usability Performance and Usage Intention

Finally, possible associations between task performance and the absolute and relative (design – initial) usage intention scores were computed.

Examining the absolute intention scores, the results for the Data Only design were not significant, $r = .040$ ($p > .05$). For the Advanced search similarly, $r = .038$ ($p > .05$). However, the simple search did obtain a marginally significant positive correlation, although a weak strength: $r = .148$ ($p = .049$).

Examining the relative usage intention scores in contrast, the Data only design still had no association between performance and intention to use $r = .063$ ($p > .05$). The search interfaces showed marginally significant correlations: $r = .128$ ($p = .089$) for the Simple Search and $r = .138$ ($p = .067$) for the Advanced Search.

6.8.4. Comparison of Metrics

Usability (attitude and performance) and intention scores for the three eStatements interface designs (and initial intention scores) were compared using normalised scales, see Figure 6.14 illustrating the differences in responses toward the three interfaces. The error bars indicate the 95% CI of the means.

The means and CIs displayed by the performance metric show the overlap between the interfaces, clearly illustrating the lack of perceived differences in ability to perform statement searches with the alternative designs. Conversely, differences between usability attitude scores were not apparent for the Simple search and Data Only designs, but attitudes are clearly more positive for the Advanced search interface. Similarly, confidence intervals around the absolute intention scores indicate some overlap between initial perceptions and those of the Data Only interface, but clear increases for the two search designs.

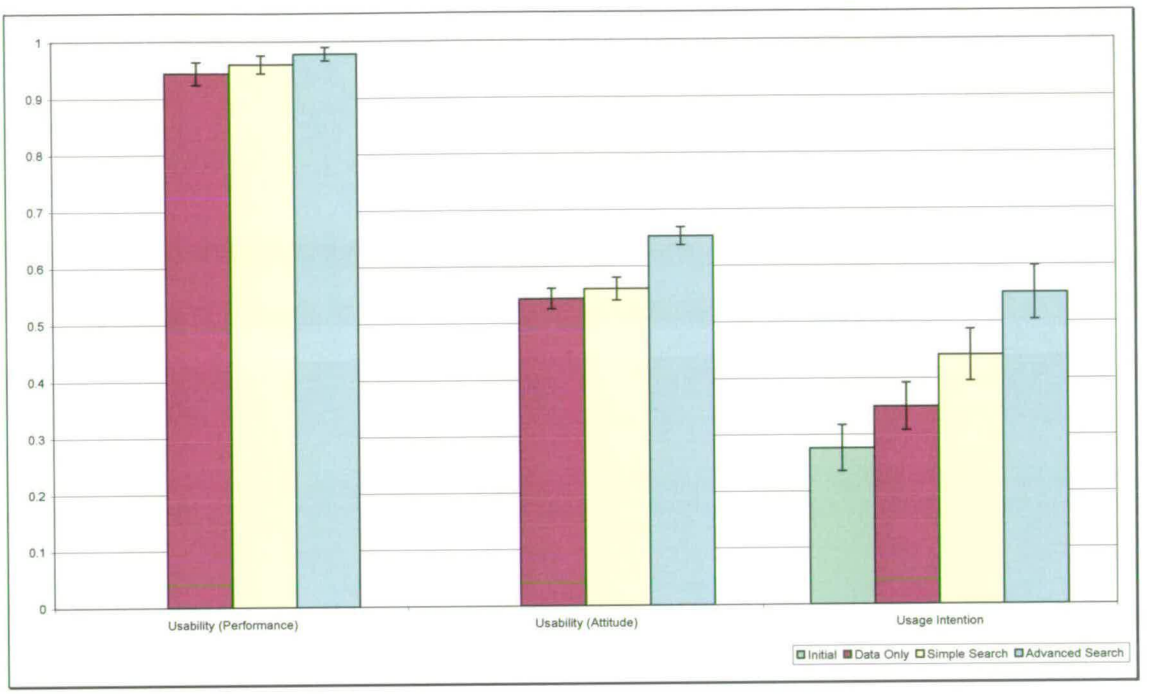


Figure 6.14. Comparison of Normalised Usability and Intention Metrics for eStatement Designs

6.8.5. Hypotheses Summary

Hypothesis and results for Experiment 3 – eStatement Functionality

Hypothesis H_0 E3a: The different eStatement interfaces will not result in different usability attitude and performance scores.

Rejected:

The different eStatement design variants resulted in significantly different user attitudes toward usability and some significant differences in task performance.

Hypothesis H_0E3b: The different eStatement interfaces will not result in different usage intentions.

Rejected: The different eStatement design variants resulted in significantly different usage intentions.

Hypothesis H_0E3c: There will be no relationship between the usability measures of performance, attitude or usage intention.

Partially Rejected:

There was a *consistently significant and strong positive* correlation between measures of usability attitude and usage intentions.

There was an positive weak but not consistently significant relationship between measures of usability attitude and performance.

There was a consistent weak, not significant (at best marginal) correlation between measures of performance and usage intent.

6.8.6. Real World Usage

A usage metric related to intentions was developed in order to investigate the association between usability (evaluated at an early stage) and potential success in the real world. In an attempt to demonstrate that human factors and usability practices yield benefits worthwhile to business (Chapanis, 1991), intent to switch from a traditional to an online service was investigated in relation to usability and at three different levels of functionality. Although real world usage cannot be determined at early evaluation stages, a comparison with actual usage of the production interface can offer reflections on the merit of the data provided by the intention metric.

When a wide range of interface characteristics, ease of use, usefulness and quality issues are considered in a subjective usability attitude measure, usability is positively related to usage intent. Indeed, measured usability differences between different interfaces followed the same patterns as the measures of switching intentions.

Although the experiments showed a small number of people willing to adopt without the incentive of a specific service offering, most found the incentive of high functionality services and good usability influenced their desire to switch. By measuring a rating of intentions towards switching behaviour, extreme views on this scale were examined to provide a conservative estimate of potential usage for the business case, whilst intention measures showing positive or negative tendency can show a more optimistic view of the potential interest in the marketplace.

The experiment data provided a conservative estimate that 13% of customers would switch to eStatements for the advanced search service (p.217). This was increased to 54% when considering switching tendencies using positive or negative indications on the intention scale (p.218).

The production eStatements search is shown in Figure 6.15. If this design is assumed to be roughly equivalent in functionality to the Advanced Search design used in the experimental evaluation, the actual uptake figures can be compared to the adoption predictions inferred from the experimental data.

Figures from the Case Bank indicate that uptake was 538,713 accounts having paper statements switched off (some 436,466 unique customers). They calculated that 23% of accounts were switched (Internal communication, 2005) This figure of 23% of accounts indicates the adoption of the online search in preference to paper statements – a measure of real-world success of the interface design project. This figure lies between the two estimates

based on conservative (extreme) scores of intention on the scale and tendencies (positive or negative) from neutral on the intention scale.

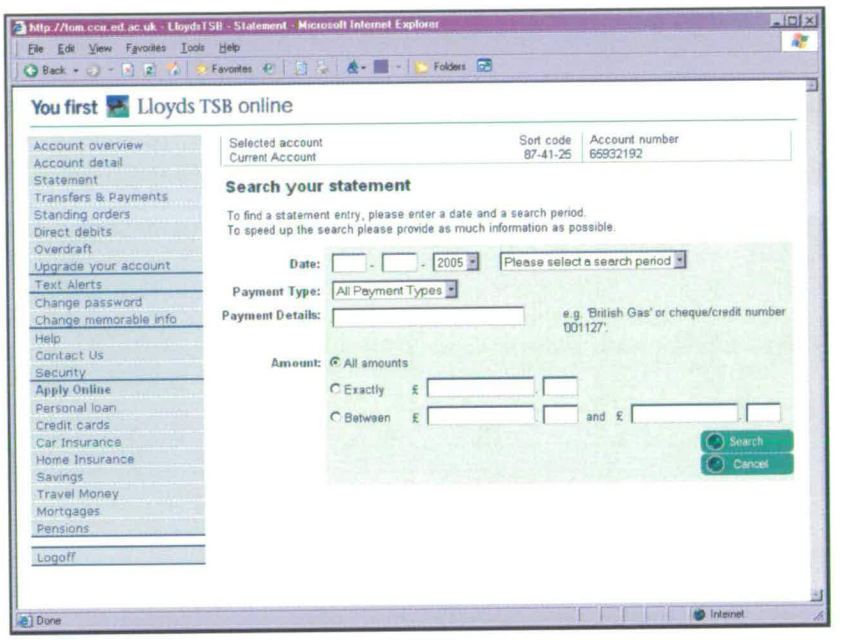


Figure 6.15. Production eStatements Search Design

Although usage intention scores taken during early evaluations are limited in their comparability with real-world take up data (due to being estimated based on an early design prototype, rather than for the optimised production interface), comparing these estimates with real world usage data determines the value of a self-report intention scale in usability and adoption research.

The research findings allowed the Case Bank to produce a highly usable and successful design for eStatements search using the prototype Advanced Search option (evaluated as the most promising of the options tested in this experiment). This design was further optimised and tested following the design recommendations resulting from this evaluation, e.g. the promotion (to nearer the top of the search options) of the keyword search, with a more appropriate field label of 'Payment Details' and the inclusion of examples to aid its use, as shown in Figure 6.15. The high numbers of accounts and customers switching from traditional paper statements to the online service demonstrates the benefit of carrying out the usability engineering methods developed in this thesis.

6.9. Discussion

As utility increased, usability attitude and performance scores also increased and so did intentions to switch. Subjective evaluation of usability and intention to switch were positively correlated at each level of the statement functionality evaluated; this provides evidence that usability (measured as attitudes toward a broad range of usability issues) is a good indicator of usage intention. This finding supports the work of other researchers (Davis, 1989; Dillon and Morris, 1999). Usability engineering typically involves early prototypes so actual adoption measures cannot be gathered. Nevertheless, at this early stage, subjective usability and self-report intentions are capable of providing statistically robust and detailed data which can inform redesign and choice of interface utility level. This can be used to build a case for additional functionality to be developed.

For the service provider to gain maximum advantage, the most promising early design must be optimised with a view to maximising usage intentions. Subjective and objective usability (attitude and performance) and usage intentions were all highest for the Advanced search. Search logs and observations also showed that the Advanced search appeared to be the most effective and efficient at finding the target data. Only the Advanced search design achieved mean switching intentions of greater than the midpoint of the scale. In contrast to market research (Graeber et al, 2003), age, gender and experience of using eBanking had no effect on intention scores.

Some participants were willing from the outset to switch to online statements, and in this group, each increase of functionality raised the levels of these intentions. Others were not so willing to switch, and remained so despite the functionality changes. Observations and comments indicated that some participants (just over 10% of this sample) were committed to keeping paper statements for tax and record keeping purposes. Many desired to search their online statements, but requested both paper and online access. Extending an online statement service to eBankers could reduce the need for branch or telephone involvement in statement search tasks, even if paper statements were also retained. Forcing a choice between paper and online statements may not be the best strategy for the service provider.

For the purposes of estimating potential adoption levels for the service provider, participants giving the most extreme values of intent to switch (0% and 100% on the scale) were considered to be displaying a complete willingness or reluctance to switch. This was consistent with observations and participants remarks. This data provides an indication of the

proportions who might switch to the various eStatements services. The advanced search was seen to have most effect on increasing proportions willing to switch and decreasing proportions of those reluctant. Unfortunately, the relationship between intention and real world usage has not been sufficiently investigated in previous studies, neither was it able to be performed in this research.

Qualitative data collected also helped diagnose and propose alternatives to various interface issues. One topic that stood out was the issue of the 'particulars' field for the Advanced search. 'Particulars' was chosen to match the name of the statement column where the search engine looked for matching terms. Consistency between the field name and the column heading complied with usability heuristics (e.g. Faulkner, 2000, pp. 188-9). This is an example of the limitations of user-interface design following heuristics and principles, but without consulting users directly. In this case, consistency in the use of terms was secondary to matching the expectations of users. The column heading called particulars on the statement has not been identified as any concern in real life and previous usability tests. Yet in the context of tasks involving statement search, the particulars field was misunderstood in two ways:

- 'Particulars' was a poor choice of label in that it was not recognised by participants, who were generally unaware of the particulars column on the statement. It would be more appropriate to call them both either 'transaction description' or 'description'.
- Others assumed the search field was a keyword search (like Google). A general assumption was that it would search the whole statement.

Changing both labels to include the word 'Description' and explicitly state 'Search the transaction descriptions', with the inclusion of an example would reduce these problems. Using a keyword search of the whole statement record might be more consistent with expectations. Placing the keyword search at the top of the advanced set of search criteria might also increase its usage and in turn reduce search results and facilitate more efficient information retrieval.

Comments and observations also indicated that the Search button could be relocated at the top of the statement page to increase the prominence of the new service. Grouping search, print and paging functions together might help users navigate their statements with less hesitation. Placing these initial functions at the top of the page would also be expected to increase their salience and usage (Lidwell et al, 2003).

Further research into the sheet metaphor for paper statements would be helpful to determine how statements should be displayed online. It would be interesting to explore whether or not this metaphor is necessary for an eStatement service, as rolling data or personalised chunks could be offered instead. Typically a match between various banking channels was valued by eBankers in previous research (Weir et al, 2006), however eStatements may be an application where this does not hold true. Individual personalisation (by the user) of eBanking might offer the best solution to some of these display and preference options. Demographic (or alternative) automated customisations might also be appropriate to alter data displays and further research in these areas is required.

Overall, there was strong evidence to support the case for the service provider to develop a comprehensive eStatements interface. Data predicted reasonable intentions to switch from paper statements in the case of an advanced search provision. The relationship between usability and usage intentions has been established in this experiment. Replication of this relationship under a wide range of conditions, and showing that usage intention relates to real world usage, would make a stronger case for involving usability engineering from an early stage of interface and functionality design for a range of critical systems.

6.9.1. Limitations

A significant limitation of the study was the inability to capture real-world usage data for the interfaces in an early evaluation phase. The real world usage data came from the production interface, which was slightly different from any of the tested designs. This limits the ability to relate intention and different types of usability metrics (subjective and objective) to actual interface success. A further limitation in the intention metric was that of a forced switch (p.201), where perhaps the more traditional intent to use (in addition to or instead of paper) question may have resulted in different scores and conclusions.

In defence of the forced switch, this measure of intent could be considered more conservative than an intention to use as it requires both the intention to use the new service and to adopt it instead of the current method. In addition, in practical and commercial settings this could extend to the uptake of e.g. a new technology for securing Internet transactions, or the use of a new eCommerce Website rather than a competitor or a physical store. In the real world both intention to use *in addition to* current method and intention to *switch from* the current method could be considered as important. Therefore there is still

scope for further research on different methods to measure intentions to use in extension to this work.

6.9.2. Outcome and Structure of the eBanking Usability & Utility Questionnaire

The structure of the usability attributes included in the eStatements usability questionnaire is proposed as shown in Figure 6.16. These usability attributes relate to account information retrieval tasks within an eBanking service. The reliability scores relate to the full N=178 cohort using the Advanced Search design [AS] and a combined value for the mean score for all three interfaces [ALL]. There were two miscellaneous attributes which did not correlate well with any other group, *text size* and printout *authenticity*. See Appendix G for the full annotated matrix and the examination of sub-group correlations.

In the experiment using eStatements, some key interaction components were all highly correlated and therefore were grouped under the general term of 'interaction'. The interaction component was made up of some IA, structure and navigation aspects but also included some affect aspects (stress, fluster) as well as some cognitive aspects of concentration and understanding. Interaction components had high associations with intent to use. This group accounts for 12.4% of the variance of the usage intent score for the Advanced search.

Another group of highly associated attributes was classed as 'fulfilment'. This related to several key aspects of success, pleasure and performance characteristics. Fulfilment attributes represent a combination of key satisfaction (affect) qualities with the utility components and the intent (*use again*) attribute. This group accounts for 30.0% of the variance of the usage intention score for the Advanced search.

Two smaller groups were concerned with visual attributes – page density, graphic design and improvement needed. This group accounts for only 7.7% of the variance of the usage intent score for the Advanced search. Finally, a small, less related group of credibility and integrity components was apparent – although this group was less clear for the Advanced Search (winning) interface than on the mean of all interfaces. This group accounts for 5.7% of the variance of the usage intent score for the Advanced search. Further reliabilities for other interfaces and for the mean of both search interfaces together are shown in Appendix G.

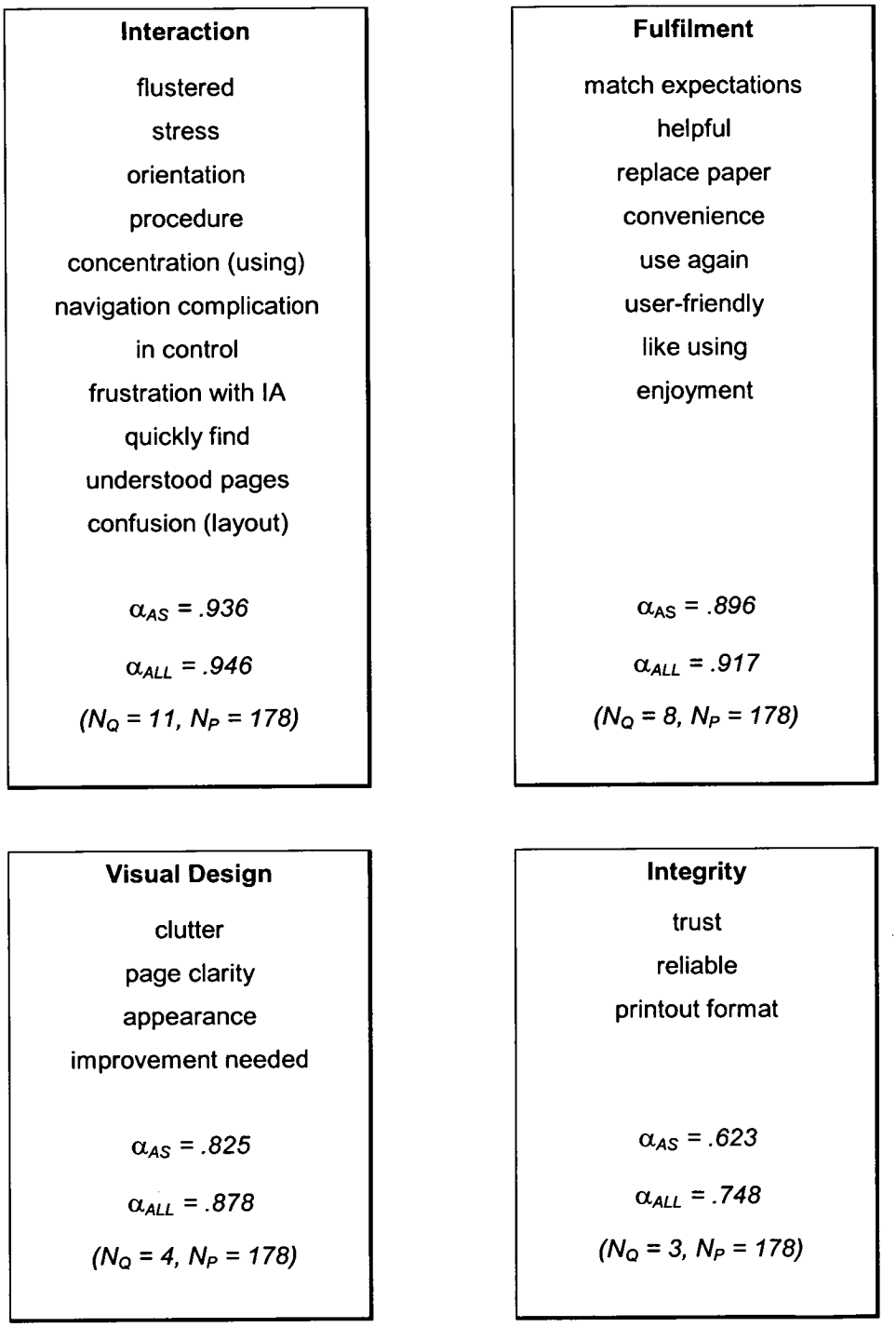


Figure 6.16. eBanking Usability – Subjective Evaluation Components

The model of subjective usability, measured by attitudes towards a broad range of interaction, fulfilment, visual design and integrity factors as they relate to usage intention can be described for the Advanced Search (the most appropriate interface design) as shown in Figure 6.17.

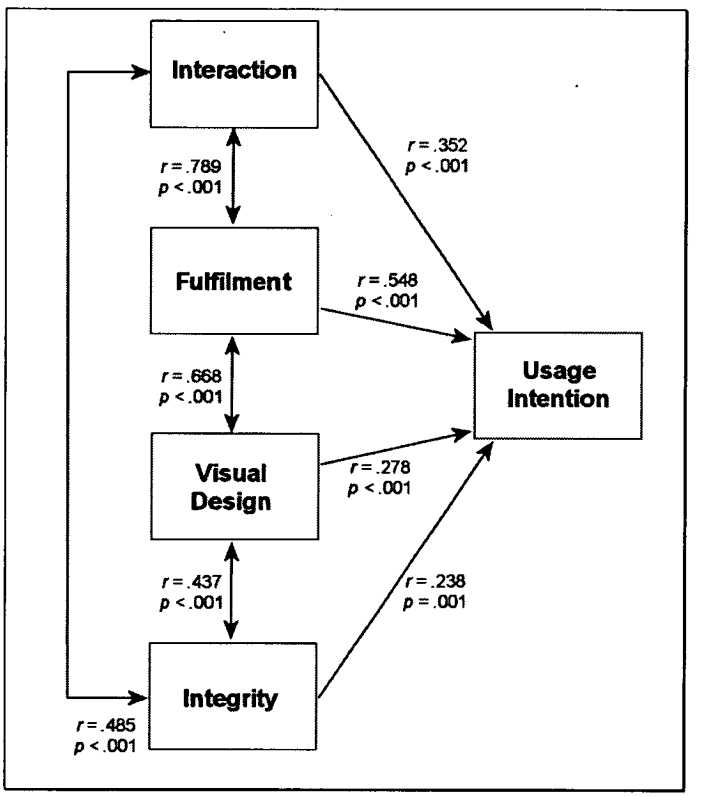


Figure 6.17. Relationships between Usability Attitude and Intention

The full correlation matrix shows very high inter-group correlations for the usability attitude measures and is shown in Table 6.16. There were high, significant correlations between subgroups in the usability questionnaire, and generally high and significant correlations between each subgroup and intentions to switch (both absolute and relative intention scores) compared to objective measures. There were weak but significant correlations between Interaction and Visual Design constructs, and performance, but no corresponding relationships between performance and fulfilment, and integrity constructs. This shows the interplay of structure and interaction with performance in the traditional (ISO) definition of usability. However, in extending usability to consider potential usage – fulfilment and integrity were not affected by performance, but were highly associated with usage intentions.

		Attitudes				Intentions		Performance
		Inter-action	Fulfilment	Visual design	Integrity	BI	ΔBI	
Interaction	*r*	1	.	.	.	.	.	.
	p	-	.	.	.	.	.	.
Fulfilment	*r*	.789	1	.	.	.	.	.
	p	<.001	-	.	.	.	.	.
Visual Design	*r*	.739	.668	1	.	.	.	.
	p	<.001	<.001	-	.	.	.	.
Integrity	*r*	.485	.547	.437	1	.	.	.
	p	<.001	<.001	<.001	-	.	.	.
BI	*r*	.352	.548	.278	.238	1	.	.
	p	<.001	<.001	<.001	.001	-	.	.
ΔBI	*r*	.377	.485	.325	.193	.640	1	.
	p	<.001	<.001	<.001	.010	<.001	-	.
Performance	*r*	.164	.111	.150	.076	.038	.138	1
	p	.029	.140	.046	.311	.618	.067	-

Table 6.16. Relationships between Usability Attitude Subgroups, Performance and Usage Intention.

Fulfilment attributes account for the largest intention variations, but do not typically include any interface or interaction specific details for integrating into an iterative design process. These considerations are present in the interaction and visual design groups.

Integrity factors seemed to be less important in explaining intentions. They are important to include in the eBanking and related financial contexts due to representing key Banking Code requirements.

The overall mean of these individual subjective evaluations offered the ability to differentiate between designs to select the most appropriate and potentially successful option to develop. The mean value of the subjective evaluation of usability accounted for some 15-25% of the variation in intentions to use (from these interfaces). Performance scores were less helpful in determining intent, although aspects of interaction and visual design were weakly yet positively associated with these objective measures of usability

Chapter 7. Discussion & Conclusion

Uncertainty about the relationship between usage intentions and usability, combined with the need to further understand the components and correlations between objective and subjective usability metrics lead to the work presented in the previous chapters.

The research has presented, modified and developed a reliable questionnaire and method for measuring usability subjectively. The method includes the capture of qualitative reports and objective measures of performance. Finally, the usability metrics were compared to comparative measures of preferences in four of the experiments and this was extended to measuring interface-specific usage intentions for the final experiment.

7.1. Summary of Evidence

Table 7.1 summarises the correlations between usability metrics (attitude and performance) and preferences, with page numbers corresponding to the earlier presentation of these results.

	Attitude vs. Preference			Attitude vs. Performance			Performance vs. Preference		
Ranking	-	1st	2nd	-	1st	2nd	-	1st	2nd
Interfaces	Δ	Best A/C	B	Δ	Best A/C	B	Δ	Best A/C	B
Web Portals[a]	.830***	.727***	.685***	.438*	.519**	.255	.237	.377	.150
	p.83	p.84	p.84	p.87	p.87	p.87	p.89	p.89	p.89
Clutter[b]	.676***	.466***	.555***	.150	.503***	-.003	.232	.147	-.069
	p.106	p.106	p.106	p.109	p.109	p.110	p.110	p.110	p.110
Interfaces	Δ	F	S/PL	Δ	F	S/PL	Δ	F	S/PL
Metaphor[c]	.789***	.635***	.470**	.497**	.164	.572**	.500***	.269	.486***
	p.149	p.149	p.149	p.153	p.153	p.153	p.154	p.154	p.154
Dialogue[d]	.616***	.630***	.519**	.299	.328	.315	.315	.379*	.437*
	p.167	p.168	p.168	p.171	p.172	p.172	p.172	p.172	p.172

Table 7.1. Relationships between Usability Attitude, Performance and Preference

Notes:
Letters summarise the experiments – usability (attitude) and preference results, utility rating and performance levels:
[a] A > B, A +ve, B –ve, matched utility, medium task performance
[b] C > B, C +ve, B –ve, matched utility, medium task performance
[c] F > S, F +ve, S +ve, matched utility, high task performance
[d] F = PL, F +ve, PL +ve, matched utility, high task performance
*** $p < .001$; ** $p < .01$; * $p < .05$

In the Web Portal and Design Clutter studies, utility was closely matched, yet task performance was only moderate (70-75%) and did not differentiate between designs. Despite these similarities, attitudes toward usability questionnaire statements indicated significant differences in usability perceptions towards the different designs (48-59%), as did preference (quality) ratings (40-63%). At this level of performance, attitude scores and preference rating were not highly positive. These data refer to Figure 4.8, p.89 and Figure 4.14, p.111.

In the eBanking transaction experiments, utility was also roughly matched and the interfaces achieved high levels of performance (91-97%) which again did not differentiate between the designs. Yet, despite no performance differences, attitude measures were significantly

different for the alternative interaction metaphors (61-68%), but not different dialogues (both 66%). Similarly, preferences were significantly different in for metaphors (56-74%), but not for dialogues (67-68%). These data refer to Figure 5.12, p.155 and Figure 5.16, p.173.

Attitude and preference (quality) scores were higher than those obtained for the Web portals with lower performance results, see Figure 7.1. Similarly, usage intentions were higher (81-100%, see p.147 & p.164).

A summary of the usability (attitude and performance) and intention correlations are shown in Table 7.2.

	Attitude vs. Intention			Attitude vs. Performance			Performance vs. Intention		
Ranking	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**	**1st**	**2nd**	**3rd**
Interfaces	**AS**	**SS**	**DO**	**AS**	**SS**	**DO**	**AS**	**SS**	**DO**
eStatements	.419***	.441***	.367***	.158*	.265***	.033	.038	.148*	.040
	p.225	p.225	p.225	p.229	p.229	p.229	p.231	p.231	p.231
Comparisons	**ΔAS-SS**	**ΔAS-DO**	**ΔSS-DO**	-	-	-	-	-	-
eStatements	.569***	.413***	.592***	-	-	-	-	-	-
	p.333	p.333	p.333	-	-	-	-	-	-

Table 7.2. Relationships between Usability Attitude, Performance and Usage Intention

Notes:
Summary of experiments – usability (attitude), utility rating and performance levels:
AS > SS & DO, AS +ve, SS +ve, DO +ve, increasing utility AS > SS > DO, high performance
*** $p < .001$; ** $p < .01$; * $p < .05$

In the final experiment, three levels of utility were tested: in each case task performance was high (94%-98%) significantly differentiating between the best and worst rated interfaces (although not between the second best and worst). This was mirrored in attitude measures which were high only for the high utility interface (54%-65%). Intentions to use were also different between levels of utility and associated attitude toward usability – with intention only positive on average for the high utility interface (35-55%). These data refer to Figure 6.14, p.232. See Figure 7.2 for an illustration.

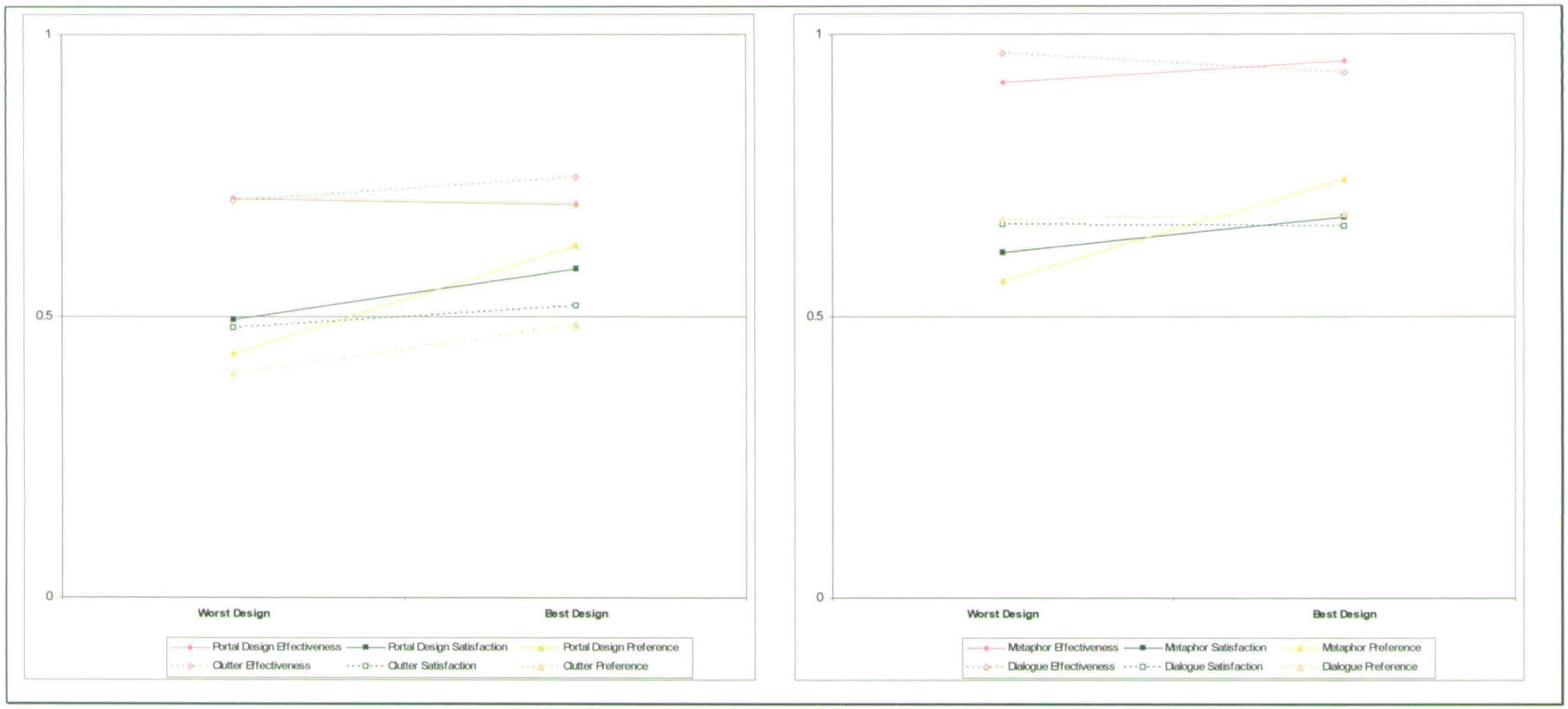


Figure 7.1. Comparison of Usability Metrics: Portal Design, Clutter, Metaphor and Dialogue Experiments

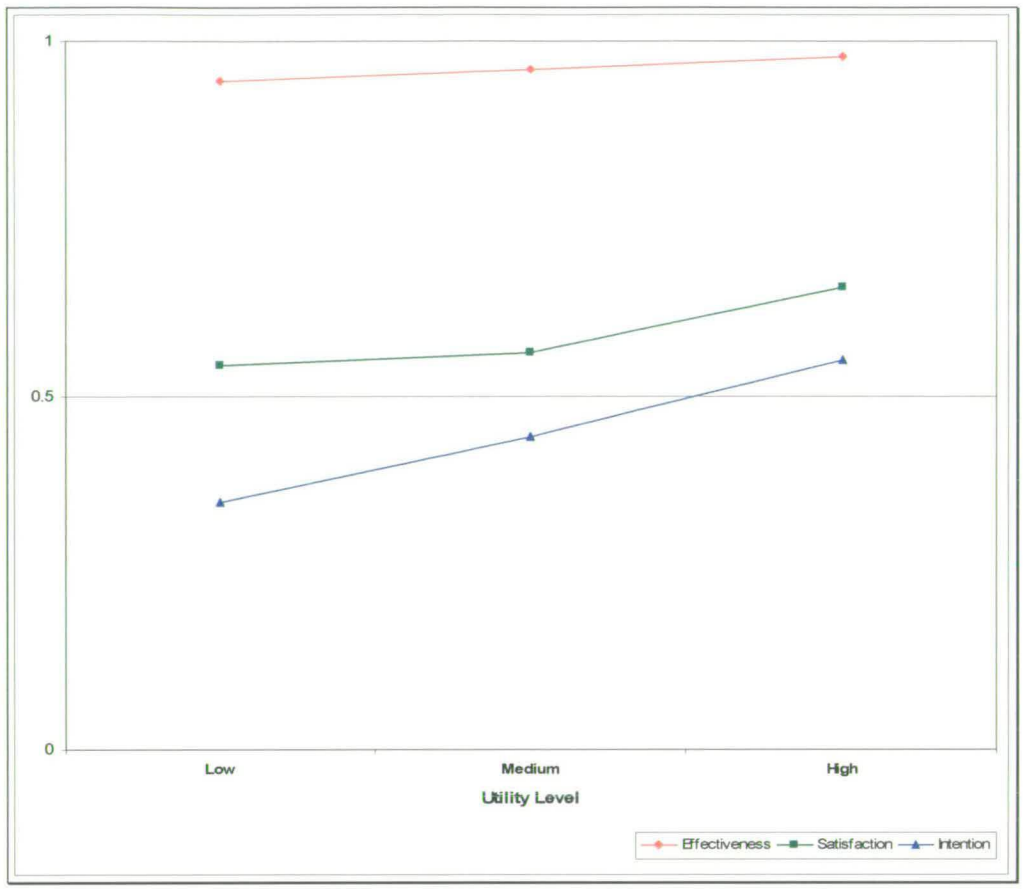


Figure 7.2. Illustration of Usability and Intention Scores as Utility Increases

7.2. Discussion

The P3 model (Dillon & Morris, 1999) that was the starting point for this research proposed that power (or functionality) in the interface, perception (attitudes) toward use of the interface and performance (ability) in completing tasks on the interface would inter-relate, and relate to usage intentions (p.19). However, these proposed relationships have not been empirically validated. There was also little evidence that these subjective and objective components of usability are related, or that they relate to intention. In related work on adoption using the TAM (Davis, 1989), behavioural intent has been empirically related to PU and PEOU in a wide range of tasks and environments, although very rarely with regard to real world usage data. However these TAM constructs do not directly map to most usability definitions. For example PU typically contains perceptions of efficiency,

effectiveness and utility statements, whilst PEOU includes perceptions of satisfaction, efficiency and effectiveness (see Appendix B, p.286-7). In addition they lack the diagnostic qualities required for early prototype usability experiments.

This research has investigated the relationships between overall usability attitudes, attitudes toward usability and related interface characteristics, performance, preference, utility and intent in various Banking information and transaction tasks using the Internet and eBanking. In five individual repeated measures experiments using the proposed metrics, the design with most potential for success was identified, areas for improvement were highlighted and qualitative reports were used to guide recommendations for redesign.

In investigating the relationships proposed by the P3 model, this research assumed that preference would be related to intent and could thus be used as an indirect measure of that construct. This was confirmed in four experiments where preference was strongly positively correlated with attitudes toward reuse. Preference is a valid proxy for intent and allows other relationships in the P3 model to be evaluated.

Considering the direct measure of intent, per interface, as measured in the final experiment, and comparing this to attitudes toward intent constructs (see p.199, 200) in the usability questionnaire (*use again, convenient* and *good replacement*), significantly strong, positive relationships were found. These contribute to validating the new intention metric.

Intent and preference both were suitable in acting as an experimental metric to relate to possible real world use. However the intent scale offered the opportunity to predict potential adoption (and non-adoption) by examining extreme views at either end of the scale. This shows the value of using intent rather than a preference for this metric when adoption rates are a concern, for example, in putting forward a firm case for development.

Performance and preference are also proposed to be related in the P3 model. For Web Portal designs there was a generally weak positive relationship which was somewhat weaker than expected. By contrast in eBanking the relationship was stronger but this may be due to the context as in eBanking the interface is more task-orientated and task completion is critical.

Performance and intent however were weakly positive but not significantly related. The weaker relationship for the eStatement tasks and designs may be because they lacked the critical nature that was noted for transactions. Performance and attitude toward reuse showed a generally weak positive relationship.

Attitudes were related to both preference and intent ratings consistently in a strong positive direction. Increasing levels of utility also caused highly significant usability attitude,

performance and intent differences, all in the positive direction, confirming the proposed model.

Finally, the P3 model also predicts that performance and attitudes should be inter-related. The relationships found were generally positive, supporting this suggestion. If people have good experiences using Websites and Web-based services such as eBanking and eCommerce facilities, then it is likely that they will continue to use them. The usability questionnaire created for the Web and used in this research was able to reliably quantify user experiences as they relate to preference formation and future usage intent.

The research has added strength to the assertion that usability is a good predictor of intention to use and real world usage. It provides empirical evidence that usable and useful designs benefit users and therefore service providers. The empirical validation of the relationship between usability, usage intentions and real world use still needs to be established. In an attempt to demonstrate a secondary use of the intent scale, adoption predictions were compared to real world usage data.

7.2.1. Relationship of Intentions to Real World Usage

In the final evaluation, Usability (attitude) evaluations were highly associated with intentions to adopt and use the technology. Conservative evaluations of usage intent to use gave a realistic guide to real world usage of the production interface of 13% (see Table 6.10, p.218). Although there was higher potential uptake of 54% (see Table 6.11, p.218) in the market using tendencies to be positive or negative in regard to switching behaviour, only 50% of this potential was realised in the final real world design at the outset of implementation where some 23% of accounts had been switched, see p.234). However, the figure from the live (production) search interface corresponded to nearly twice as much as was estimated from a conservative evaluation counting only those with extreme intentions to switch based on an early prototype.

7.3. Implications

7.3.1. Implications for Usability Metrics

In comparison to other findings, usability (attitude) and preference measured in this research were related, .830 – .470 (Table 7.1, p.244), representing some 22 to 70% variance in one score accounted for by the other. This is somewhat lower than the range computed by Agarwal and Venkatesh (2002) between their usability satisfaction instrument and some overall usability perceptions (.71 to .93). In the meta-analysis by Hornbaek and Law (2007) the correlation between satisfaction and preference was .526 to -.036, which is generally much lower than the range found in the experiments presented here.

Attitudes and performance were correlated in the range -.003 – .572 (p.244), representing zero to 33% variance in one score accounted for by the other. This is somewhat lower than the range computed by Sauro and Kindlund (2005a and 2005b) who had mainly strong positive or negative associations between objective and subjective measures (-.33 to .35 and .38 to .45, see p.61). In the meta-analysis by Hornbaek and Law (2007) the correlation between satisfaction and effectiveness metrics was in the range of .401 to .012, which is a smaller range but of similar bounds to the correlations reported here. The suggestion is that there is a generally positive association, but it is very interface dependent. It may also be domain or task dependent and further work would be needed to determine whether the relationship generally does or does not hold.

In terms of performance and preference, the range found was .500 to -.069, representing some zero to 25% variance in one score accounted for by the other. The range also includes the value presented by Nielsen and Levy (1994) in their early meta-analysis (.44) of errors vs. preferences (p.61).

In usability evaluations, there is a tendency to substitute a subjective evaluation of usability for an objective measure. The results found tend to support the need for a wide-ranging subjective evaluation instrument, which includes perceptions about effectiveness and efficiency, and that this can be used on its own to predict potential preferences and should extend to predicting success in the final product. However, the measures of objective usability did hold some importance, particularly in determining what was good enough in terms of performance ability. Utility also had an effect. Further work would be needed to determine whether a threshold of performance and utility is required before which the

proposed model of attitude and usage intention will not hold, certainly it held here for the lowest utility interface.

In terms of subjective measures, interaction and fulfilment components such as information structure, navigation, content quality, affect, utility and relative advantage components accounted for higher variance of preference scores, and these are potential differentiators between alternative interface designs. Although attitude scores did mirror preferences, performance measures did not consistently do so.

The attitude questionnaire was shown to be highly internally reliable using Cronbach's alpha (p.37). This indicates that all the question statements contributed to the questionnaire mean providing a good measure of overall usability by subjective means. In addition, the test-retest consistency was shown in Experiments 1 and 2, with highly comparable, not significantly different scores obtained by two different participant samples, performing different tasks (bank account transactions) but evaluating the same eBanking design.

7.3.2. Implications for Measuring Usage Intention

The range of association between usability metrics and usage intentions found in these experiments is summarised and compared to previous published TAM findings. Usability (attitudes) and usage intention correlated in the range .592 – .367, representing some 13 to 35% variance in one score accounted for by the other (p.245). This is within the range published by Szajna (1996) for PU and PEOU on BI (.29 to .72). It is also consistent with reports from Davis (1989) that PEOU and self-reported usage are strongly correlated (.25 to .59) although the exact constructs are different (see p.64).

Performance related to usage intention in the range .148 – .038, representing only zero to 2% variance in one score accounted for by the other. Again, as performance is typically not studied in TAM studies, there is no real comparison to make with this measure, but considering that PU contains performance and efficiency ratings (perceptions rather than objective measures), these values are more typical of those found by Szajna (1996) of .09 PU to self report usage than Davis who obtained strong positive correlations for PU and self report usage (.56 to .85). Again, this is only a very rough comparison due to the different measures used (see p.63, 64).

In terms of subjective measures, fulfilment components such as affect, utility and relative advantage accounted for much higher variances in usage intentions than the other

components of interaction, visual design and integrity. Yet interaction was also highly significantly correlated with intention at $r > .3$ indicating a medium effect, and visual design and integrity were highly significantly correlated. The inter-group correlations indicate that the structure could have been composed in other ways, for example referring to the full annotated correlation matrix for the Advanced Search in Appendix G, it is clear that affect qualities may also have been placed in the Interaction group, or the Visual design group with equal ease.

The implication from this research is that a broad range of characteristics included in a usability questionnaire can be very helpful in predicting potential success from an early prototype interface design. This goes some way in proving the value of early usability evaluations as the additional resources required are minimal at this stage but should allow the identification of the most potentially successful design before development work commences. A substantial benefit of the questionnaire attributes is that they also provide detailed guidance to problem areas in designs. When combined with qualitative reports and objective measures of usage this is helpful in making recommendations for iterative redesign.

This research implies that at least in the case of eStatements, it would be worth developing a high-functionality interface to increase the likelihood of effecting switching behaviour. In the final interface, similar levels of utility (no usability or intention data available), were associated with almost 50% more switching than predicted by the conservative estimate - however, the final interface did fail to capitalise on the intention tendencies of 54% who were positive in terms of intent with the Advanced Search design from the prototype phase. While there are likely to be some individual reasons for real world behaviour and different timescales within which people will adopt such services (i.e. innovators and laggards), the final interface had high enough utility and usability to be a successful upgrade to the eBanking service.

7.4. Limitations

Limitations of the individual experiments have been discussed in detail in the relevant chapters. In summary, the limited time and one-off exposures to interfaces restricted the depth of study in terms of learning curves and objective efficiency measures. Yet, at least in

the eBanking context, intuitive design was important given the diverse user base and the sporadic nature of usage.

The major limitation of the work is the laboratory setting and the fact that in this setting there is no way to measure real world use. Indeed, the use of early prototypes means that the early usability and final usage data link can never be definitively computed. Usability evaluation is best applied early and therefore large scale studies like those presented in this work will not be comparable to finished production designs..

A final major limitation of the work is that by using the Pearson's correlation to analyse relationships, no cause or effect conclusions can be drawn. Although it may be logical to suggest that the usability objectively measured and attitudes toward interface design caused changes in usage intentions, this can not be confirmed with these designs or analysis. In fact, as attitudes and behaviour were monitored in a hands-on usage session, it is conceivable that usage also influenced perceptions and performance. For example, there is commonly a learning curve associated with performing the same or similar tasks on the same interface. This effect is also apparent in usability evaluations in terms of usability attitude scores and performance measures. These metrics typically increase steeply within the first 3 to 5 experiences. The frequency and scope of each experience with a new technology will therefore influence the perception of its usability, and this in turn may affect future usage intentions. This is the same notion as trialability in diffusion of innovations theory: learning and experience will probably influence future use.

So the model of usability and intent is not causal, but rather cyclic in nature: Usability influences potential usage, which in turn influences future perceptions of usability, and future performance.

The only item which was not apparently involved in the cyclic usability-usage model is the level of functionality, the utility, or fit of task to interface. Functionality of an interface is determined and specified by early task analysis within a usability, design and development lifecycle. There were data to support the conclusion that functionality had to be of a certain standard in order for usability to predict usage, but further studies would be required to confirm what levels of utility and performance are required.

7.5. Further Work

Potential for further work has been discussed throughout the chapters presented. In summary, it would be of interest to explore levels of expertise in further work on eBanking and Web usability, but difficulties in judging experience may hinder this.

Further research should also continue to develop the questionnaire and determine a core set of usability attributes relevant for a wider range of activities online, such as shopping, information gathering and making product comparisons.

A comparison of the three questionnaires proposed and used in the research is shown in Table 7.3. The full questionnaire wordings are included in a similar table in Appendix H.

Several of these usability attributes were consistently working as significant differentiators between designs evaluated in comparison. These related to interaction and fulfilment components of structure and flow, orientation, control, fluster and stress as well as matching expectations. These components and designs which offer controlled comparisons in these dimensions would be of interest to study in terms of usability and usage intention.

Banking Portals (Pilot A & B)	Group v1	eBanking Transactions (Experiments 1 & 2)	Group v2	eStatements (Experiment 3)	Group v3
Navigation complication	Structure	Navigation complication	Structure	Navigation complication	Interaction
Liked using	Affect	Liked using	Affect	Liked using	Fulfilment
In control	Affect	In control	Affect	In control	Interaction
Use again	Quality	Use again	Quality	Use again	Fulfilment
Helpful	Content	Helpful	Content design	Helpful	Fulfilment
Page clarity	Page design	Page clarity	Content design	Page clarity	Visual design
Orientation	Structure	Orientation	Structure	Orientation	Interaction
Flustered	Affect	Flustered	Affect	Flustered	Interaction
Understood pages	Content	Understood pages	Content design	Understood pages	Interaction
User-friendly	Affect	User-friendly	Affect	User-friendly	Fulfilment
Frustration with IA	Structure	Frustration with IA	Structure	Frustration with IA	Interaction
Enjoyment	Affect	Enjoyment	Affect	Enjoyment	Fulfilment
Stress	Affect	Stress	Affect	Stress	Interaction
Reliable	Integrity	Reliable	Misc.	Reliable	Integrity
Quickly find	Structure	Quickly find	Structure	Quickly find	Interaction
Improvement needed	Quality	Improvement needed	Misc.	Improvement needed	Visual design
Procedure	Structure	Procedure	Structure	Procedure	Interaction
Confusion (layout)	Page design	Confusion (layout)	Content design	Confusion (layout)	Interaction
Attractive	Misc.	Attractive	Misc.	Appearance	Visual design
Text size	Content	Text size	Misc.	Text size	Misc.
Concentration (reading)	Content	Concentration (reading)	Content design	*Concentration (using)*	Interaction
Clutter	Page design			Clutter	Visual design
Matched Expectations	Structure			Matched Expectations	Fulfilment
Trust	Integrity			Trust	Integrity

Banking Portals (Pilot A & B)	Group v1	eBanking Transactions (Experiments 1 & 2)	Group v2	eStatements (Experiment 3)	Group v3
Graphics & pictures	Misc.				
Link content	Content				
Link visibility	Page design				
		Change arrangements	Quality		
		Words & phrases	Misc.		
				Replace paper	Fulfilment
				Convenience	Fulfilment
				Suitability Printout	Integrity
				Authenticity	Misc.

Table 7.3. Comparison of Questionnaire Statements

7.6. Summary of Relationships

These relationships are summarised in Figure 7.3.

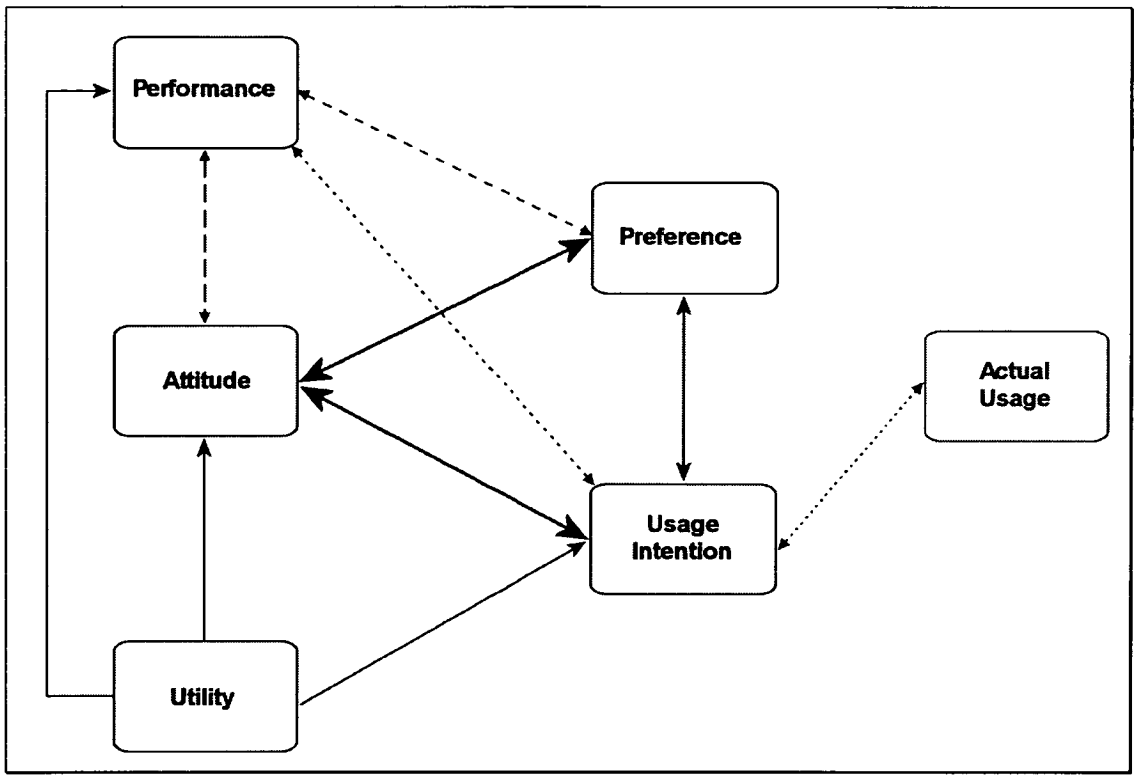


Figure 7.3. Illustration of Positive Relationships and Effects Revealed by Experiments

Notes:
Double arrows indicate correlations; Full lines represent strong positive relationships; Dashed lines represent weak positive relationships.
Directional arrows indicate ANOVA results where utility was manipulated and at higher levels increased performance, attitude and intent levels.

Attitude-Preference

Throughout the first four experiments, for a range of tasks and interfaces for Banking information and transactions in eBanking services, the relationship between usability attitudes and preferences was established to be strong, positive and highly significant.

Attitude-Performance

The relationship between objective and subjective usability metrics (attitudes and performance) was not consistent. There was a strong positive relationship for the preferred Web portal in both cases. However, it was not consistently significant in relation to the differences between interfaces or the least preferred Website. There were also inconsistent results in this regard for eBanking transaction interfaces, with generally positive relationships but at varying strengths and significance. Finally, the relationship was also weakly positive and significant for the higher utility interfaces with search facilities, although not established for the low utility version.

Performance-Preference

The relationship between performance and preference was not consistent. There were weak positive (but not significant) relationships for the Web Portals of banking information. Although when critical eBanking tasks were considered there were stronger, more consistently significant relationships for individual designs and differences.

Attitude-Usage Intention

The relationship between usability attitude and usage intention was established to be strong, positive and highly significant for information search using eStatements in eBanking services.

Performance-Usage Intention

The relationship between usability performance and usage intention was also not consistent, they were mainly very weak, positive and not significant.

Intention-Actual Usage

The correspondence between the intention predictions and actual usage of the production interface showed that actual usage was well within the range predicted, but closer to conservative estimates using extreme scores than for generally positive intention tendencies.

7.7. Conclusion

Usage intentions were higher for high utility interfaces, high attitudes and high performance levels. The usability questionnaire when augmented by relative advantage (convenience and replacement) components, and considered a broad range of usability issues and interface characteristics provided a suitable metric related to potential uptake success. Moreover, it was identifying areas where usability might be improved in the design. The metric was suitable to be used at an early stage of the development process, when objective measures such as time and steps to completion may be less informative due to learning effects.

The P3 model proposed by Dillon and Morris (1998) is the basis for the thesis that usability evaluation metrics will associate with usage intentions, and by extension preferences. The results were used to develop and refine a modern definition of usability as it relates to intent to use, and by extension success. The resulting definition of usability is:

> ***Usability is the utility of the system, the users' performance during operation and their attitudes towards a range of interaction, fulfilment, visual design and integrity components during use of the system.***

In this definition, subjective metrics of usability, as measured by the attitude questionnaire designed for the research, were positively associated with usage intentions in an eBanking service context. The components of usability as they relate to probable success in the marketplace were dominated by fulfilment constructs including *convenience*, *good replacement*, *used again* and affect qualities of *like using* and *enjoyment*. The subjective metrics also positively associated with preferences between designs measured on a comparative scale, indicating that preferences can act as a reasonable proxy measure of intent at least in an eBanking context.

The usability model presented here of Web banking success could be extended to applications in eCommerce transactions, information retrieval on a variety of platforms and for different service providers such as business, banking and government. It may also extend to eLearning and other applications.

It is not sufficient to ask whether users thought they would want to use a design again as an attitude statement in usability research, as this statement alone will not necessarily differentiate between designs. However, a broad range of usability related statements are able to make this distinction. Similarly, preferences are able to differentiate between designs but won't help focus redesign efforts. The metrics presented in this thesis identify problem areas in designs while the qualitative reports collected in parallel help formulate solutions.

The empirical data presented in this research serve to support the thesis that attitude measurement of usability which includes components of interaction, fulfilment, visual design and integrity relate to preference formation and usage intention in performing online information retrieval and eBanking tasks.

References

Agarwal, R. and Venkatesh, V., 2002, Assessing a Firm's Web Presence: A Heuristic Evaluation Procedure for the Measurement of Usability, Information Systems Research, Volume 13, Issue 2, pp.168-186.

Aladwani, A. M., 2001, Online banking: a field study of drivers, development challenges, and expectations, International Journal of Information Management, Volume 21, Issue 3, pp. 213-225.

Alty, J. L., Knott, R. P., Anderson, B. and Smyth, M., 2000, A framework for engineering metaphor at the user interface, Interacting with Computers, Volume 13, Issue 2, pp. 301-322.

Anderson, N.H., 1961, Scales and Statistics: Parametric and Non-parametric, Psychological Bulletin, pp. 305-316

Baeker, R.M., Grudin, J., Buxton, W.A.S., and Greenberg, S., 1995, Readings in Human-Computer Interaction: Toward the Year 2000, Second Edition, Morgan Kaufmann Publishers Inc., USA.

Bailey, R.W., 1993, Performance vs. preference, in Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting (HFES), pp. 282-285.

Baroudi, J.J., Olson, M.H., and Ives, B., 1986, An empirical study of the impact of user involvement on system usage and information satisfaction, Communications of the ACM, Volume 29, Issue 3, pp. 232-238.

BBC News, 2004, Changing banks 'still a rarity', Published 18/06/2004, © BBC MMIV, available online: http://news.bbc.co.uk/1/hi/business/3818151.stm, accessed on 03/05/2007.

BBC News, 2005, Reluctance to switch banks fades, Published 28/11/2005, © BBC MMV, available online: http://news.bbc.co.uk/1/hi/business/4385842.stm, accessed on 03/05/2007.

BBC News, 2006, Barclays bans jargon in makeover, Published 03/02/2006, © BBC MMVI, available online: http://news.bbc.co.uk/1/hi/business/4674548.stm, accessed on 03/05/2007.

Bernard, M. L., Chaparro, B. S., Mills, M. M., & Halcomb, C. G., 2003, Comparing the effects of text size and format on the readibility of computer-displayed Times New Roman and Arial text, International Journal of Human-Computer Studies, Volume 59, Issue 6, pp. 823-835.

Bevan, N., 1995, Measuring usability as quality of use, Software Quality Journal, Volume 4, pp. 115-150

Blakeslee, A. F., 1914, Corn and Men: The Interacting Influence of Heredity and Environment—Movements for Betterment of Men, or Corn, or Any Other Living Thing, One-sided Unless They Take Both Factors into Account, Journal of Heredity, Volume 5, Number 11, pp. 511-518.

British Bankers' Association, 2005, The Banking Code, BBA Enterprises Ltd.

Boran, M.T., Ramey, J., 2000, Thinking aloud: reconciling theory and practice, IEEE Transactions on Professional Communication, Volume 43, Number 3, pp 261-78.

Bryman, A. and Cramer, D, 2001, Quantitative Data Analysis with SPSS Release 10 for Windows - A Guide for Social Scientists, Routledge.

Burgess, R. G., 1993, Research Methods, Nelson.

Canny, J., 2006, The Future of Human Computer Interaction, ACM Queue, Volume 4, Number 6, pp.24-32.

Carroll, J.M., Mack, R.L. and Kellogg, W.A., 1988, Interface metaphors and user interface design, In: Helander, M., Editor, 1988, Handbook of Human–Computer Interaction, North-Holland, Amsterdam, pp. 45–65.

CCIR, 1992, Intelligent Dialogues for Multimedia Services - Dialogues for Systems Report on Experiment 4, Commercial in Confidence Research Report submitted to BT.

Ceaparu, I., Lazar, J., Bessiere, K., Robinson, J., Shneiderman, B., 2004, Determining Causes and Severity of End-User Frustration, International Journal of Human-Computer Interaction, Volume 17, Issue 3, September, pp. 333 – 356.

Centeno, C., 2004, Adoption of Internet services in the Acceding and Candidate Countries, lessons from the Internet banking case, Telematics and Informatics, Volume 21, Issue 4, pp.293-315.

Chapanis, A., 1991, The business case for human factors in Informatics, Chapter 3 in: Human Factors for Informatics Usability, Shackel, B., and Richardson, S. J., Cambridge University Press, pp. 39 - 71.

Chau, P.Y., 1996, An empirical assessment of a modified technology acceptance model. Journal of Management Information Systems, Volume 13, Number 2, pp. 185-204.

Chin, J. P., Diehl, V. A., and Norman, K. L., 1988, Development of an instrument measuring user satisfaction of the human-computer interface, Proceedings of the ACM Human Factors in Computing Systems Conference (CHI 88), ACM Press, New York, NY, pp. 213-218.

Chuan-Chuan Lin, J. and Lu, H., 2000, Towards an understanding of the behavioural intention to use a Web site, International Journal of Information Management, Volume 20, pp. 197-208.

Cochran, W. G. and Cox, G. M., 1957, Experimental design, Wiley.

Colman, A. M., 2001, A Dictionary of Psychology, Oxford University Press, Oxford.

Coolican, H., 1990, Research Methods and Statistics in Psychology, Hodder & Stoughton, GB.

Cooper, A., 1999, The inmates are running the asylum, Indianapolis.

Cox, D. R., 1958, Planning of experiments, Wiley.

Dandapani, K., 2004, Success and Failure in Web-based Financial Services, Communications of the ACM, Volume 47, Number 5, pp. 31-33.

Davies, M., 1996, Image problems with financial services: some considerations for improvement, Management Decision, Volume 34, Issue 2, pp. 64-71

Davis, F.D., 1989, Perceived usefulness, perceived ease of use and user acceptance of information technology, MIS Quarterly, Volume 13, Number 3, pp. 319-340.

Davis, F.D., Bagozzi, R.P. and Warshaw, P.R., 1989, User Acceptance of Computer Technology: A Comparison of Two Theoretical Models, Management Science, Volume 35, Number 8, pp. 982-1003.

Davis, F.D., Bagozzi, R.P., and Warshaw, P.R., 1992, Extrinsic and intrinsic motivation to use computers in the workplace. Journal of Applied Social Psychology, Volume 22 (1992), pp. 1111–1132.

Davis, F.D., 1993, User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. International Journal of Man-Machine Studies, Volume 38, Number 3, pp. 475-487.

Davis, F.D. and Venkatesh, V., 2004, Toward preprototype user acceptance testing of new information systems: implications for software project management, IEEE Transactions on Engineering Management, Volume 51, Number 1, pp.31-46.

Delone, W.H. and McLean, E.R., 2003, The DeLone and McLean Model of Information Systems Success: A Ten-Year Update, Journal of Management of Information Systems, Volume 19, Issue 4, April, pp. 9-30.

Dillon, A. and Morris, M.G., 1996, User acceptance of new information technology: theories and models, in Williams, M. E., Eds. Annual Review of Information Science and Technology, Chapter 31, pp. 3-32.

Dillon, A. and Morris, M.G., 1998, From "can they" to "will they?": Extending usability evaluation to address acceptance. In Hoadley, E.D. and Benbasat, I., Eds. Proceedings Association for Information Systems Conference, Baltimore, MD.

Dillon, A. and Morris, M.G., 1999, Power, Perception and Performance: From Usability Engineering to Techonology Acceptance with the P3 Model of User Response, Proceedings of 43rd Annual Conference of the Human Factors and Ergonomics Society, Santa Monica, CA, pp. 1017-1021.

Dillon, A., and Gushrowski, B., 2000, Is the home page the first digital genre?, Journal of the American Society for Information Science, Volume 51, Number 2, pp. 202-205.

Dillon, A., 2002, Beyond usability: process, outcome and affect in human-computer interactions. Canadian Journal of Library and Information Science, Volume 26, Number 4, pp. 57-69.

Dishaw, M.T. and Strong, D.M., 1999, Extending the technology acceptance model with task-technology fit constructs, Information & Management, Volume 36, Issue 1, pp. 9-21.

Dix, A., Finlay, J., Abowd, G.D. and Beale, R., 2004, Human-Computer Interaction, 3rd Edition, Prentice-Hall, Pearson.

Dumas, J. S., and Redish, J. C., 1993, A practical guide to usability testing, Norwood, NJ: Ablex Publishing.

Dutton, R.T., Foster, J.C., Jack, M.A., and Stentiford, F.W.M., 1993, Identifying usability attributes of automated telephone services, Proceedings of the European Conference on Speech Communication and Technology (Eurospeech 93), pp. 1335-1338.

Dutton, R.T., Foster, J.C., and Jack M.A., 1999, Please mind the doors – do interface metaphors improve the usability of voice response services?, BT Technology Journal, Volume 17, Number 1, pp.172-177.

Edwards A.L. and Kenney, K.C, 1967, A comparison of the Thurstone and Likert techniques of attitude scale construction, In. M. Fishbein (Ed.), Readings in attitude theory and measurement, New York, Wiley, pp. 249–256.

Elling, S. Lentz, L. and de Jong, 2007, Website Evaluation Questionnaire: Development of a Research-Based Tool for Evaluating Informational Websites, Lecture Notes in Computer Science, Volume 4656/2007, Springer Berlin / Heidelberg, pp. 293-304.

Erickson, 1995, Working with Interface Metaphors, in Baecker, R. M., 1995, Readings in human-computer interaction: toward the year 2000. San Francisco, California, Morgan Kaufmann Publishers, pp. 147-151.

Faulkner, L., 2003, Beyond the five-user assumption: Benefits of increased sample sizes in usability testing, Behavior Research Methods, Instruments and Computers, Volume 35, Number 3, pp. 379-383.

Faulkner, X., 2000, Usability Engineering, Macmillan Press Ltd, London, UK.

Field, A., 2005, Discovering Statistics Using SPSS, 2nd Edition, Sage Publications, GB.

Fishbein, M. (Ed), 1967, Readings in attitude theory and measurement, John Wiley, London.

Fishbein, M and Ajzen, I., 1985, Belief, Attitude, Intention and Behaviour: An Introduction to Theory and Research. Reading, MA: Addison-Wesley.

Fisher, R. A., 1971, The design of experiments, Statistical methods, experimental design, and scientific inference, 9th Edition, Hafner, New York, NY.

Frøkjær, E., Hertzum, M., and Hornbæk, K., 2000, Measuring Usability: Are Effectiveness, Efficiency and Satisfaction Really Correlated?, CHI Letters, Volume 2, Issue 1, pp. 345-352.

Furnell, S., 2004, E-commerce security: a question of trust, Computer Fraud & Security, 2004, Issue 10, pp. 10-14.

Gefen, D., Karahanna, E., Straub, D.W., 2003, Trust and TAM in online shopping: an integrated model, MIS Quarterly, Volume27, Issue 1, pp. 51-90.

Goodhue, D.L. and Thompson, R.L., 1995, Task-Technology Fit and Individual Performance, MIS Quarterly, Volume19, Number 2, pp. 213-236.

Gopalakrisnan, S., Wischnevsky, J.D. and Damanpour, F., 2003, A Multilevel Analysis of Factors Influencing the Adoption of Internet Banking, IEEE Transactions on Engineering Management, Volume 50, Number 4, pp. 413-426.

Gould, J.D., 1988, How to Design Usable Systems, in Helander, M. (Ed.), Handbook of Human Computer Interaction, Amsterdam: NorthHolland, pp. 757-789.

Gould, J.D., 1995, How to Design Usable Systems, in Baeker, R.M., J. Grudin, W.A.S. Buxton, and S. Greenberg, Readings in Human-Computer Interaction: Toward the Year 2000, 2nd Edition, Morgan Kaufmann Publishers Inc., USA, pp. 93-121.

Grabner-Kräuter, S., and Kaluscha, E. A., 2003, Empirical research in on-line trust: a review and critical assessment, International Journal of Human-Computer Studies, Volume 58, Issue 6, pp. 783-812.

Graeber, C., Shevlin, R. and Sweeney, J., 2003, Canadian Customers Are Ready For eStatements, Forrester Research.

Guildford, J.P., 1967, Response bias and response sets, Chapter 30, In. M. Fishbein (Ed.), Readings in attitude theory and measurement, New York, Wiley, pp 277-281.

Guttman, L., and Suchman, E.A., 1967, Intensity and a zero point for attitude analysis, Chapter 29, In. M. Fishbein (Ed.), Readings in attitude theory and measurement, New York, Wiley, pp. 267-276.

Hartson, R., 1998, Human-Computer Interaction: Interdisciplinary Roots and Trends, Journal of Systems and Software, Volume 43, pp. 103-118.

Helander, M. (Ed), 1988, Handbook of human-computer interaction, Amsterdam: North-Holland.

Helander M.G. and Khalid, H.M., 2000, Modeling the customer in electronic commerce, Applied Ergonomics, Volume 31, Issue 6, pp. 609-619

Hoaglin, D.C., Mosteller, F. & Tukey, J.W., 1983. Understanding robust and exploratory data analysis, Wiley, New York.

Holland, C.P. and Westwood, J.B., 2001, Product-market and technology strategies in banking, Communications of the ACM, Volume 44, Number 6, pp. 53-57.

Hornbæk, K., 2006, Current practice in measuring usability: Challenges to usability studies and research, International Journal of Human-Computer Studies, Volume 64, Issue 2, pp. 79-102.

Hornbæk, K. and Law, E.L., 2007, Meta-Analysis of Correlations Among Usability Measures, Proceedings of CHI 2007, ACM Press, pp. 617-626.

Horton R.P., Buck T., Waterson, P.E., Clegg C.W., 2001, Explaining intranet use with the technology acceptance model, Journal of Information Technology, Volume 16, December, pp. 237-249.

Howell, D.C., 1997, Statistical methods for psychology, Duxbury Press, Belmont CA.

Howell, W.C., 1985, Engineering Psychology, Chapter 10 in Altmaier, E.M. and Meyer, M.E. (Eds.), Applied Specialties in Psychology, Random House, Lawrence Erlbaum Associates, NJ.

Hudson, W., 2001, How Many Users Does it Take to Change a Web Site?, SIGCHI Bulletin, Volume 33, Number 3, pp. 6.

Hudson, W., 2002, The Lost World of E-Banking, SIGCHI Bulletin, Volume 34, Number 5, (September/October 2002), pp. 7.

Igbaria, M., Schiffman, S.J., & Wieckowski, T.J., 1994, The respective roles of perceived usefulness and perceived fun in the acceptance of microcomputer technology, Behavior and Information Technology, Volume 13, pp. 349–361.

Igbaria, M., Iivari, J. and Maragahh, H., 1995, Why do individuals use computer technology? A Finnish case study, Information & Management, Volume 29, Issue 5, pp. 227-238.

Igbaria, M. and Tan, M., 1997, The consequences of information technology acceptance on subsequent individual performance, Information and Management, Volume 32, Issue 3, pp. 113-121.

Internal communication, 2005, from contacts at Lloyds TSB.

International Organization for Standardization, ISO 9241-11, 1998, Ergonomic requirements for office work with visual display terminals (VDTs) Part II: Guidance on Usability.

Ives, B., Olson, M. H., and Baroudi, J. L., 1983, The measurement of user information satisfaction, Communications of the ACM, Volume 26, Issue 10, pp. 785-793.

Jack, M.A., Foster, J.C., and Stentiford, F.W.M., 1993, Usability analysis of intelligent dialogues for automated telephone services, Proceedings of the Joint ESCA/NATO workshop on Applications of Speech Technology, pp. 149-152.

Jamal, A. and Naser, K., 2002, Customer satisfaction and retail banking: an assessment of some of the key antecedents of customer satisfaction in retail banking, The International Journal of Bank Marketing, Volume 20, Number 4, pp. 146-160.

Jayawardhena, C., and Foley, P., 2000, Changes in the banking sector – the case of Internet Banking in the UK, J. Internet Research: Networking and Policy, Volume 10, Number 1, pp. 19-30.

Jeyaraj, A.; Rottman, J.W.; Lacity, M.C., 2006, A review of the predictors, linkages, and biases in IT innovation adoption research, Journal of Information Technology, Volume 21, Number 1, Palgrave Macmillan, pp. 1-23.

Kalback, J. and Bosenick, T., 2003, Web Page Layout: A Comparison Between Left- and Right-justified Site Navigation Menus, Journal of Digital Information, Volume 4, Issue 1, Article No. 153, 2003-04-28, available online at: http://jodi.tamu.edu/Articles/v04/i01/Kalbach/ (accessed 16th February 2007).

Karat, J., 1988, Software Evaluation Methodologies, in Helander, M. (Ed.), Handbook of Human Computer Interaction, Amsterdam: NorthHolland, pp. 891-903.

Karat, J., 1997, Evolving the scope of user-centered design, Communications of the ACM, Volume 40, Issue 7, pp. 33-38.

Kerlinger, F., 1973, Foundations of Behavioural Research, Second Edition.

Kim, J., and Moon, J. Y., 1998, Designing towards emotional usability in customer interfaces – trustworthiness of cyber-banking system interfaces, Interacting with Computers, Volume 10, Issuel, pp. 1–29.

Kim, K. and Prabhakar, B., 2000, Initial trust, perceived risk, and the adoption of internet banking, In Proceedings of the Twenty First international Conference on information Systems, Association for Information Systems, pp. 537-543.

King, W.R. and He, J., 2006, A meta-analysis of the technology acceptance model, Information & Management, Volume 43, Issue 6, pp. 740-755.

Kirakowski, J., 1994, The use of questionnaire methods for usability assessment, available online at: http://sumi.ucc.ie/sumipapp.html (accessed 22nd Feb 2005).

Kline, 2004, Beyond Significance Testing: Reforming Data Analysis Methods in Behavioural Research, APA.

Kuniavsky, M., 2003, Observing the User Experience: A Practitioner's Guide to User Research, Morgan Kaufmann.

Labovitz, S., 1967, Some observations on measurement and statistics, Social Forces, Volume 46, pp. 151-160.

Lai, V. S., and Li, H., 2005, Technology acceptance model for internet banking: an invariance analysis, Information and Management, Volume 42, Issue 2, pp. 373-386.

Landauer, T.K., 1988, Research Methods in Human Computer Interaction, in Helander, M. (Ed.), Handbook of Human Computer Interaction, Amsterdam: NorthHolland, pp. 905-928.

Lane, D.M., 1999, Hyperstat, Second Edition, Atomic Dog Publishing.

LeCompte, D., 1999, Seven, plus or minus two, is too much to bear: Three (or fewer) is the real magic number, Proceedings of the Human Factors and Ergonomics Society, pp. 289-292.

Lee, Y. and Kim, J., 2002, From design features to financial performance: a comprehensive model of design principles for online stock trading sites, Journal of Electronic Commerce Research, Volume 3, Number 3, pp. 128-143.

Legris, P., Ingham, J. and Collerette, P., 2003, Why do people use Information Technology? A critical review of the technology acceptance model, Information and Management, Volume 40, pp. 191-204.

Lewis, J. R., 2002, Psychometric evaluation of the PSSUQ using data from five years of usability studies, International Journal of Human-Computer Interaction, Volume 14, Issues 3-4, pp. 463-488.

Lewis, J., and Sauro, J., 2006, When 100% really isn't 100%: Improving the accuracy of small-sample estimates of completion rates, Journal of Usability Studies, Volume 1, Number 3, pp. 136-150.

Liao, Z., and Cheung, M.T., 2002, Internet based e-banking and consumer attitudes: an empirical study, Information and Management, Volume 39, pp. 283-295.

Lichtenstein, S. and Williamson,, K., 2006, Understanding Consumer Adoption of Internet Banking: An Interpretive Study in the Australian Banking Context, Journal of Electronic Commerce Research, Volume 7, Number 2, pp. 50-66.

Lidwell, W., Holden, K., and Butler, J., 2003, Universal Principles of Design, Rockport Publishers, Inc.

Likert, R., 1932., A technique for the measurement of attitudes, Archives of Psychology, Volume 140, pp. 5-55.

Likert, R., 1967, A method of constructing an attitude scale, Chapter 11, In. M. Fishbein (Ed.), Readings in attitude theory and measurement, New York, Wiley, pp. 90-95.

Lindgaard, G., Fernandes, G., Dudek, C. & Brown, J., 2006, Attention web design-ers: You have 50 milliseconds to make a good first impression!, Behaviour & Information Technology, Volume 25, pp. 115–126.

Ling, J. & van Schaik, P., 2002, The effect of text and background colour on visual search of Web pages, Displays, Volume 23, Number 5, pp. 223-230.

Ling, J. & Schaik, P.V., 2006, The influence of font type and line length on visual search and information retrieval in web pages, International Journal of Human-Computer Studies, Volume 64, Number 5, pp. 395-404.

Love, S., Dutton, R.T., Foster, J.C., Jack, M.A., Nairn, I.A., Vergeynst, N.A., and Stentiford, F.W.M., 1992, Towards a usability measure for automated telephone services, Proceedings of the Institute of Acoustics, Speech and Hearing Workshop, Volume 14, Number 6, pp. 553-559.

Love, S., Dutton, R.T., Foster, J.C., Jack, M.A., and Stentiford, F.W.M., 1994, Identifying salient usability attributes for automated telephone services, Proceedings of the International Conference on Spoken Language Processing (ICSLP-94), pp.1307-1310.

Love, S., 1997, The Role of Individual Differences in Dialogue Engineering for Automated Telephone Services, PhD thesis, University of Edinburgh.

Marcus, A., 1995, Principles of effective visual communication for graphical user interface Design, in Baeker, R.M., J. Grudin, W.A.S. Buxton, and S. Greenberg, Readings in Human-Computer Interaction: Toward the Year 2000, Second Edition, Morgan Kaufmann Publishers Inc., USA, pp.

Marshall, D., Foster J.C., and Jack, M.A., 2001, User performance and attitude towards schemes for alphanumeric data entry using restricted input devices, Behaviour and Information Technology, Volume 20, Number 3, pp.167-188.

Mathieson, K.1991. Predicting user intentions: Comparing the technology acceptance model with the theory of planned behaviour. Information Systems Research, pp. 173-191.

Mattila, M., Karjaluoto, H., Pento, T., 2003, Internet banking adoption among mature customers: early majority or laggards?, Journal of Services Marketing, Volume 17, Number 5, pp. 514-528.

McBreen, H.M., Shade, P., Jack M.A. and Wyard, P.J., 2000, Experimental Assessment of the Effectiveness of Synthetic Personae for Multi-Modal E-Retail Applications, Proceedings of the Fourth International Conference on Autonomous Agents,pp.39-45.

McGrath, J.E., 1995, Methodology Matters: Doing Research in the Behavioural and Social Sciences, in Baeker, R.M., J. Grudin, W.A.S. Buxton, and S. Greenberg, Readings in Human-Computer Interaction: Toward the Year 2000, Second Edition, Morgan Kaufmann Publishers Inc., USA, pp. 152-169.

McInnes, F., 2001, Lies, damned lies...Principles and pitfalls of statistics, Notes for CCIR Discussion Group.

McInnes, F., 2005, Statistical Analysis of Experiment Data, Chapter 9 in The Principles and Practice of Usability Engineering, CCIR Workshop Notes, pp.86-104.

Molich, R., Ede, M.R., Kaasgaard, K. and Karyukin, B., 2004, Comparative usability evaluation, Behaviour & Information Technology, Volume 23, Number 1, pp. 65-74.

Moore, D. S., 2001, Statistics: concepts and controversies, 5th edition, WH Freeman and Co, NY.

Morkes, J. and Nielsen, J., 1988, Applying Writing Guidelines to Web Pages, Conference on Human Factors in Computing Systems (CHI 88), ACM Press, pp. 321-322.

Nah, F.F-H. and Davis, S., 2002, HCI Research Issues in Electronic Commerce, Journal of Electronic Commerce Research, Volume 3, Number 3, pp. 98-113.

National Statistics, 2006, First Release: Internet Access, Households and Individuals, Issued by National Statistics, 1 Drummond Gate, London (23rd August), available online at: www.statistics.gov.uk/pdfdir/inta0806.pdf, accessed on 20/07/2007.

Nelson, B.C. and Smith, T., 1990, User interaction with maintenance information: a performance analysis of hypertext versus hard copy formats, In: Proceedings of the Human Factors and Ergonomics Society 34th Annual Meeting (HFES), pp. 229–233.

Nichols, T.A., Rogers, W.A., Fisk, A.D. and West, L.D., 2001, How old are your participants? An investigation of age classifications as reported in human factors, Proceedings of the Human Factors and Ergonomics Society 45th Annual Meeting, pp. 260-261.

Nielsen, J., 1993a, Usability Engineering, Academic Press, USA.

Nielsen, J., 1993b, Iterative User-Interface Design, IEEE Computer , Volume 26, Number 11, pp. 32-41.

Nielsen, J., and Levy, J., 1994, Measuring usability: preference vs. performance, Communications of the ACM, Volume 37, Number 4, pp. 66-75.

Nielsen, J., 1996, Top Ten Mistakes in Web Design, published on Jakob Nielsen's Alertbox, available online at: http://www.useit.com/alertbox/9605.html (accessed 11th June 2007).

Nielsen, J., 2000, Designing Web Usability: The Practice of Simplicity, New Riders Publishing, Indianapolis.

Nilsson, M., Adams, A., and Herd, S., 2005, Building security and trust in online banking, In Extended Abstracts on Human Factors in Computing Systems (CHI '05), ACM Press, New York, NY, pp. 1701-1704.

Norman, D., 1988, The Design of Everyday Things, Basic Books, USA.

Nunnally, J.C., 1978, Psychometric theory, McGraw-Hill.

O'Brien Holt, P., 2006, "The Human Factor" Human Factors Engineering and Artefacts, Interactive Systems Research Group, School of Computing, Robert Gordon University.

Oppenheim, A.N., 1992, Questionnaire Design, Interviewing and Attitude Measurement (New Edition), Pinter, London.

Pikkarainen, T., Pikkarainen, K., Karjaluoto, H., Pahnila, S., 2004, Consumer acceptance of online banking: an extension of the technology acceptance model, Internet Research: Electronic Networking Applications and Policy, Volume 14, Number 3, pp.224-35.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., & Carey, T., 1994, Human-computer interaction, Wokingham, UK: Addison-Wesley.

Preece, J., Rogers, Y. and Sharp, H., 2002, Interaction Design: Beyond Human-Computer Interaction, John Wiley & Sons Inc., NY.

Premkumar, G. and Bhattacherjee, A., 2008, Explaining information technology usage: A test of competing models, Omega, Volume 36, Issue 1, pp. 64-75.

Quesenbery, W., 2002, Using the 5Es to understand users, available online at: http://www.wqusability.com/articles/getting-started.html, accessed: 12th Feb 2005.

Raskin, J., 2000, The Humane Interface, ACM Press.

Roberts, P. and Henderson, R., 2000, Information technology acceptance in a sample of government employees: a test of the technology acceptance model, Interacting with Computers, Volume 12, Issue 5, pp. 427-443.

Robson, C., 1983, Experiment, Design and Statistics in Psychology: An Introduction, Pelican Books, GB.

Rogers, E., 1993, Diffusion of Innovations, 2nd Edition.

Root, R. W. & Draper, S., 1983, Questionnaires as a software evaluation tool, Proceedings of the ACM Conference on Human Factors and Computer Systems (CHI 83), New York, ACM Press, pp. 83-87.

Rosenfeld, L. and Morville, P., 2002, Information Architecture for the World Wide Web, Second Edition, O'Reilly.

Sarel, D. and Marmorstein, H., 2003, Marketing online banking services: The voice of the customer, Journal of Financial Services Marketing, Volume 8, Number 2, pp. 106-118, Palgrave Macmillan.

Sauro, J. and Kindlund, E., 2005a, A method to standardize usability metrics into a single score, In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '05), ACM, New York, NY, 401-409.

Sauro, J. and Kindlund, E., 2005b, Using a Single Usability Metric (SUM) to Compare the Usability of Competing Products, in Proceeding of the Human Computer Interaction International Conference (HCII 2005), Las Vegas, USA.

Sauro, J., and Lewis, J., 2005, Estimating Completion Rates from Small Samples using Binomial Confidence Intervals: Comparisons and Recommendations, Proceedings of the Human Factors and Ergonomics Society Annual Meeting HFES, Orlando, FL.

Scheffé, H., 1959, The Analysis of Variance, John Wiley & Sons Inc.

Schubert, P. and Dettling, W., 2002, Extended Web Assessment Method (EWAM) – Evaluation of E-Commerce Applications from the Customer's Viewpoint, Proceedings of 35th Hawaii International Conference on System Sciences (HICSS-35'02), IEEE Computer Society.

Schulte-Mecklenbeck, M., & Huber, O., 2003, Information search in the laboratory and on the web: With or without an experimenter, Behavior Research & Methods, Instruments & Computers Volume 35, Number 2, pp. 227-235.

Shackel, B., 1986, Ergonomics in design for usability, In Proceedings of the Second Conference of the British Computer Society, Human Computer interaction Specialist Group on People and Computers: Designing For Usability (BCS-HCI 86), M. D. Harrison and A. F. Monk, Eds, Cambridge University Press, New York, pp. 44-64.

Shackel, B., 1990, Human Factors and Usability, in: Preece, J. and Keller, L.S. (Eds.), Human-Computer Interaction: Selected Readings, Prentice Hall, UK, pp. 27-41.

Shackel, B., 1991, Usability – Context, Framework, Definition, Design and Evaluation, Chapter 1 in: Human Factors for Informatics Usability, Shackel, B., and Richardson, S. J., Cambridge University Press, pp. 39 - 71.

Shackel, B, 2000, People and computers – some recent highlights, Applied Ergonomics, Volume 31, pp. 595-608.

Shaw, M., Gardner, D. and Thomas, H., 1997, Research opportunities in electronic commerce, Decision Support Systems, Volume 21, pp. 27-33.

Sheth, J.N. and Sisodia, R.S., 1997, Consumer behavior in the future, in Peterson R.A. (Ed.), Electronic Marketing and the Consumer, Sage Publications, Thousand Oaks.

Sheppard, B.H., Hartwick, J., & Warshaw, P.R, 1988, The theory of reasoned action: A meta-analysis of past research with recommendations for modifications and future research, Journal of Consumer Research, Volume 15, pp. 325-343.

Shih, Y-Y. and Fang, K., 2004, The use of a decomposed theory of planned behavior to study Internet banking in Taiwan, Internet Research: Electronic Networking Applications and Policy, Volume 14, Number 3, pp. 213-223.

Shneiderman, B., 1987, Designing the user interface: Strategies for effective human-computer interaction, Reading, MA: Addison-Wesley.

Shneiderman, B., 1992, Designing the User Interface: Strategies for Effective Human-Computer Interaction, Addison-Wesley.

Shneiderman, B., 1995, A Taxonomy and Rule Base for the Selection of Interaction Styles, in Baeker, R.M., Grudin, J., Buxton, W.A.S. and Greenberg, S., Readings in Human-Computer Interaction: Toward the Year 2000, Second Edition, Morgan Kaufmann Publishers Inc., USA, pp. 401-410.

Shneiderman, B., 2002, interviewed in Preece, J., Rogers, Y. and Sharp, H., Interaction Design: Beyond Human-Computer Interaction, John Wiley & Sons Inc., NY, 2002.

Shneiderman, B. and Plaisant, C., 2005, Designing the User-Interface: Strategies for Effective Human Computer Interaction, 4th Edition, Pearson, Addison Wesley.

Siddarth, S. and Chattopadhyay, A., 1998, To zap or not to zap; A study of the determinants of channel switching during commercials, Marketing Science, Volume 17, Issue 2, pp. 124-138.

Sohail, M. S. and Shanmugham, B., 2003, E-banking and Customer Preference in Malaysia: An Empirical Investigation, Information Sciences, Volume 150, pp. 207-217.

Spool, J., Scanlon, T., Snyder, C., DeAngelo, T., and Schroeder, W., 1997, Web Site Usability: A Designer's Guide, Morgan Kaufmann, San Francisco.

Spool, J.M. and Schroeder, W., 2001, Testing Web Sites: Five Users is Nowhere Near Enough, Extended Abstracts of CHI 2001, in Jacko, J. and Sears, A. (Eds), Conference on Human Factors in Computing Systems (CHI '01), ACM Press, pp. 285-286.

Stamoulis, D., Kanellis, P. and Martakos, D., 2002, An approach and model for assessing the business value of e-banking distribution channels: evaluation and communication, International Journal of Information Management, Volume 22, Issue 4, pp. 247-261.

Stanton, N.A., and Young, M.S., What price ergonomics?, 1999, Nature, Volume 399, pp. 197 – 198

Suh, B., and Han, I., 2002, Effect of trust on customer acceptance of Internet banking, Electronic Commerce Research and Applications, Volume 1, Issues 3-4, pp. 247-263.

Szajna, B., 1994, Software evaluation and choice: predictive evaluation of the Technology Acceptance Instrument, MIS Quarterly, Volume 18, Number 3, pp. 319-324.

Szajna, B., 1996, Empirical Evaluation of the Revised Technology Acceptance Model, Management Science, Volume 42, Number 1, pp. 85-92.

Tan, M and Teo, T., S., H., 2000, Factors Influencing the Adoption of Internet Banking, J. Association for Information Systems, Volume 1, Article 5, July 2000, pp. 1-42.

Taylor , S. and Todd, P., 1995, Assessing IT Usage: The Role of Prior Experience, # MIS Quarterly, Vol. 19, No. 4 (Dec. 1995), pp. 561-570.

Teague, R., De Jesus, K., Ueno, M.N., 2001, Concurrent vs. post-task usability test ratings, In Extended abstracts on Human factors in computing systems (CHI '01), pp. 289 - 290.

Thimbleby, H., 1990, User-Interface Design, ACM Press.

Thurstone, L.L., 1967, Attitudes can be measured, Chapter 10, In. M. Fishbein (Ed.), Readings in attitude theory and measurement, New York, Wiley, pp. 77-89.

Tognazzini, B., 2005, Design for Usability, Chapter 3 in: Security and Usability, Eds. Cranor and Garfinkel, O'Reilly, pp. 31-46.

Tohidi, M., Buxton, W., Baecker, R., and Sellen, A., 2006, Getting the Right Design and the Design Right: Testing Many Is Better Than One, In: Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems 2006. pp. 1243-1252.

Toledano, D.T., Fernandez Pozo, R., Hernandez Trapote, A., and Hernandez Gomez, L., 2006, Usability evaluation of multi-modal biometric verification systems, Interacting with Computers, Volume 18, Issue 5, pp. 1101-1122.

Tzanidou, E., Minocha, S., Petre, M., & Grayson, A, 2005, Revisiting Web Design Guidelines by Exploring Users' Expectations, Preferences and Visual Search Behaviour, In People and Computers XIX: The Bigger Picture: Proceedings of HCI2005, Volume 1, pp. 421-438.

van Den Haak, M.J., De Jong, M.D.T., Schellens, P.J. (2003). Retrospective Versus Concurrent Think-Aloud Protocols: Testing the Usability of an Online Library Catalog, Behavior & Information Technology, Volume 22, Number 5, pp. 339–351.

Vaughan, M. and Courage, C., 2007, SIG: capturing longitudinal usability: what really affects user performance over time?. In Extended Abstracts on Human Factors in Computing Systems (CHI '07), ACM Press, New York, pp. 2149-2152.

Venkatesh, V. and Davis, F.D., 2000, A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies, Management Science, Volume 46, Number 2, pp. 186-204.

Virzi, R.A., 1992, Refining the test phase of usability evaluation: How many subjects is enough?, Human Factors, Volume 34, Number 4, pp. 457-468.

Walker, M.A., Fromer, J., Di Fabbrizio, G., Mestel, C., and Hindle, D., 1998, What can I say?: evaluating a spoken language interface to Email, In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'98), Karat, C., Lund, A., Coutaz, J., and Karat, J., Eds., ACM Press/Addison-Wesley, New York, NY, pp. 582-589.

Weir, C.S., Douglas, G. and Jack, M.A., 2005, Examining Usability and Security in 2-Factor Authentication for Internet Banking, Commercial in Confidence Research Report submitted to Lloyds TSB Group plc. (A paper based on this work has been accepted to be published in the journal Computers & Security, *In Press*, September 2008).

Weir, C.S., Anderson, J.A., and Jack, M.A., 2006, On the role of metaphor and language in design of third party payments in eBanking: usability and quality, International Journal of Human-Computer Studies, Volume 64, Issue 8, pp. 771-785.

Weir, C.S., McKay, I. and Jack, M.A., 2007, Functionality and Usability in Design for eStatements in eBanking services, Interacting with Computers, Volume 19, Issue 2, pp. 241-256.

Wenham, D., Zaphiris, P., 2003, User interface evaluation methods for internet banking web sites: a review, evaluation and case study, in Proceedings of the Human–Computer Interaction International Conference (HCII'03), Lawrence Erlbaum, pp. 721–725.

Whiteside, J., Bennett, J. and Holtzblatt, K., 1988, Usability Engineering: our experience and evolution, in Helander, M. (Ed.), Handbook of Human Computer Interaction, Amsterdam: NorthHolland, pp. 791-817.

Whiteside, J. and Wixon, D., 1985, Developmental theory as a framework for studying HCI, Chapter 2, In, Hartson, H. R. (Ed.), Advances in Human-Computer Interaction, Volume 1, pp.29-48.

Wright, A., 2002, The changing competitive landscape of retail banking in the e-commerce age, Thunderbird International Business Review, Volume 44, Issue 1, pp. 71-84.

Yi, M.Y. and Hwang, Y., 2003, Predicting the use of web-based information systems: self-efficacy, enjoyment, learning goal orientation, and the technology acceptance model, International Journal of Human-Computer Studies, Volume 59, Issue 4, pp. 431-449.

Yousafzai, S. Y., Pallister, J. G., and Foxall, G. R., 2003, A proposed model of e-trust for electronic banking, Technovation, Volume 23, Issue 11, pp. 847-860.

Zhang, P. and von Dran, G. M., 2000, Satisfiers and dissatisfiers: a two-factor model for website design and evaluation, Journal of the American Society for Information Science, Volume 51, pp. 1253-1268.

Ziefle, M., 2002, The influence of user expertise and phone complexity on performance, ease of user and learnability of different mobile phones, Behavior & Information Technology, Volume 21, Issue 5, pp. 303-311, Taylor & Francis.

Zwass, V., 2003, Electronic commerce and organisational innovation: aspects and opportunities, International Journal of Electronic Commerce, Volume 7, Number 3, pp. 7-38.

Appendices

Appendix A – Statistics Examples

A1: Repeated-Measures ANOVA

A1.1: Example 1 – Questionnaire Means, Pilot 1.

2-cells, repeated measures, within subjects design, with age group (a) and gender (g) as between-subject factors (2 levels each): N = 24; C = 2; $k_a = 2$; $k_g = 2$; $N_T = 48$.

Where:

N = Number of participants.

C = Number of repeated measures conditions.

N_T = Total number of data points (so N x C = N_T).

k_i = Number of levels for between-participant factor 'i'.

SPSS output and example degrees of freedom calculations:

Within-Subjects Factors

Interface	Dependent Variable
Design A	AttMean_A
Design B	AttMean_B

Between-Subjects Factors

		N
Age	<30	9
	30+	15
Gender	F	12
	M	12

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Interface	Sphericity Assumed[6]	5.703	**1**	5.703	**5.938**	**.024**
Interface * Age	Sphericity Assumed	<.001	1	<.001	<.001	.989
Interface * Gender	Sphericity Assumed	<.001	1	<.001	<.001	.987
Interface * Age * Gender	Sphericity Assumed	.990	1	.990	1.031	.322
Error(Interface)	Sphericity Assumed	19.209	**20**	.960		

[6] With a 2-cell design, there is no sphericity calculation. Therefore the within-subjects effects table shown does not include the Greenhouse-Geisser or other corrections for violations of sphericity. These will be discussed in the third example of a 3-cell design.

Degrees of freedom (df)

Main effect df *Interface design*: Number of conditions (C) – 1 = 2 – 1 = 1. Similarly, df *Interface by Age*: (C – 1) x (k_a – 1) = 1 x 1 = 1; and df *Interface by Age and Gender*: (C – 1) x (k_a – 1) x (k_g – 1) = 1 x 1 x 1 = 1, etc.

Error term df: (C – 1)N – (sum of effect dfs) = (1 x 24) – (1 + 1 + 1 + 1) = 24 – 4 = 20.

Example of reporting: There was a main effect for interface design, F(1, 20) = 5.938, p = .024 (see p.75).

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	841.159	1	841.159	902.042	.000
Age	2.076	1	2.076	2.226	.151
Gender	.029	1	.029	.031	.861
Age * Gender	3.716	1	3.716	3.985	.060
Error	18.650	20	.933		

For between subject effects, the results do not depend on the number of interfaces, rather the mean over all interfaces. The intercept, the grand mean of all data points has df = 1.

For the age/gender interaction (marginally significant here): Age * Gender effect df: (k_a – 1) x (k_g – 1) = (2 – 1) x (2 – 1) = 1; Similarly, for the other between-subject factors, Age effect df = (k_a – 1) = 2, etc.

Error term df: N – (sum of effect dfs) = 24 – (1 + 1 + 1 + 1) = 24 – 4 = 20

Example of reporting: The between-subject age and gender interaction, F(1, 20) = 3.985, p = .060 (not a significant finding in this case).

A1.2: Example 2 – Questionnaire Means, Experiment 3.

3-cells, repeated measures, within subjects design, with order (o), age group (a) and gender (g) as between-subject factors. The order variable involves 6 levels; age group involves 3 levels and gender again is 2 levels:

N = 178; C = 3; k_o = 6; k_a = 3; k_g = 2; N_T = 534.

SPSS output and example degrees of freedom calculations:

Within-Subjects Factors

Interface	Dependent Variable
DO	AttMean_DO
SS	AttMean_SS
AS	AttMean_AS

Between-Subjects Factors

		N
Age	<30 years	65
	30-49 years	45
	50 years +	68
Gender	Female	81
	Male	97
Order	DO -> SS -> AS	25
	DO -> AS -> SS	34
	SS -> DO -> AS	35
	SS -> AS -> DO	30
	AS -> DO -> SS	27
	AS -> SS -> DO	27

Mauchly's Test of Sphericity

Within-Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Interface	.976	3.520	2	**.172**	.976	1.000	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

[a] May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

In this example Mauchly's test was not significant, therefore sphericity is assumed and no adjustment to the degrees of freedom is needed for the within-subject effects.

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Interface	Sphericity Assumed	60.266	**2**	30.133	**49.456**	**.000**
Interface * Age	Sphericity Assumed	4.821	4	1.205	1.978	.098
Interface * Gender	Sphericity Assumed	.252	2	.126	.207	.813
Interface * Order	Sphericity Assumed	14.937	10	1.494	2.452	.008
Interface * Age * Gender	Sphericity Assumed	1.524	4	.381	.625	.645
Interface * Age * Order	Sphericity Assumed	7.666	20	.383	.629	.890
Interface * Gender * Order	Sphericity Assumed	5.286	10	.529	.868	.564
Interface * Age * Gender * Order	Sphericity Assumed	9.040	18[7]	.502	.824	.671
Error(Interface)	Sphericity Assumed	174.254	**286**	.609		

[7] The degrees of freedom for the complex 4-way interaction are modified due to missing participants in the Female, 30-49 years age group for the order of experience AS -> SS -> DO (for all three interfaces). Therefore N=3 combinations of factor levels are missing from a full-factorial design. From the partitioning of the experimental effect, two degrees of freedom are subtracted from the within-subjects effect df (20 – 2 = 18). Similarly for the between-subjects 3-way interaction, where the third df is subtracted from the expected df of 10 (resulting in a df for the 3-way interaction of 9).

Degrees of freedom (df)

Main effect df *Interface design*: Number of conditions (C) – 1 = 3 – 1 = 2. Similarly, df *Interface by Order*: (C – 1) x (k_o – 1) = 2 x 5 = 10; and df *Interface by Age and Gender*: (C – 1) x (k_a – 1) x (k_g – 1) = 2 x 2 x 1 = 4, etc.

Error term df: N(C – 1) – (sum of effect dfs) = (2 x 178) – (2 + 4 + 2 + 10 + 4 + 20 + 10 + 18) = 356 – 70 = 286.

Example of reporting: There was a main effect for interface design, $F(2, 286) = 49.456$, $p < .001$. There was also a significant interaction between interface design and order, $F(10, 286) = 2.452$, $p = .008$ (see p.207).

Notice that the F-test determines that there are significant differences between interfaces. It does not indicate where within the three interfaces the significant differences lie. These effects are determined by the pairwise comparisons, as discussed on p.35 and illustrated for these data on p.207.

Tests of Between-Subjects Effects

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	9946.868	1	9946.868	6440.606	.000
Age group	4.681	2	2.341	1.516	.223
Gender	1.447	1	1.447	.937	.335
Order	36.011	**5**	7.202	**4.663**	**.001**
Age group * Gender	.868	2	.434	.281	.755
Age group * Order	8.076	10	.808	.523	.872
Gender * Order	5.159	5	1.032	.668	.648
Age group * Gender * Order	11.946	9*	1.327	.859	.563
Error	220.849	143	1.544		

* see footnote 7 (previous page)

For between subject effects, the results do not depend on the number of interfaces, rather the mean over all interfaces. The intercept, the grand mean of all data points has df = 1.

For the order effect (significant for this example): Order effect df: (k_o – 1) = (6 – 1) = 5; Similarly, for the other between-subject factors, Age effect df = (k_a – 1) = 2; Age/Order effect df = (k_a – 1) x (k_o – 1) = 2 x 5 = 10, etc.

Error term df: N – (sum of effect dfs) = 178 – (1 + 2 + 1 + 5 + 2 + 10 + 5 + 9) = 178 – 35 = 143.

Example of reporting: There was a significant between-subject order interaction, $F(5, 143) = 4.663$, $p = .001$ (see p.207).

A1.3: Example 3 – Relative Intention, Expt. 3 – Sphericity Violated.

In this example, from the same experiment design specifications as example 2 (above), Mauchly's test was significant, therefore the assumption of sphericity was violated, and the correction to the degrees of freedom (Greenhouse-Geisser) was applied to the within-subject effects as shown below and in the table overleaf.

SPSS output and example degrees of freedom calculations:

Within-Subjects Factors

Interface	Dependent Variable
DO	RelBI_DO
SS	RelBI_SS
AS	RelBI_AS

Mauchly's Test of Sphericity

Within-Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[a]		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Interface	.941	**8.630**	**2**	**.013**	**.944**	1.000	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

[a] May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

Degrees of freedom (df)

The main effect df *Interface design* (Number of conditions (C) – 1 = 3 – 1 = 2) are corrected using the Greenhouse-Geisser ε = .944: 2 x .944 = 1.889.

Similarly, the Error term df is also corrected: 286 x .944 = 270.074.

Example of reporting: Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(2) = 8.630$, $p = .013$. Therefore the degrees of freedom were corrected using the Greenhouse-Geisser estimates of sphericity (ε = .944). There was a main effect for interface design, $F(1.889, 270.074) = 54.335$, $p < .001$. (see p.216).

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
RelBI	Sphericity Assumed	440.397	2	220.198	54.335	.000
	Greenhouse-Geisser	440.397	**1.889**	233.183	**54.335**	**.000**
	Huynh-Feldt	440.397	2.000	220.198	54.335	.000
	Lower-bound	440.397	1.000	440.397	54.335	.000
RelBI * Age	Sphericity Assumed	20.170	4	5.043	1.244	.292
	Greenhouse-Geisser	20.170	3.777	5.340	1.244	.293
	Huynh-Feldt	20.170	4.000	5.043	1.244	.292
	Lower-bound	20.170	2.000	10.085	1.244	.291
RelBI * Gender	Sphericity Assumed	1.394	2	.697	.172	.842
	Greenhouse-Geisser	1.394	1.889	.738	.172	.830
	Huynh-Feldt	1.394	2.000	.697	.172	.842
	Lower-bound	1.394	1.000	1.394	.172	.679
RelBI * Order	Sphericity Assumed	70.988	10	7.099	1.752	.069
	Greenhouse-Geisser	70.988	9.443	7.517	1.752	.074
	Huynh-Feldt	70.988	10.000	7.099	1.752	.069
	Lower-bound	70.988	5.000	14.198	1.752	.127
RelBI * Age * Gender	Sphericity Assumed	38.913	4	9.728	2.400	.050
	Greenhouse-Geisser	38.913	3.777	10.302	2.400	.054
	Huynh-Feldt	38.913	4.000	9.728	2.400	.050
	Lower-bound	38.913	2.000	19.457	2.400	.094
RelBI * Age * Order	Sphericity Assumed	91.459	20	4.573	1.128	.319
	Greenhouse-Geisser	91.459	18.886	4.843	1.128	.322
	Huynh-Feldt	91.459	20.000	4.573	1.128	.319
	Lower-bound	91.459	10.000	9.146	1.128	.345
RelBI * Gender * Order	Sphericity Assumed	66.553	10	6.655	1.642	.094
	Greenhouse-Geisser	66.553	9.443	7.048	1.642	.099
	Huynh-Feldt	66.553	10.000	6.655	1.642	.094
	Lower-bound	66.553	5.000	13.311	1.642	.153
RelBI * Age * Gender * Order	Sphericity Assumed	74.446	18	4.136	1.021	.436
	Greenhouse-Geisser	74.446	16.998	4.380	1.021	.436
	Huynh-Feldt	74.446	18.000	4.136	1.021	.436
	Lower-bound	74.446	9.000	8.272	1.021	.427
Error(RelBI)	Sphericity Assumed	1159.045	286	4.053		
	Greenhouse-Geisser	1159.045	**270.074**	4.292		
	Huynh-Feldt	1159.045	286.000	4.053		
	Lower-bound	1159.045	143.000	8.105		

Appendix B – Questionnaires

B1. TAM Questions

B1.1. Original Published Questions

The original set of attributes constructed for the Technology Acceptance Model (TAM) (Davis, 1989). The attributes in the questionnaire were scored between 'extremely likely' and 'extremely unlikely' on a 7-point Likert style scale.

Perceived Usefulness (U)

- Using <system> in my job would enable me to accomplish tasks more quickly.
- Using <system> would improve my job performance.
- Using <system> in my job would increase my productivity.
- Using <system> would enhance my effectiveness on the job.
- Using <system> would make it easier to do my job.
- I would find <system> useful in my job.

Perceived Ease of Use (EOU)

- Learning to operate <system> would be easy for me.
- I would find it easy to get <system> to do what I want it to do.
- My interaction with <system> would be clear and understandable.
- I would find <system> to be flexible to interact with.
- It would be easy for me to become skilful at using <system>.

Usage

The original study (Davis, 1989) relied on self-reports of current usage using a six-position categorical scale:

- Don't use at all.
- Use less than once each week.
- Use about once each week.
- Use several times a week.
- Use about once each day.
- Use several times each day.

B1.2. Updated TAM Questions

Further studies allowed the questions to be condensed into this set typically included in most TAM research (Davis & Venkatesh, 2004).

Intention to Use

- Assuming I had access to <system>, I intend to use it.
- Given that I had access to <system>, I predict that I would use it.

Perceived Usefulness

- Using <system> will improve my performance on my job.
- Using <system> in my job will increase my productivity.
- Using <system> will enhance my effectiveness in my job.
- I find <system> would be useful in my job.

Perceived Ease of Use

- My interaction with <system> will be clear and understandable.
- Interacting with <system> will not require a lot of my mental effort.
- I find <system> will be easy to use.
- I will find it easy to get <system> to do what I want it to do.

Responses on a 7-point Likert scale as before.

B2: Usability for Automated Telephone Services (ATS)

The original, proven usability attitude questionnaire from CCIR's extensive research into automated telephony systems (Dutton et al, 1993) was used as a base for adaptation to Web services.

COGNITIVE EFFORT & STRESS

I found the service confusing to use
I had to concentrate hard to use the service
I got flustered when using the ATS
I felt under stress while using the ATS
I felt frustrated while using the ATS
I thought the ATS was too complicated

FLUENCY

When I was using the ATS I didn't know what I was expected to do
I felt out of control while using the ATS
The ATS was too fast for me

TRANSPARENCY

The ATS was easy to use
I thought the voice was very clear
I would prefer to give my number to a human being

QUALITY OF INTERFACE

I would be happy to use the ATS again
I felt that the ATS was reliable
I thought the ATS was efficient
I feel that the ATS needs a lot of improvement

CONVERSATIONAL MODEL & FRIENDLINESS

The ATS was friendly
I liked the voice
I enjoyed using the ATS
I thought the ATS was polite

Appendix C – Qualitative Techniques

C1: Training in Facilitation Techniques

The following training guidelines were provided in advance of the user-sessions to ensure all facilitators had sufficient knowledge in what qualitative data was to be collected and how to obtain it. This session was followed by training in the priming and detailed documentation, materials and interfaces to be used in each particular experiment. Typically two days of training was allowed for in the schedule. A variety of facilitators were used in each experiment, all were fully trained and many had years of usability, facilitation and interviewing experience.

General Code of Conduct

- Participants should always be treated with respect.
- Be friendly and try to put the participant at ease.
- Dress code is smart, no blue jeans or trainers.

The appropriate attitude for a researcher is one of professional detachment and neutrality. Using encouraging terms like "Good" or "Well done" may give the impression that the participant, rather than the system, is being evaluated.

Avoid the temptation to finish participants' sentences for them, or to tell them what you think is on their minds. Instead, **maintain your silence, listen, and be attentive**

Beginning the Session:

- Inform the participant that THEY are not being tested, the *system* is.
- That they will be told what they will be doing at each stage.
- That participation is voluntary.
- That they can stop the session at any time.

Ensure that the participant is comfortable before you begin. Ask them if they have any questions. Emphasise that their opinions are important to us, and will be helpful in designing the final product or service.

Information

What you CAN tell the participants (if they ask or if you feel you need to)

- Typically a session will last 45 minutes to an hour (or specified period).
- They should try each task, but if they can't figure it out, they should feel free to move on to another.
- For the think aloud protocol, you may need to remind them periodically to think aloud and talk through what they are doing and why.
- If the session will be logged automatically (mouse movements, eye tracking, clicks and keystrokes etc.).
- That you will be taking notes about what they do and say.
- Reassure them that you will not include their name in any notes you make.
- Reassure them that the information they provide us is confidential and anonymous.

- Explain that we use their and other participant data to provide statistics and representative comments which are reported back to the sponsor – their name and personal details will not be disclosed.
- Reassure them that we are impartial observers, as such we will not be offended if they do not like something or find something difficult.
- Stress the importance that they answer questions honestly.

What you MUST NOT tell the participants:

- Do not explain the system, what it does or how it works.
- Do not tell participants what to do, where to click, what to look for etc.
- Do not agree or disagree with the participants opinions.
- Do not make any value judgements – such as "this bit is good because...", or "I like this..." etc.

During the Sessions

- Participants may be reminded and encouraged to *think aloud* during the session (if appropriate).
- The researcher should be supportive of the participants taking their own notes.
- Participants are invited to make general comments/suggestions throughout.
- Do not include participant names in any notes you make.
- Folders are marked with ID codes not participant names – they are the lifeblood of the experiment and need to be carefully organised and processed.
- Experimental control is very important. You must give each participant the correct sequence – follow the experiment blueprint.

Script

- You should follow the formal script with each participant in order to ensure that all participants receive the same information.
- Without a script it is *very easy* to give too much or too little information.
- The script will be written in a conversational style. It is not necessary to read it verbatim – and certainly it should not be read aloud from the paper.
- The script should be studied carefully in advance such that each participant is given the same information and NO MORE.
- Stress that we're testing the system, and not the participant's abilities:
- *"If you find parts of the product difficult to use or understand, so will other people, and it is our job to find out how we can improve it."*
- Explain any protocol being used in the session – e.g. the 'think aloud' protocol:
- *"As you work, tell me what you are thinking and doing, what you are looking for. If you don't know what something is for, please say 'I don't know what this is for', or something similar. I may also prompt you from time to time to ask you what you are thinking."*
- Always ask if they have any questions before beginning the session.
- It is good practice to read task sheets out loud as well as offering a printed sheet to the participants. This ensures that the participant has a clear idea of the whole task before they begin to use the system.

Ending the session

- Finish the session by thanking the participant.
- It may be appropriate to ask them if they can participate in further research, for example, discussion groups.
- When asking about participation in discussion groups make sure that you tell participants if the session will be video taped and that they will be asked to sign a consent form.
- Give the participant their payment at the end of the session.
- *Participants* must sign the cheque-signing sheet to confirm receipt of the cheque.

Notes

- Take detailed notes about what the participant says.
- Record events in as much detail as possible—to the level of keystrokes and mouse clicks if necessary.
- **Do not prompt participants** unless it is clearly necessary to do so. Make detailed notes about *any help* you have had to give and report this on the observations sheet.

Exit Interview, Debrief and Final Questions

- Write down everything that the participant tells you.
- Make sure you ask "why?" and get reasons for participants preferences, opinions etc. where possible.
- Add participants' comments to the data entry sheets using their own words, i.e. write: *"I like the directory"* NOT *"This participant said he/she liked the directory"*.

Data

During an experiment, you may facilitate many different and sequential participant sessions. It is often helpful to review your notes straight after each session if time permits.

Data must be transferred from the folders into the computer. If possible, entering data directly after each session helps to maintain the accuracy of the data.

- **Be meticulous** in data entry.
- Have another researcher check 10% of your data: take some folders at random from the finished set and check the written notes with the entered data for accuracy. Count any mistakes and change values in the data sheet.
- Log the error rate in the data sheet.
- Don't rush through data entry – accuracy is crucial to our work.

Technical Issues

When things do go wrong remember, "The show must go on". Log the problem and details on the observations sheet but try to continue with the experiment if possible.

Problems

- Double bookings – if two participants arrive for the same slot, try to get one of them to book in again – see the researcher in charge of bookings. When a participant has turned up, they should always be paid and sign the cheque signing sheet, whether or not they participate.
- If you feel uncomfortable in the session with a participant – if they are being personally inappropriate or unreasonable, make up a plausible excuse (software glitch), pay them and let them go. Always tell the researcher in charge, who will ensure they are not called again.
- If a participant does not want to continue, is uncomfortable or has changed their mind about participating, again they must be paid and sign the sheet. Their names should also be given to Douglas so they are not phoned again.
- As a last resort, there are Panic Buttons in each experiment room. If you feel physically threatened and are not able to terminate the experiment yourself, press the button and another member of staff will intervene.

Observations by Researchers

You should record any general observations and note any patterns you saw in the testing sessions. You will have a sheet in each folder to record your general observations and report problems with the session.

This will be in addition to the detailed progress sheets to complete during a participants interaction with the system and tasks

Along with your data entry, you will be expected to write a 1-2 page document detailing your observations. The document should summarise the main themes that you have noticed in the experiment session. It should also record any aspects of the experiment that were unexpected or not covered in the question and answer sessions.

Use the observation sheet to note down general feelings and problems. Use the progress sheets to note down 'think aloud' data, comments and step-by-step descriptions of the interaction where appropriate.

Observations might include:

- What worked well?
- Was anything consistently misunderstood – what?
- Any mistakes consistently made – what?
- Did they miss anything – what?
- Which tasks caused the most trouble?
- When did they seem frustrated? what were they doing?
- Where did the service fail them?
- What made them confused?
- What help you gave participants and why this was necessary.

In your general observations it is important to clarify whether or not participants (in general) were examples of actual users (such as a branch or the Internet), or if participants seemed intimidated by the equipment they were using in the sessions.

Whether or not each participant could use a computer is very important information for Internet experiments, as only current computer users would be likely to take up an Internet service. As such, there will be a column at the beginning of the data sheet for important per

participant observations. There will be a second column to note major per participant problems (such as NOT A COMPUTER USER or DIDN'T COMPLETE EXPERIMENT etc.). You should only write important per participant comments in the data sheet that might influence whether or not a participants data is analysed in the set.

The Observations document is used to summarise the general issues – these will not be per participant remarks.

The Think Aloud Protocol

The principles of think aloud are that the participant tries to offer a verbal blow-by-blow description of their thoughts whilst interacting with a service. Participants often find this difficult and may start off describing their thought processes, but may soon trail off into silence.

Techniques

Reminding and prompting participants to talk through what they are doing helps. Other techniques can also provide useful information without drastically interrupting them.

Good Prompts

The following questions can help remind the participant and get information without introducing bias:

- What is your goal?
- What did you expect when you did that?
- How did you expect that to work?
- Can you tell me what you were thinking?
- What do you want to accomplish here?
- Describe the steps you are going through here.
- How did you feel about that process?
- Tell me about your thinking here.
- What did you expect to happen when you . . . .?

The Echoing Technique

Repeating the participant's own words or phrases back to them as a question: "That message is confusing?" Usually, participants will be happy to clarify their thoughts.

Signal that you are there, you are interested, but that it is still their turn to talk (mmm, hmm, etc ...).

Always Ask "Why?"

This is the most simple, unbiased question to ask, and often produces very useful information.

Whenever it is appropriate to do so, ask the participants for more information about their statements: "Why?".

Non-verbal Cues

When a participant reacts physically to an experience, but does not say anything, it can be helpful to pick them up on it and ask if there was anything particular causing them to frown, smile etc.

Focus the Discussion

Participants will stray off topic. Allowing people some freedom to steer the discussion into topics that are important to them is good practice. However, participants who are commenting off-topic and don't seem able to make their point quickly, may need to be interrupted. Try to refocus the discussion, or move on to the next question/task.

Probe Expectations

Before participants perform an action (click a link, check a box etc.) they will have some expectation of what will happen. By asking "What do you think will happen when you click there?" it is possible to find out about such expectations. In addition, it is also useful to ask: "Is that what you expected?" immediately after an action.

Personal Experience

Some participants tend to try and generalise their experience to include those they know: husbands, wives, children, parents and friends. They often will suggest: "although I don't find this useful, others will". Their personal experiences during the usability study are by far the more useful comments than their expectations of how other people may use things.

If someone tells you: "I think it may be useful to someone, someday", then ask him or her: "is it useful to you, now?"

Appendix D - Pilot Experiments 1 & 2

D1: Experiment Materials

D1.1. Tasks

Task Sheet 1

1. You would like to get a Credit Card. Find out how old you have to be to register for a Lloyds TSB Asset Card?
2. You are interested in registering for Lloyds TSB Telephone Banking. Find out what security features are used by the PhoneBank Express service.
3. Whilst looking for a loan you decide to check the insurance options. What does Loan Protection provide?
4. You want to find out about the Lloyds TSB Internet Banking service. During what hours is the service available?
5. Look into getting an overdraft facility on your account. Which account offers the lowest interest rate for an agreed overdraft?

Task Sheet 2

1. What banking services are available to Lloyds TSB customers when they go to the Post Office?
2. You want to find information from Lloyds TSB about their Mortgage products. What is Mortgagesure?
3. By what method can you buy and sell shares with Lloyds TSB Stockbrokers' Sharedeal Direct?
4. You will be getting a Cashpoint card with your new account. How many Lloyds TSB Cashpoint machines are there in the UK?
5. If you had £2,500 to put in a savings account, which Lloyds TSB account would give you the highest interest rate?

D1.2. Interview Questions

Now thinking about your preferences, I would like you to rate the two Websites you have used today. Place the markers along the scale between Best and Worst, I will record the order and positions.

Can you tell me more about which features of the websites you liked or disliked?

Do you have any suggestions for improving the websites?

Would your experience today encourage you to look at financial information on the Internet? (Y/N/don't know)

Any other comments?

D2: Qualitative Data

The interface designs evaluated:

Design A (Pilot 1) – Indexed interface.

Design B (Pilots 1 and 2) – Cluttered interface.

Design C (Pilot 2) – Reduced Clutter version of Design B – holding all information and structure constant.

D2.1. Interview Responses from Pilot 2

Likes:

- "The information on this one (C) is less cluttered and comes across better, clearer"

Dislikes:

- The middle of the page is not used to the best advantage (B)"
- "I disliked the flashy and annoying adverts (B)"
- "This one (C) is a bit drab in terms of appearance"
- "The home page was awful (B) just full of advertising and blurb, cluttered, trying to sell not provide – it gives a bad impression."

Intention to use:

Those who were not interested or unsure about using the Internet for financial information were either not really interested in looking at banking information, or were not convinced of the security of Internet banking and therefore didn't really look at financial information online either.

D2.2. Observations & Comments from Pilot Studies 1 & 2

Design B and its variants caused more confusion for participants than Design A. Participants commented that the layouts of the pages were not user friendly. Many participants were instantly drawn to the central menu options (in the content pane), but found them of limited use; in fact many deemed them completely irrelevant.

Some general problems with all the designs were the size of text on some of the menus, and the use of menus instead of using inline hyperlinks on the page content. There were frequent comments that "the information seems to be spread around the screen". "It was frustrating reading the word I was looking for in the centre of the screen but not being able to click on it to find out more." Participants generally were trying (usually in vain) to find hyperlinked words in the sections they were reading, they found it very confusing that they had found a relevant part of the site, but then had to search around for a link in order to continue.

Design A (Indexed)

The home page was generally considered very simple to look at, as there were only four options that they had to deal with. Navigation was fast with this design and most participants appreciated the A-Z index on offer, although some voiced that it "did not list things in a logical manner". Some participants who tended to use the index for everything commented that: "Once I have found something that works on a Web site, I'll just keep using it."

The top right hand (very small font text) menu was commented on frequently in terms of (smallness of) font size and location – it was thought to be too small and did not grab attention.

Facilitators thought that participants were noticeably faster on certain tasks with this design. Most participants were happier with the menu structure and options, although gave several negative comments made about the appearance: "The use of pale green links" and "black on grey links" that were "hard to read". "It wasn't always apparent what was a clickable link and what was not". They usually voiced a preference for the alternative design in terms of appearance.

Designs B and C (General)

These design variants were both based around the hierarchical organisational metaphor of the Design B and included the same options.

A few comments suggested that these designs in general seemed liked a collection of leaflets thrown together - more of an information site, rather than allowing the user to ask questions, cross-compare accounts/rates etc. or cross-link between related topics.

Not many participants noticed the RH menu at the beginning of the tasks but did gradually notice it as the experiment progressed and they persistently had to search for further information on tasks – in this way some noted that with "a bit more time I would know my way around the site without much problem", whilst others realised that "At home, I would probably have given up by now".

Participants were observed screwing their eyes up and leaning towards the screen, struggling to read the text of some menus. Similarly, menus which appeared intermittently went unnoticed by some participants, while others were simply confused when it appeared or disappeared. "This needs to be more prominent".

Several people mentioned that they didn't like scrolling on the home page, as all information should be available 'at a glance'.

The 'Our rates' menu was not used that much or it took a long time for participants to find. Those who saw the link were pleased. Some participants had found it very hard to compare rates with the Design A – wanting a specific feature for rate comparison. However, relatively few participants did notice the link on the redesign, and some who did still could not find the rates they were looking for amongst the long list. Many suggested that having different sections for Credit Cards, Savings etc. would have been better than using the account names.

Design B (Cluttered)

- "There is just too much clutter, all this information, writing and pictures all compressed onto the screen."
- "It felt like I was stumbling across things at random, just like a collection of leaflets thrown together – not using the internet to its advantage – should allow more cross-comparison and linking."
- "I felt like it was designed for someone who had a Lloyds TSB account already – not good as a front end site."
- "There were too many things on screen."

Design C (Reduced Clutter)

When the clutter was removed, participants generally commented very favourably on this issue. Many were convinced that it was better and easier to navigate around:

- "It is much easier to see what is going on, now that the clutter is gone."
- "I am much happier looking at this one, the white space on the right makes me feel more in control, not nearly so many different elements fighting for my attention."

D3: Correlations and Questionnaire Structure

For Pilot Study 2, the sample size was large enough (N = 44) to examine and begin to structure the usability questionnaire.

The correlations for the most appropriate interface – the reduced clutter Design C are shown. The matrices show high inter-item correlations (highly significant), for semantically related attributes.

		Like Using	User-Friendly	Enjoyment	Flustered	Stressful	In Control
Like Using	*r*	1	.725	.815	.555	.507	.636
	p	-	<.001	<.001	<.001	<.001	<.001
User-Friendly	*r*	.725	1	.773	.577	.599	.646
	p	<.001	-	<.001	<.001	<.001	<.001
Enjoyment	*r*	.815	.773	1	.543	.475	.555
	p	<.001	<.001	-	<.001	.001	<.001
Flustered	*r*	.555	.577	.543	1	.811	.615
	p	<.001	<.001	<.001	-	<.001	<.001
Stressful	*r*	.507	.599	.475	.811	1	.709
	p	<.001	<.001	.001	<.001	-	<.001
In Control	*r*	.636	.646	.555	.615	.709	1
	p	<.001	<.001	<.001	<.001	<.001	-

Table D3.1. Proposed Group – Affect

		Match Expectations	Frustration with IA	Navigation Complication	Quickly Find	Procedure	Orientation
Match Expectations	*r*	1	.612	.536	.654	.615	.511
	p	-	<.001	<.001	<.001	<.001	<.001
Frustration with IA	*r*	.612	1	.640	.796	.537	.400
	p	<.001	-	<.001	<.001	<.001	.007
Navigation Complication	*r*	.536	.640	1	.822	.610	.580
	p	<.001	<.001	-	<.001	<.001	<.001
Quickly Find	*r*	.654	.796	.822	1	.648	.551
	p	<.001	<.001	<.001	-	<.001	<.001
Procedure	*r*	.615	.537	.610	.648	1	.502
	p	<.001	<.001	<.001	<.001	-	.001
Orientation	*r*	.511	.400	.580	.551	.502	1
	p	<.001	.007	<.001	<.001	.001	-

Table D3.2. Proposed Group – Structure

		Confusion (layout)	Clutter	Page Clarity	Link Visibility
Confusion (layout)	r	1	.593	.682	.529
	p	-	<.001	<.001	<.001
Clutter	r	.593	1	.518	.494
	p	<.001	-	<.001	.001
Page Clarity	r	.682	.518	1	.365
	p	<.001	<.001	-	.015
Link Visibility	r	.529	.494	.365	1
	p	<.001	.001	.015	-

Table D3.3. Proposed Group - Page Design

		Helpful	Link Content	Concentration (reading)	Understood Pages	Text Size
Helpful	*r*	1	.566	.523	.403	.281
	p	-	<.001	<.001	.007	.065
Link Content	*r*	.566	1	.448	.547	.412
	p	<.001	-	.002	<.001	.006
Concentration (reading)	*r*	.523	.448	1	.535	.581
	p	<.001	.002	-	<.001	<.001
Understood Pages	*r*	.403	.547	.535	1	.359
	p	.007	<.001	<.001	-	.017
Text Size	*r*	.281	.412	.581	.359	1
	p	.065	.006	<.001	.017	-

Table D3.4. Proposed Group – Content

		Reliable	Trust
Reliable	*r*	1	.619
	p	-	<.001
Trust	*r*	.619	1
	p	<.001	-

Table D3.5. Proposed Group – Integrity

		Use Again	Improvement Needed
Use Again	*r*	1	.799
	p	-	<.001
Improvement Needed	*r*	.799	1
	p	<.001	-

Table D3.6. Proposed Group – Quality

Graphics and Pictures did not correlate highly (or significantly) with Appearance, *r* .263 (*p* = .084). These two items also did not fit well into any other proposed group.

Appendix E - Experiment 1: Metaphors

E1: Experiment Materials

E1.1. Personas

Participants used fictitious personal and bank account data in all the eBanking experiments. This allowed full control of the data offered, different accounts, statements, payment arrangements and all other aspects of the banking service being used. Participants all used the same persona login details, had the same balances and transactions and therefore their experiences could be compared.

E1.2. Tasks

Tasks were selected to represent real world tasks, presented in a convincing scenario including persona details, login information, account numbers for accessing the experimental interfaces. The tasks all concentrated on the section of the interface where participants paid various monies to companies and other people.

Task Sheet 1:

1. You pay for music lessons with money from your savings account. The arrangement has been set up to pay Mrs Jones for this months' lessons. She's going on holiday this month and you've taken a break from your lessons so you need to cancel this payment.
2. You pay the baby-sitter, Linda Hannay with money from your current account directly using the online service. She was baby-sitting last night and you need to send her £25.
3. You have just changed to a new mobile phone company. You used to pay your Vodafone bills using the online service from your current account. Now you can cancel this arrangement.
4. You get your electricity from Npower and pay the bills online using your current account. A bill has just arrived, you need to pay £45 this month.

Task Sheet 2:

1. You pay off your HSBC Visa card online using funds from your current account. You want to pay £60 toward the bill this month.
2. You were paying for your TV licence using an arrangement from your current account. You've decided that you'd prefer to pay for the licence at the post office, so you can cancel this arrangement.
3. You sometimes send money to Oxfam from your current account directly using the online service. You want to donate £25 to them today.
4. You arranged to send £500 from your usual savings account to your High Interest savings account at the end of the month. However, you've had a plumbing problem and now need the money for repairs instead so you need to cancel this payment.

E1.3. Interview Questions

Now thinking about your preferences, I would like you to rate the two services you have used today. Place the markers along the scale between Best service and Worst service, I will record the order and positions. {Note any comments}

What differences did you notice between the two services?

Show pictures of typical screens (form/spreadsheet main page). What aspects of the services did you like?

What aspects did you dislike?

Would you like to be able to send money to other accounts on the Internet Banking service? (Y/N/don't know)

E2: Qualitative Data

The interface designs evaluated:

Design F – Form metaphor.

Design S – Spreadsheet metaphor.

E2.1: Interview Responses

Asked what differences they noticed between the designs, participants commented:

- "Different layout. Seemed the second (S) had more on each page, the first (F) seemed more straightforward."
- "It was much easier to delete payments on the second version (S) but I rated it lower as it was more confusing to use, you couldn't click on "HSBC Visa' for example."
- "First (F) talked you through it more than the second (S)."
- "Had to read the instructions on the second (S) and think about it more."
- "This site (F) was far more intuitive."

Although many participants said they noticed slight differences, only few were able to describe the differences accurately. As detailed comments about the interfaces individually were desired, reminder screen pictures were given to participants before asking them what they liked, disliked and could suggest improving. Typical comments about what they liked about the designs:

- "I liked this one (F) as I felt it was always guiding your next actions, thus it was easier to use."
- "Both were very user-friendly. Remarkably easy to pay money into accounts."
- "Idiot proof."
- "It was functional."
- "The first one (F) was really easy."
- "Easy to get used to the layout of the pages - especially in the first version (F)."
- "Liked this site (S), as it was easy to see all the information at once."
- "Not too much text in the first (S)."
- "I liked the table (S) - convenient and quick to use."

What participants disliked about the designs:

- "The buttons, like 'make payment', 'delete payment' etc., were not obvious at the start."
- "Would like the site to use less banking terms and more colours."
- "There wasn't enough explanation of the different options available on the menu."
- "A lot of the time I found myself not knowing what to click on next to get my request processed. It needs to be clear, as there is no one to help you when you're using the Internet."
- "Sometimes when there is too much information on the one screen, it becomes very confusing, which was the case with the first version (S) I tried."
- "The success page, should tell you what you have succeeded in, like a receipt."
- "Some of the terms were obscure."
- "Too much text in second (S). Having two boxes for pounds and pence - should be just one."
- "Dislike tabbing, pre-filled dates would have been useful."

E3: Correlations and Questionnaire Structure

Correlations between attribute 'Use Again' (as Intent construct) and mean usability questionnaire score:

	Differences (F-S)	Form (F)	Spreadsheet (S)
Pearson's *r*	.752	.676	.804
p	<.001	<.001	<.001
N	32	32	32

Table E3.1. Attribute Use Again – Correlations with Mean Usability Questionnaire

Questionnaire Structure

Again, the structure the usability questionnaire was proposed by examining the correlation matrices. The most appropriate interface for transacting in eBanking was determined to be the Form design. The correlation matrices for this design are shown. The matrices show medium-high inter-item correlations (frequently highly significant), for semantically related attributes (N = 32).

		User-Friendly	Like Using	Enjoyment	Flustered	Stressful	In Control
User-Friendly	*r*	1	.703	.814	.595	.691	.495
	p	-	.000	.000	.000	.000	.004
Like Using	*r*	.703	1	.808	.425	.473	.321
	p	.000	-	.000	.015	.006	.074
Enjoyment	*r*	.814	.808	1	.615	.752	.342
	p	.000	.000	-	.000	.000	.056
Flustered	*r*	.595	.425	.615	1	.624	.396
	p	.000	.015	.000	-	.000	.025
Stressful	*r*	.691	.473	.752	.624	1	.357
	p	.000	.006	.000	.000	-	.045
In Control	*r*	.495	.321	.342	.396	.357	1
	p	.004	.074	.056	.025	.045	-

Table E3.2. Proposed Group – Affect

		Frustration with IA	Navigation Complication	Quickly Find	Procedure	Orientation
Frustration with IA	*r*	1	.294	.403	.317	.294
	p	-	.102	.022	.077	.102
Navigation Complication	*r*	.294	1	.391	.334	.427
	p	.102	-	.027	.062	.015
Quickly Find	*r*	.403	.391	1	.440	.375
	p	.022	.027	-	.012	.034
Procedure	*r*	.317	.334	.440	1	.621
	p	.077	.062	.012	-	.000
Orientation	*r*	.294	.427	.375	.621	1
	p	.102	.015	.034	.000	-

Table E3.3. Proposed Group – Structure

		Confusion (layout)	Page Clarity	Helpful	Concentration (reading)	Understood pages
Confusion (layout)	*r*	1	.328	.354	.320	.358
	p	-	.067	.047	.074	.044
Page Clarity	*r*	.328	1	.158	.357	.508
	p	.067	-	.387	.045	.003
Helpful	*r*	.354	.158	1	.046	.325
	p	.047	.387	-	.801	.069
Concentration (reading)	*r*	.320	.357	.046	1	.166
	p	.074	.045	.801	-	.364
Understood pages	*r*	.358	.508	.325	.166	1
	p	.044	.003	.069	.364	-

Table E3.4. Proposed Group – Content Design

		Use again	Change details
Use again	*r*	1	.501
	p	-	.003
Change details	*r*	.501	1
	p	.003	-

Table E3.5. Proposed Group – Quality

Appendix F - Experiment 2: Dialogue Style

F1: Experiment Materials

F1.1. Interface Language Changes

Table F1.1 describes the changes in dialogue for the two versions of the Internet banking service.

Formal Language (Jargon)	Plain Language Version
Payments and transfers	**Payments and transfers**
To set up a new payment click here. To make, view, amend or cancel a payment, select a beneficiary from the list below. If you have instructed us to make a payment to a beneficiary on a date in the future, the payment due and the date due column will be completed.	To make, view, change or delete a pending payment, click on the recipient name below. To add a new recipient to the list below, click here. If you have previously instructed us to make a payment to a recipient on a date in the future, the Amount and the Date due columns will be completed. If you have instructed us to make regular payments, the Frequency and (where specified) the Expiry date of the payment will also be displayed.
Confirm Payment Type	**Confirm New Recipient**
Do you want to pay a bill to a company, or make a payment to another person? * Pay a bill * Make a payment to another person [OK] [Cancel]	If you want to pay a bill, search for a recipient from our list of companies. Otherwise, make a payment to a specific account by providing its sort code and account number. * Search the list of companies * Specify recipient account details [OK] [Cancel]

Table F1.1. Language Alterations

Formal Language (Jargon)	Plain Language Version
Setting up a new bill payment arrangement	**Searching for a new payment recipient**
Step 1 You can set up a bill payment arrangement to most companies who've given you a unique customer reference number (this can usually be found on your bill). Paying other people To set up a payment to another person, please click here.	Step 1 You can set up a new payment for most companies who've given you a unique customer reference number (this can usually be found on your bill).
Step 2 If the company you want is not on the list below, and you entered the name exactly as it appears on your bill, you'll need to call us and we'll set up the bill payment arrangement for you.	Step 2 If the company you want is not on the list above, and you entered the name exactly as it appears on your bill, you can call us and we'll set up the bill payment arrangement for you. Otherwise click here to enter the account details yourself.
Step 3 Pay to: XYZYX When you want to make a payment to this company it will show as XYZYX. You may want to make a note of this. Please enter your unique customer reference number in both boxes below and click on 'Set up arrangement'. Your reference number should be on your bill. If not, please ask the company you want to pay. [Set up arrangement]	Step 3 Pay to: XYZYX When you want to make a payment to this company it will show as XYZYX. You may want to make a note of this. Please enter your unique customer reference number in both boxes below and click on 'Set up payment recipient'. Your reference number should be on your bill. If not, please ask the company you want to pay. [Set up payment recipient]
Your request has been successfully received	**Your request has been successfully received**
Your new bill payment has been set up successfully.	Your new payment recipient has been set up successfully.

Table F1.1 (continued). Language Alterations

Formal Language (Jargon)	Plain Language Version
Your payment details	**Your payment details**
Making a payment Fill in the amount and date you want the payment to be made. A date can only be specified up to a maximum of 31 days in the future. Please allow 4 working days for the payment to be received from the date specified. Re-enter your password and select 'Make payment'. Amending or cancelling a payment If the amount and due date fields are already completed you have a payment due. To change this payment, alter the amount or due date details, re-enter your password and select 'Change payment'. To delete this payment, re-enter your password and select 'Delete payment'. Pay to: XYZYX Sort code: 16-82-65 Account number: 455646547 Reference number (if any): 1234567890 Amount: £_____.___ Pence Date: * As soon as possible * Payment date / / Re-enter your password: [Make payment] [Change payment] [Delete payment]	Making a payment To make a new payment, specify the amount and date when you want the payment to be made, then re-enter your password and select 'Make payment'. (The date specified must be within the next 31 days.) Please allow 4 working days for the payment to be received from the date specified. Deleting this payment recipient To delete this recipient from your payment list, re-enter your password and select 'Delete payment'. Pay to: XYZYX Sort code: 16-82-65 Account number: 455646547 Reference number (if any): 1234567890 Amount: £_____.___ Pence Date to make payment: / / Frequency of payment: [drop down selector] Expiry: * Until further notice * Date / / Re-enter your password: [Make payment] [Delete payment]

Table F1.1 (continued). Language Alterations

Formal Language (Jargon)	Plain Language Version
Making a payment to another person	**Making a new payment**
Fill in all the details below. It may take up to 4 working days for the payment to be received from the date specified. Pay to: Sort code: - - Account number: Amount: £_____.___ Pence Re-enter your password: [Make payment]	Fill in all the details for the recipient below. The 'Reference' field is optional. It may take up to 4 working days for the payment to be received from the date when the payment is made. Pay to: Sort code: - - Account number: Reference (if any): Amount: £_____.___ Pence Date to make payment: / / Frequency of payment: [drop down selector] Expiry: * Until further notice * Date / / Re-enter your password: [Make payment]
Transferring money	**Transferring money**
Transfer money to: [drop down selector] Amount: £_____.___ Pence Re-enter your password: [Transfer]	Transferring money to another account Please specify the account to receive the money and the amount to transfer, then re-enter your password and click 'Transfer'. Transfer money from: This account (specified) Transfer money to: [drop down selector] Amount: £_____.___ Pence Re-enter your password: [Transfer]

Table F1.1 (continued). Language Alterations

Formal Language (Jargon)	Plain Language Version
Confirm Delete	**Confirm Delete**
Do you want to delete just this payment, or delete all details of this beneficiary from your payment list? * Delete only this payment * Delete the beneficiary from my payment list [OK] [Cancel]	Do you want to delete just this payment, or delete all details of this recipient from your payment list? * Delete only this payment * Delete the recipient from my payment list [OK] [Cancel]
Your request has been successfully received	**Your request has been successfully received**
The beneficiary has been deleted from your payment list successfully.	The recipient has been deleted from your payment list successfully.

Table F1.1 (continued). Language Alterations

F1.2. Personas

Participants used fictitious personal and bank account data in all the eBanking experiments. This allowed full control of the data offered, different accounts, statements, payment arrangements and all other aspects of the banking service being used. Participants all used the same persona login details, had the same balances and transactions and therefore their experiences could be compared.

F1.3. Tasks

Tasks were selected to represent real world tasks, presented in a convincing scenario including persona details, login information, account numbers for accessing the experimental interfaces. The tasks required making the initial arrangements to start making payments or transfers in online banking, concentrating on areas where language details had been changed.

Task Sheet 1:

1. You have decided to get cable TV and want to pay for it using online banking. Set up a payment to Telewest from your current account. Your reference number from your bill is 987456.
2. You have been saving some money for a new car. This month you've saved £200, arrange to transfer it from your current to your savings account.
3. Your daughter Sarah has passed all her exams and as a gift you've decided to send her £50. Use your savings to pay her account directly, the details she's given you are:
   Sort Code 12-45-72, Account number 87343205.

Task Sheet 2:

1. You are buying your friends old PC for £150. Pay the money to his account directly using your current account. The details he's given you are:
   Gavin Livingston, Sort Code 23-73-49, Account number 00457823.
2. You have been saving some money for home improvements. This month you're going to begin the work, arrange to transfer £250 from your savings to your current account to cover the first bills.
3. You have decided to pay your gas bills using online banking. Set up a payment to Scottish Gas from your current account. Your reference number from your bill is 2567985.

F1.4. Interview Questions

Now thinking about your preferences, I would like you to rate the two services you have used today. Place the markers along the scale between Best service and Worst service, I will record the order and positions. {note any comments}

What differences did you notice between the two services?

Show pictures of typical screens (showing language change in instructions). What aspects of the services did you like?

What aspects did you dislike?

Would you like to be able to send money to other accounts on the Internet Banking service? (Y/N/don't know)

F2: Qualitative Data

F2.1: Interview Responses

Asked what differences they noticed between the designs, participants commented:

- "There was very little difference between the two sites, mainly the wording of some of the options."
- "Version A had fewer options on the left, actual processes were fairly similar"
- "The ability to make a payment frequency choice option (PL)."
- "Language was more basic on version two (F) it seemed less formal and more accessible."
- "The second site (PL) was clearer, the information seemed to flow better from page to page and I could scan the page to pick out the information I needed without having to read it all"

What they liked about the designs:

- "I liked the range of facilities on offer."
- "The clearer wording in the first site (F)."
- "Liked that you could define the frequency of payments on one of the sites (PL)."
- "It wasn't confusing, did everything it had to."
- "Explained what to do when stuck (both)."
- "Instructions were really clear, liked that it asked if it was a one-off payment (PL)."

- "Very clear, logical."
- "I like the thought of just using the Internet rather than going to the bank."
- "Easy to set up bill payments - surprised. It had all the info you needed there which makes it a lot more comfortable."

What was disliked about the designs:

- "The buttons (make, amend and delete) looked too similar, helpful if they were different colours."
- "Difficult to read, text small."
- "Pop-up windows were not clearly worded."
- "Making spelling mistakes resulted in the company name not being matched. It would be helpful to add those which almost match my spelling so it's obvious I've made a typo, rather than thinking that the company doesn't exist."
- "Make it clearer what account you are using."
- "Not very obvious which menu option to choose."
- "I would like to be able to have a printout of all the transactions I have carried out over the Internet, so that I can keep track of what I have already done."
- "Payments and transfers should be separate."
- "I worry about making mistakes in entering details."
- "The left hand menu could be done in black rather than grey for greater contrast."
- "Not clear how to change to another account or which account you are in on the success page."

F3: Correlations and Questionnaire Structure

Correlations between attribute 'Use Again' (as Intent construct) and mean usability questionnaire score:

	Differences (F-PL)	Formal (F)	Plain Language (PL)
Pearson's *r*	.379	.483	.461
p	.042	.008	.012
N	29	29	29

Table F3.1. Attribute Use Again – Correlations with Mean Usability Questionnaire

Questionnaire Structure

The most appropriate language for transacting in eBanking was determined to be the Formal style. The correlation matrices for this design are shown with medium-high inter-item correlations (frequently highly significant), for semantically related attributes (N = 29).

		User-Friendly	Like Using	Enjoyment	Flustered	Stressful	In Control
User-Friendly	*r*	1	.484	.350	.479	.382	.674
	p	-	.008	.063	.009	.041	.000
Like Using	*r*	.484	1	.310	.351	.119	.363
	p	.008	-	.102	.062	.537	.053
Enjoyment	*r*	.350	.310	1	.151	.468	.327
	p	.063	.102	-	.434	.011	.083
Flustered	*r*	.479	.351	.151	1	.515	.508
	p	.009	.062	.434	-	.004	.005
Stressful	*r*	.382	.119	.468	.515	1	.482
	p	.041	.537	.011	.004	-	.008
In Control	*r*	.674	.363	.327	.508	.482	1
	p	.000	.053	.083	.005	.008	-

Table F3.1. Proposed Group – Affect

		Frustration with IA	Navigation Complication	Quickly Find	Procedure	Orientation
Frustration with IA	*r*	1	.217	.720	.375	.681
	p	-	.259	.000	.045	.000
Navigation Complication	*r*	.217	1	.243	.274	.231
	p	.259	-	.204	.150	.227
Quickly Find	*r*	.720	.243	1	.625	.876
	p	.000	.204	-	.000	.000
Procedure	*r*	.375	.274	.625	1	.624
	p	.045	.150	.000	-	.000
Orientation	*r*	.681	.231	.876	.624	1
	p	.000	.227	.000	.000	-

Table F3.2. Proposed Group – Structure

		Confusion (layout)	Page Clarity	Helpful	Concentration (reading)	Understood pages
Confusion (layout)	***r***	1	.610	.625	.579	.620
	p	-	.000	.000	.001	.000
Page Clarity	***r***	.610	1	.769	.547	.612
	p	.000	-	.000	.002	.000
Helpful	***r***	.625	.769	1	.617	.666
	p	.000	.000	-	.000	.000
Concentration (reading)	***r***	.579	.547	.617	1	.298
	p	.001	.002	.000	-	.116
Understood pages	***r***	.620	.612	.666	.298	1
	p	.000	.000	.000	.116	-

Table F3.3. Proposed Group – Content Design

Appendix G – Experiment 3: eStatements

G1: Experiment Materials

G1.1. Personas

Participants used fictitious personal and bank account data in all the eBanking experiments. This allowed full control of the data offered, different accounts, statements, payment arrangements and all other aspects of the banking service being used. Participants all used the same persona login details, had the same balances and transactions and therefore their experiences could be compared.

G1.2. Tasks

Task Sheet 1

Working with your Current Account, using the online Statement service, try the following tasks:

1. Using your statement, see how much you paid with your Direct Debit to onetel.co.uk in June 2002 and in February 2002.
2. Find out what date a debit card payment of £43.99 to John Lewis cleared your account.
3. Print out a copy of your statement for the period 5th – 27th April 2002, statement Sheet Number 17.
4. You remember paying the electrician roughly £50 in May 2002. Find the cheque (number 00475) in your statement and see what the exact amount was.
5. Print out a record of all the Direct Debits of £19.38 per month to Dixons from 1st May 2002 to 30th July 2002.

Task Sheet 2

Working with your Current Account, using the online Statement service, try the following tasks:

1. Using your statement, see how much you paid with your Direct Debit to Scottish Gas in May 2002 and in March 2002.
2. Find out what date a debit card payment of £56.99 to Waterstones cleared your account.
3. Print out a copy of your statement for the period 5th – 27th April 2002, statement Sheet Number 17.
4. You remember paying roughly £40 for a meal in May 2002. Find the cheque (number 00476) in your statement and see what the exact amount was.
5. Print out a record of all the Direct Debits of £15.99 per month to Currys from 1st April 2002 to 30th June 2002.

Task Sheet 3

Working with your Current Account, using the online Statement service, try the following tasks:

1. Using your statement, see how much you paid with your Direct Debit for Council Tax in July 2002 and in March 2002.
2. Find out what date a debit card payment of £35.99 to John Lewis cleared your account.
3. Print out a copy of your statement for the period 5th – 27th April 2002, statement Sheet Number 17.
4. You remember paying roughly £100 for a DVD player in May 2002. Find the cheque (number 00477) in your statement and see what the exact amount was.
5. Print out a record of all the Direct Debits of £10.99 per month to Telewest from 1st April 2002 to 30th June 2002.

G1.3. Interview Questions

Do you have any comments the different statement services you used today?

Which of the three versions of the statement service did you most prefer?

(A/B/C)

Using reminder screen-shots participants were asked for:

- Comments on the various search criteria they used in the tasks (for each interface)
- Comments on the printouts

G2: Combining the Questionnaires

Combining the Pilot Web Usability questionnaire, with the eBanking Usability Questionnaires from Experiments 1 and 2 resulted in an eStatements Usability Questionnaire.

In order to include the most effective attributes from the questionnaires utilised thus far in the Web banking research, the full pool of attributes from both versions of the Web-usability questionnaire were examined. A tally of important results from each experiment and interface resulted in a proposed list for the eStatements usability questionnaire. The tally included:

- Salient usability and interaction attributes from each interface evaluated – where scores were high on the attitude scale (typically above 5).
- Attributes which were significant differentiators between different interfaces evaluated in comparison.
- Attributes which significantly correlated with preference scores.

The resulting tallies are shown in the following table resulting in a summary of the questions selected for experiment three.

Banking Portals (Pilot A & B)	PA	PB	eBanking Transactions (Experiments 1 & 2)	E1	E2	E3	N?	eStatements (Experiment 3)
Navigation complication	4	1	Navigation complication	3	3	11	4	Navigation complication
Liked using	3	3	Liked using	3	1	10	4	Liked using
In control	3	2	In control	3	1	9	4	In control
Use again	3	2	Use again	3	1	9	4	Use again
Helpful	3	2	Helpful	2	1	8	4	Helpful
Page clarity	3	1	Page clarity	3	1	8	4	Page clarity
Orientation	2	1	Orientation	2	3	8	4	Orientation
Flustered	4	1	Flustered	2	1	8	4	Flustered
Understood pages	3	1	Understood pages	2	2	8	4	Understood pages
User-friendly	3	1	User-friendly	3	1	8	4	User-friendly
Frustration with IA	3	1	Frustration with IA	1	2	7	4	Frustration with IA
Enjoyment	2	1	Enjoyment	2	2	7	4	Enjoyment
Stress	1	2	Stress	3	1	7	4	Stress
Reliable	3	1	Reliable	1	1	6	4	Reliable
Quickly find	3	-	Quickly find	2	3	8	3	Quickly find
Improvement needed	4	2	Improvement needed	-	1	7	3	Improvement needed
Procedure	3	-	Procedure	1	-	4	2	Procedure
Confusion (layout)	3	-	Confusion (layout)	-	-	3	1	Confusion (layout)
Attractive	-	1	Attractive	-	-	1	1	Appearance
Text size	-	-	Text size	1	-	1	1	Text size
Concentration (reading)	-	1	Concentration (reading)	-	-	1	1	*Concentration (using)*
Clutter	2	2				4	.	Clutter
Matched Expectations	3	-				3	.	Matched Expectations
Trust	1	1				2	.	Trust
Graphics & pictures	-	-				-	.	
Link content	-	-				-	.	
Link visibility	-	-				-	.	
			Change arrangements	3	.	3	.	
			Words & phrases	.	2	2	.	
								Replace paper
								Convenience
								Suitability Printout
								Authenticity

Table G2.1. Analysis to Create the Final Usability Attitude Questionnaire for eStatements

G3: Search Logs – Data and Summary

Search behaviour on the two search engine design was analysed for patterns.

Task type	Method	Simple Search	Advanced Search
1	Used Search	36.1%	72.1%
	Used Paging	62.8%	26.8%
2	Used Search	81.4%	91.3%
	Used Paging	17.5%	7.7%
3	Used Search	47.5%	48.1%
	Used Paging	47.0%	45.9%
4	Used Search	36.1%	73.8%
	Used Paging	62.3%	25.1%
5	Used Search	84.7%	94.5%
	Used Paging	13.7%	4.9%

Table G3.1. Methods adopted to complete tasks for the two search designs

Table G3.1 shows the methods adopted in completion of the five banking tasks with search facilities on offer. Participants either performed tasks via the standard sheet-centric linear navigation method (paging) or by use of the offered search facility. The Advanced Search design encouraged searching more than the Simple Search design, which was used in combination with paging to complete tasks. The amount search (type 2) and printing a record of three sequential Direct Debits (type 5) drove most use of the two search features respectively (see task type details, p.197).

Finding the statement sheet (task type 3) was straightforward on the Simple Search variant where this type of search was one of the three options (refer back to Figure 6.2, p.192). In the Advanced Search it had to be performed using an inputted date range (refer back to Figure 6.3, p.193). This mismatch between the task wording (using the word "Sheet") and the advanced search interface (no sheet search specified) resulted in a drop in searches, participants reverting to paging to navigate the statements instead. Evidence gathered through observation and participant commentary indicated that users were unaware of the (paper-based) 'sheet' organisation. Participants talked about statements representing their finances in monthly chunks instead.

The logs from both search variants were also compared in terms of mean, mode and median searches performed for the five tasks. The optimum paths to search and find the information required for task completion were also mapped. Finally, the numbers of results returned were compared for each individual search. Table G3.2 compares the search logs for the two different design variants, Simple and Advanced.

Data gathered qualitatively indicated that when the Simple Search returned one result it was often not the result expected or required for task completion, whereas often when the Advanced Search returned a single result it did allow task completion. Using the Simple Search often only took a participant close to the statement item they were looking for, rarely exactly to it, resulting in more scanning and paging to complete the task.

Search Logs	Simple	Advanced
Search Use		
Mean no. of searches per participant	5.6	5.9
Standard deviation	3.65	2.35
Optimum no. of searches required per task sheet	6	5
Median no. of searches per participant	5	5
Modal no. of searches per participant	2, 4 and 6	4, 5 and 6
% participants using ten or more searches per task sheet	9%	6%
Results Returned		
Median no. of results per search	1	3
Modal no. of results per search	1	1
Maximum no. of results returned in a search	27	195
Minimum no. of results returned in a search	0	0
Mean no. of results returned per search	1.56	9.73
Standard deviation	1.72	18.9

Table G3.2. Comparison of Simple and Advanced Search Log Descriptive Statistics

Simple Search Log	Summary Data
Date	
Optimum number of Date searches to complete tasks	2
Percent of searches using Date option	37%
Mean Date searches per participant	2.1
Percent of Date searches used correctly	97%
Amount	
Optimum number of Amount searches to complete tasks	2
Percent used Amount Search option	41%
Mean Amount searches per participant	2.3
Percent of Amount searches used correctly	91%
Sheet	
Optimum number of Sheet searches to complete tasks	1
Percent used Sheet Search option	21%
Mean Sheet searches per participant	1.2
Percent of Sheet searches used correctly	97%

Table G3.3. Analysis of the Simple Search Logs

In terms of specific search criteria, a breakdown of the search method required for optimum task completion, and the corresponding data from the search logs is presented for the Simple Search in Table G3.3. Higher numbers of searches were performed in each category than expected (from optimum paths), particularly for the amount search. Mistakes made in using the search criteria were usually due to searching with blank fields, thus producing no results. However, for the amount search, some mistakes were also made in keying in the relevant amount, or placing the decimal point, thus accounting for the lower percent of correct uses of the amount search.

Task Type	Expected Path per task: Advanced Search
1	Date, Type and Particulars
2	Amount (exact), Type and Particulars
3	Date range only
4	Date, Amount (range), Type and Particulars
5	Date, Amount (exact), Type and Particulars

Table G3.4. Advanced Search criteria useful for task completion

For the Advanced Search, participants could choose a number of criteria to search in one query. The options useful for completion of each task type are presented in Table G3.4. A breakdown of the search method required for optimum task completion, and the corresponding data from the search logs is presented for the Advanced Search in Table G3.5.

The use of the search criteria mainly corresponds with the expected optimum paths for date and type. The amount search was used less frequently than expected, and the particulars much less than the four times expected for optimal use. Again, the same types of mistake were made in using the search criteria with blank fields and with keying in amounts. The type search (p.197) was the most frequently used search criterion, however in 6% of searches, type was selected with the default option 'all transactions' – equivalent to not selecting a type. Half of all type searches made use of the 'all direct debits' option (appropriate for 2 tasks), just less than 20% used 'all cheques' (appropriate for 1 task) and 'all debit card transactions' (appropriate for 1 task). Some users used more general categories, such as 'all debits'. Some inappropriate selections were made that did not correspond with the tasks.

The most interesting details from the log related to the particulars search. This criterion was used much less than expected for the five tasks (for optimal search). Most participants (144, 79%) chose to use a cheque number as a keyword for cheque searching tasks. Less than 2% of keywords entered into this field were not relevant to the tasks or statement items. Some 2% of keywords were mis-spelled. Only 6% of participants who used the particulars search specified an abbreviated form of the keyword. From observations and comments, there were indications of a mismatch between the field name 'particulars' and the concept of a keyword search. In addition, some participants commented that after inputting dates, amounts and perhaps a transaction type, typing into the particulars field seemed to be excessive. These two indicators could account for the lower than expected use of the particulars criterion.

Advanced Search Log	Summary Data
Date	
Optimum number of Date searches used to complete tasks	4
Percent of searches using Date option	62%
Mean Date searches per participant	3.7
Percent of Date searches used correctly	98%
Amount	
Optimum number of Amount searches used to complete tasks	3
Percent used Amount option	43%
Mean Amount searches per participant	2.6
Percent of Amount searches used correctly	87%
Type	
Optimum number of Type searches used to complete tasks	4
Percent used Type option	66%
Mean Type searches per participant	3.9
% of searches with appropriate Type selected	92%
% of searches with default (all types) selected	6%
Particulars	
Optimum number of Particulars searches used to complete tasks	4
Percent used Particulars option	43%
Mean Particulars searches per participant	2.5
Percent of Particulars searches used correctly	96%
% of Particulars searches misspelled	2%
% of Particulars searches using irrelevant keywords	2%
% of Particulars keywords input in abbreviated form	6%

Table G3.5. Analysis of the Advanced Search Logs

G4: Qualitative Data

G4.1: Interview Responses

General Comments, likes, dislikes and suggestions:

- Although a bit unfamiliar and tricky, Search B [AS] is definitely most versatile and in time would be very useful
- The drop downs included all types you would need to know, overall Search B [AS] was very well thought out
- I would need reassurances about legality of printouts
- I would like to see a 'Clear search' button and a drop down menu for the date field –

or pop up calendar, also a drop down menu for the amount range

- Search A [SS] here is not adequate to make me move any further away from receiving a paper statement. It is more frustrating to have a service that is not up to standard than to have no search facility at all!
- If service B [AS] was available, I would register immediately
- The printout should look like it came from the screen - footer etc. should show it came from the net
- I would only use service if statement stood up in court of law, were admissible for mortgage applications. Local printout must be printer friendly, all you want is content, non-wasteful colours, not loads of ink
- The suppress button should not feature where it is located - it also shouldn't be a button as it doesn't fit in with any activity on the page, not part of the overall service
- Prefer the ability to choose multiple criteria [AS].
- The search was good, but it would be nice to click on Direct Debits and Standing Orders on the relevant pages and see a history of payments - easier than searching.
- It would be easier if the date [on the statement display] had the month name written.
- Search for all Internet payments made using the online service.
- I preferred this search [AS] a lot of thought has obviously been put into that one.
- [AS] definitely most versatile and in time would be very useful.
- I would definitely use services like these.

Comments regarding search criteria:

- "Particulars was the most handy, sheet search seems useless."
- "I didn't like the name 'particulars', don't know what it is until you try it."
- "'Particulars' is bank speak - would prefer 'keyword'."
- "I would only change 'particulars' for 'description'."

Comments regarding sheet metaphor:

- "The next, previous and recent buttons are a good method of paging through."
- "It was annoying having only one, or very few items on some sheets."
- "The sheet number is nonsense - it means nothing. You don't see money in chunks or pages, just as a continuous list."
- "1st of the month - 1st of the month. At the moment, don't know where to look at dates don't run sensibly. Your pay and other things come in calendar months - so think that way more."
- "A sheet per calendar month would require less thought and less scanning for the correct place."

Comments on printouts:

- "All you need is content, don't need it to be too formal."
- "As long as it's easy to read it doesn't matter, this one is fine."
- "As long as the figures are right, that's the only important thing."
- "It should be official-looking like a paper statement."
- "Generally need a statement as proof of address etc. so it would need to look like the postal statements otherwise it's not much use."
- "It's not acceptable as proof of identity I don't think - I would prefer to print out in the classic style."
- "It looks way too basic - like excel, like you knocked it up yourself."
- "Got enough info there to print and file for a record."

G4.2: Think Aloud Comments and Observations

The comments and observations made during the hands-on session are grouped into themes.

General comments about the interfaces were positive, although many said they thought the 'search' button should have been at the top of the statement page (it was at the bottom). Some also thought the 'print' button should have been with the next/previous buttons so as to keep all the functions together.

Some subjects thought the online statement was inadequate for what they wanted to be able to do online with their bank- they wanted to have a Financial planner or Money manager that would allow them to categorise the main areas of their accounts, e.g.: Utilities, Bills, Car, Mortgage, etc. thus personalising their account into the categories that made sense to them.

Paper statements

There were several participants who came in with the opinion that they couldn't imagine life without paper statements for security reasons, proof of account activity and proof of purchase/ transactions. Some of these participants did voice however that this was an excellent 'value added service' that the bank should be providing anyway in order to keep up with what the competition are doing and what technology is capable of.

At the other end of the spectrum of customer opinion there were some people that said they "hate" paper statements, don't see the point in them, they are a waste, and they are thrown in the bin unopened most of the time. These were typically users who were younger in age and did have an online account or regularly monitored what was going on in their account by other means, either by telephone or by gaining an ATM print out.

Data Only Version

Due to not having used the "previous page" button on the live online statement (because typically participants only had one page of statement activity available online at that time) many participants had to search quite hard in order to find out how to go back a page. Most did find their way around after that initial hesitation

In general, those who were most frustrated with using the statement history (with no search) were those who had previously experienced one or both searchable services.

- "I just get the most recent statement on my online service, I have never been able to do this – it's really good and I've always liked the site in general, but I could use it more as I can just look online and get the answer straight away".
- "Using the next and previous buttons is quick and easy, you know where you are and how to get to another page"
- "I don't really miss not having the search"
- "very simple to use"
- "Using the previous button is faster and easier than searching"
- "Would be easier if pages were month to month"
- "Could also be easier to use a dropdown to select a particular month instead of repeated use of the previous button to get back to a specific page"
- "No point to that, you're as quick using paper"
- "That wasn't user-friendly, would struggle to get groups of info together, search would be helpful"
- "Can't search or isolate individual items"

Simple Search

A great many customers did not get on well with this search as they found it did little to aid them and they "wasted" their time in trying desperately to make it find something (anything) for them. This is because many people mistook the choice of one search method to be a combination of all three.

Most customers saw searching for a "Sheet number" as being totally useless.

A lot of people resorted to the Next and Previous buttons instead of trying to use the search facility, although others persevered and occasionally had to give up on tasks. The behaviour was to typically abandon the search facility unless they had some concrete information to enter into it, like an exact amount.

Confusion would often occur with this search facility when using the 'date' method of searching. People were confused with the expression "on or prior to" (saying things like – do they mean "before"?), the return of just one transaction from date searches was also confusing to them. The general expectation was that this should return pages of results and not just a single transaction from the date specified.

A lot of customers were also confused by the "OR" nature of this search, and tried to either delete other fields they had used previously - sometimes after selecting and inputting the new criteria, the result being that the blanked field was selected when they pressed find and therefore no results were returned. Others tried to input multiple criteria - whether or not they had previously seen the advanced search.

- "The 'Search' here is not adequate to make me move any further away from receiving a paper statement. It is more frustrating to have a service that is not up to standard than to have no search facility at all!"
- "The Amount search worked – and made more sense than date searches"
- "That was rubbish! You've got to be able to put all the info into a search, they should try to fit the search into the top bit of each statement"
- "You could only use one option at a time, so couldn't give 2 areas, limiting choice. I am confused - these search results make no sense"
- "It would be much easier to put in the company name"
- "What I want to ask for is 'a sheet beginning 1st Feb'"
- "A date range would be useful here"
- "The search needs more options"
- "I was expecting something to allow me to search between two dates or from one date and move forwards, 'on or prior to' is hard to get your head round"
- "You don't typically think of sheet numbers, think in terms of dates."
- "It would be better to have something to pick a month and year, as you don't always know the date anyway as it takes time to clear, you're more likely to be able to select a month"
- "It's easier and faster to use the Next and Previous buttons than this search"

Advanced Search

Without question, the Advanced search facility was favoured greatly by the participants. Crucially this was given frequently as a caveat when asked at the end whether they were more or less inclined to opt out of paper statements.

This version tended to return multiple results, and on occasion multiple pages of results. Multiple results pages used the same "next", "previous" etc. buttons at the top – making it too similar to the statement for some people.

Overall, the reactions to the complex search were much more favourable than the other two services. Most customers seemed to be happier to use the different criteria available, mainly intuitively – none seemed to work their way through the instructions.

The problems associated with the paging of long lists of search results were the most troubling for participants, and the suggestions given tend to ask for the results to be formatted in a scrolling list.

- "'Particulars' was a bit confusing - it should be 'Name' or 'Text' search"
- "Needs a drop down menu for the date field, or a pop up calendar"
- "Really good – I would definitely use it; you can't always get to the bank, this makes a real difference"
- "This search was quite good, obviously helpful. I didn't see much difference except that being able to write in the name of what you were looking for was very useful"
- "It would be best to remove 'all' from each item in type drop-down menu - that would make it easier to scan the list alphabetically"
- "That one is the best. Customers are more sophisticated nowadays - they want more sophisticated services. Also, they are used to getting tools like this elsewhere and expect it."
- "You can key in as much info as you have available - it is more refined"
- "This is more like an internet search - much better to have a multiple search criteria"
- "The type 'cheques in/out' is useful"
- "Great, that is much less hassle than paper"

Printouts

Some participants did not want to print out the statements themselves, as the charge would fall on them for paper and printing.

The biggest concern among customers was whether other banks, credit card companies, and mortgage lenders would accept the printout. In this the print out was deemed to be a bit basic and one that "you could make yourself".

Many people realised that given the right software even the most complex of formats could be imitated - they therefore wanted it to say somewhere on the statement that this was a valid form of proof and needed the reassurance that it would be accepted by others.

One suggestion was the use of an authentication code that could be printed out by the customer using their online service then registered with the bank in some way.

Opting Out

Most participants considered that Internet Services should be free. Most noticed that the bank could save a lot of money by stopping sending paper statements, several mentioned how much they would save if they also stopped sending junk mail. For some, this was a good reason to stop getting paper statements – but only if the junk mail stopped as well.

Some also did not want charges and other letters sent to them via the post, and suggested email as an alternative.

Many saw that the bank were making less work for themselves by getting the customer to have more responsibility and making them pay for the paper and ink of their own copier. They saw this to mean that the bank didn't have to pay for the man-hours of their workers or the cost of the paper and postage – and expected some kind of monetary reward (in the form of interest rates or another incentive) for switching to the online statement service.

Some customers also wanted to receive a paper statement once in a while too – possibly every 6 months. Those customers still not interested in stopping their paper statements were a little concerned that this would become mandatory. A section of largely eBankers expressed an interest in 'paperless banking'.

The older participants understood the importance (and commented typically at length) of holding on to a paper statement for the tax return of 7 years evidence of transactions. There were some self employee younger participants that also voiced this.

G5: Attitude–Intention Correlations

G5.1: Scatter Plots – Per Interface

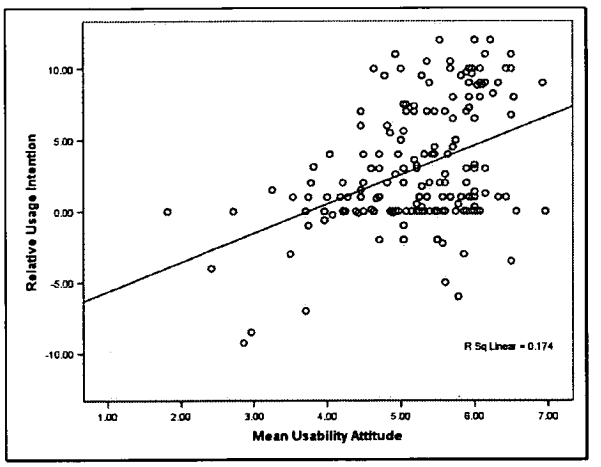


Figure G5.1a: Advanced Search Design – Usability vs. Relative Intention

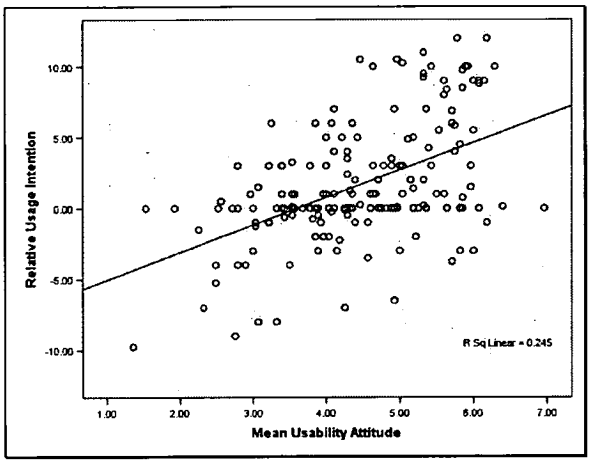


Figure G5.1b: Simple Search Design – Usability vs. Relative Intention

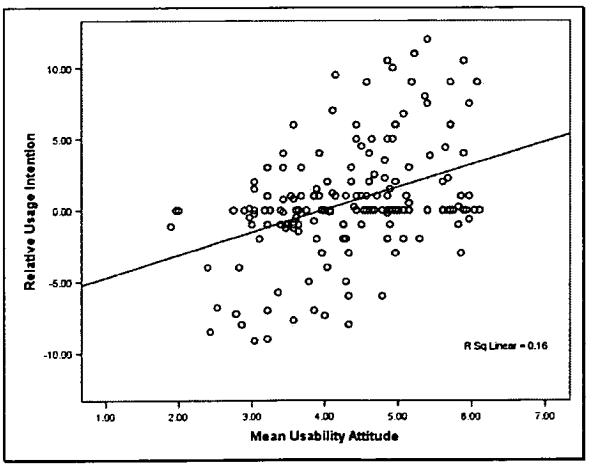


Figure G5.1c: Data Only Design – Usability vs. Relative Intention

G5.2: Scatter Plots – Interface Pairs

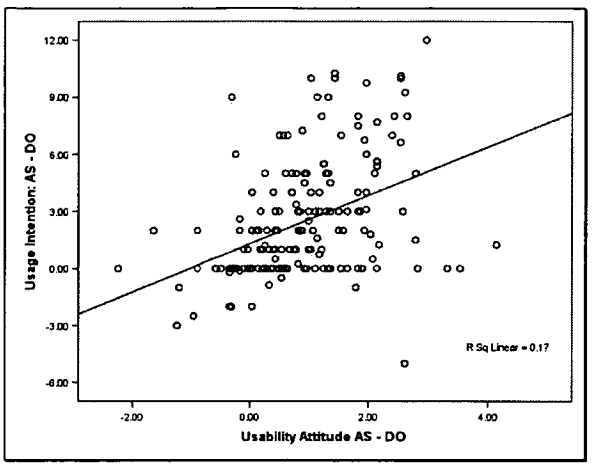


Figure G5.2a: Advanced Search – Data Only Designs

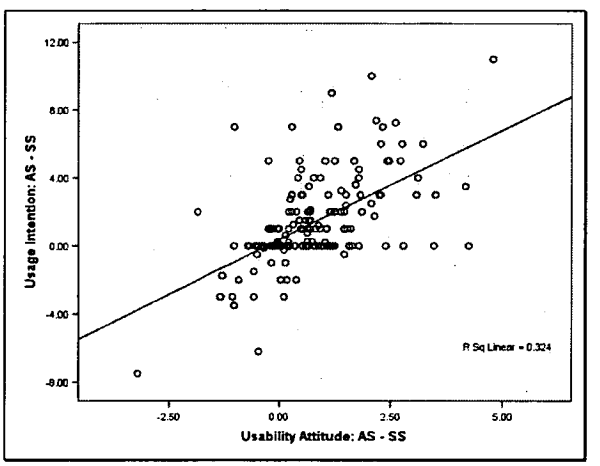


Figure G5.2b: Advanced Search – Simple Search Designs

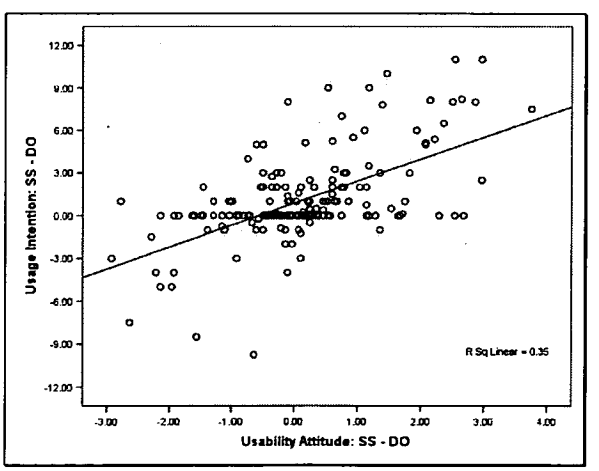


Figure G5.2c: Simple Search – Data Only Designs

G5.3: Interfaces Pairs Correlations

In addition to the absolute correlations (and significance levels) for each eStatement interface between attitude and intention, the differences between pairs were also studied:

Attribute	Advanced-Simple		Advanced-Data Only		Simple-Data Only	
	r	*p*	*r*	*p*	*r*	*p*
convenience	**.529**	**<.001**	**.309**	**<.001**	**.472**	**<.001**
like using	**.507**	**<.001**	**.419**	**<.001**	**.585**	**<.001**
match expectations	**.501**	**<.001**	**.375**	**<.001**	**.464**	**<.001**
quickly find	**.497**	**<.001**	**.334**	**<.001**	**.432**	**<.001**
user-friendly	**.486**	**<.001**	**.339**	**<.001**	**.481**	**<.001**
replace paper	**.484**	**<.001**	**.308**	**<.001**	**.532**	**<.001**
flustered	.475	<.001	.285	<.001	.379	<.001
navigation complication	.463	<.001	.235	.002	.357	<.001
understood pages	.459	<.001	.202	.007	.174	.020
use again	**.445**	**<.001**	**.419**	**<.001**	**.550**	**<.001**
stress	.439	<.001	.274	<.001	.425	<.001
in control	**.437**	**<.001**	**.362**	**<.001**	**.548**	**<.001**
frustration with IA	.429	<.001	.247	.001	.428	<.001
improvement needed	**.424**	**<.001**	**.428**	**<.001**	**.585**	**<.001**
enjoyment	**.422**	**<.001**	**.362**	**<.001**	**.544**	**<.001**
concentration	.418	<.001	.270	<.001	.420	<.001
appearance	.408	<.001	.061	.417	.332	<.001
confusion (layout)	.393	<.001	.184	.014	.319	<.001
reliable	.387	<.001	.162	.030	.330	<.001
helpful	**.382**	**<.001**	**.348**	**<.001**	**.464**	**<.001**
procedure	.349	<.001	.145	.054	.282	<.001
page clarity	.302	<.001	.173	.021	.267	<.001
clutter	.273	<.001	.101	.179	.167	.026
orientation	.259	<.001	.214	.004	.302	<.001
trust	.151	.044	.011	.885	.187	.012
suitability	.142	.059	.158	.035	.279	<.001
text size	.075	.318	.093	.215	.079	.296
printout authenticity	.071	.345	.017	.825	.082	.279
mean usability	***.569***	***<.001***	***.413***	***<.001***	***.592***	***<.001***

Table G5.3. Differences Between Interface Pairs: Attitude – Intention Relationship

G6: Individual Attribute Correlations with Mean Scores

In previous experiments, the usability satisfaction scores have been related to preference ratings as an indicator of their relationship with participants' real world selections. In other words, preferences were used as estimators of intent. In this experiment the aim was to measure intent directly. Therefore it was of interest to note how the individual attributes related to mean usability metrics, and which were the most salient in being associated with overall usability ratings.

As can be seen from the table, almost all the individual attributes were highly significantly and positively correlated with the overall mean score for each interface. For three attributes however, the correlation coefficient dropped below r = .5, equivalent to a strong effect, where at least 25% of the variance in the overall score could be attributed to variance in the attribute score. The three less salient qualities of the interfaces were therefore considered to be:

- Trust
- Text size
- Authenticity of printout

Overall however, the evidence is that all the individual attributes were significantly correlated with overall mean scores – consistent with the reliability analysis of Cronbach's alpha (p.37) which determined no strong candidates for deletion. Of further interest were the highest correlations across the three designs: Frustration with IA, Like using, In control, Enjoyment, User-friendly and Use again were the most highly correlating with the overall mean usability scores. Generally, these characteristics of the interface and perception of the interaction explained over 48% of the variance in the mean usability scores.

Attribute	Advanced Search (AS)		Simple Search (SS)		Data Only (DO)	
	r	*p*	*r*	*p*	*r*	*p*
Frustration with IA	**.815**	**<.001**	**.868**	**<.001**	**.780**	**<.001**
Like using	**.812**	**<.001**	**.883**	**<.001**	**.855**	**<.001**
Understood pages	.801	<.001	.744	<.001	.669	<.001
In control	**.792**	**<.001**	**.850**	**<.001**	**.834**	**<.001**
Flustered	.784	<.001	.713	<.001	.676	<.001
Navigation complication	.777	<.001	.783	<.001	.647	<.001
Stress	.774	<.001	.749	<.001	.683	<.001
Confusion (layout)	.769	<.001	.764	<.001	.609	<.001
Enjoyment	**.769**	**<.001**	**.864**	**<.001**	**.827**	**<.001**
Page clarity	.735	<.001	.658	<.001	.611	<.001
User-friendly	**.727**	**<.001**	**.843**	**<.001**	**.809**	**<.001**
Orientation	.716	<.001	.665	<.001	.670	<.001
Appearance	.710	<.001	.693	<.001	.617	<.001
Concentration (using)	.701	<.001	.749	<.001	.681	<.001
Procedure	.701	<.001	.639	<.001	.528	<.001
Use again	**.694**	**<.001**	**.814**	**<.001**	**.719**	**<.001**
Helpful	.683	<.001	.771	<.001	.687	<.001
Improvement needed	.669	<.001	.818	<.001	.724	<.001
Convenient	.627	<.001	.671	<.001	.578	<.001
Quickly find	.626	<.001	.766	<.001	.557	<.001
Clutter	.614	<.001	.555	<.001	.449	<.001
Replace paper	.609	<.001	.666	<.001	.604	<.001
Match Expectations	.579	<.001	.688	<.001	.578	<.001
Format (printout)	.517	<.001	.521	<.001	.431	<.001
Reliable	.514	<.001	.576	<.001	.576	<.001
Trust	.378	<.001	.395	<.001	.262	<.001
Text size	.347	<.001	.334	<.001	.231	.002
Authenticity (printout)	.268	<.001	.327	<.001	.231	.002

Table G6.1.Significant Correlations between Usability Attributes and Mean Usability Scores

G7: Correlations and Questionnaire Structure

The most appropriate eStatement service interface was an Advanced Search design. The correlation matrices for this design are shown. The matrices show high inter-item correlations (very highly significant), for semantically related attributes (N = 178).

	User-friendly	Like using	Enjoyment	Use again	Helpful	Match expectations	Replace paper	Convenient
User-friendly	1.000	.638	.572	.469	.472	.381	.409	.466
Like using	.638	1.000	.807	.684	.628	.521	.535	.574
Enjoyment	.572	.807	1.000	.646	.586	.453	.523	.542
Use again	.469	.684	.646	1.000	.659	.396	.469	.544
Helpful	.472	.628	.586	.659	1.000	.390	.443	.414
Match expectations	.381	.521	.453	.396	.390	1.000	.398	.479
Replace paper	.409	.535	.523	.469	.443	.398	1.000	.685
Convenient	.466	.574	.542	.544	.414	.479	.685	1.000

Table G7.1. Proposed Group – Fulfilment

Correlations all significant at $p < .001$

	Flustered	Stress	Orientation	Procedure	Concentration (using)	Navigation complication	Control	Frustration with IA	Quickly find	Confusion (layout)	Understood pages
Flustered	1.000	.831	.610	.576	.639	.670	.665	.633	.520	.610	.599
Stress	.831	1.000	.513	.488	.658	.650	.676	.630	.564	.528	.630
Orientation	.610	.513	1.000	.570	.517	.578	.601	.564	.491	.566	.510
Procedure	.576	.488	.570	1.000	.521	.567	.601	.540	.405	.593	.482
Concentration (using)	.639	.658	.517	.521	1.000	.602	.526	.524	.495	.516	.510
Navigation complication	.670	.650	.578	.567	.602	1.000	.618	.706	.432	.610	.643
Control	.665	.676	.601	.601	.526	.618	1.000	.614	.569	.547	.666
Frustration with IA	.633	.630	.564	.540	.524	.706	.614	1.000	.466	.736	.724
Quickly find	.520	.564	.491	.405	.495	.432	.569	.466	1.000	.500	.461
Confusion (layout)	.610	.528	.566	.593	.516	.610	.547	.736	.500	1.000	.607
Understood pages	.599	.630	.510	.482	.510	.643	.666	.724	.461	.607	1.000

Table G7.2. Proposed Group – Interaction

Correlations all significant at $p < .001$

	Clutter	Page clarity	Appearance	Improvement needed
Clutter	1.000	.590	.551	.420
Page clarity	.590	1.000	.639	.552
Appearance	.551	.639	1.000	.547
Improvement needed	.420	.552	.547	1.000

Table G7.3. Proposed Group – Visual Design

Correlations all significant at $p < .001$

	Format (printout)	Reliable	Trust
Format (printout)	1.000	.397	.362
Reliable	.397	1.000	.348
Trust	.362	.348	1.000

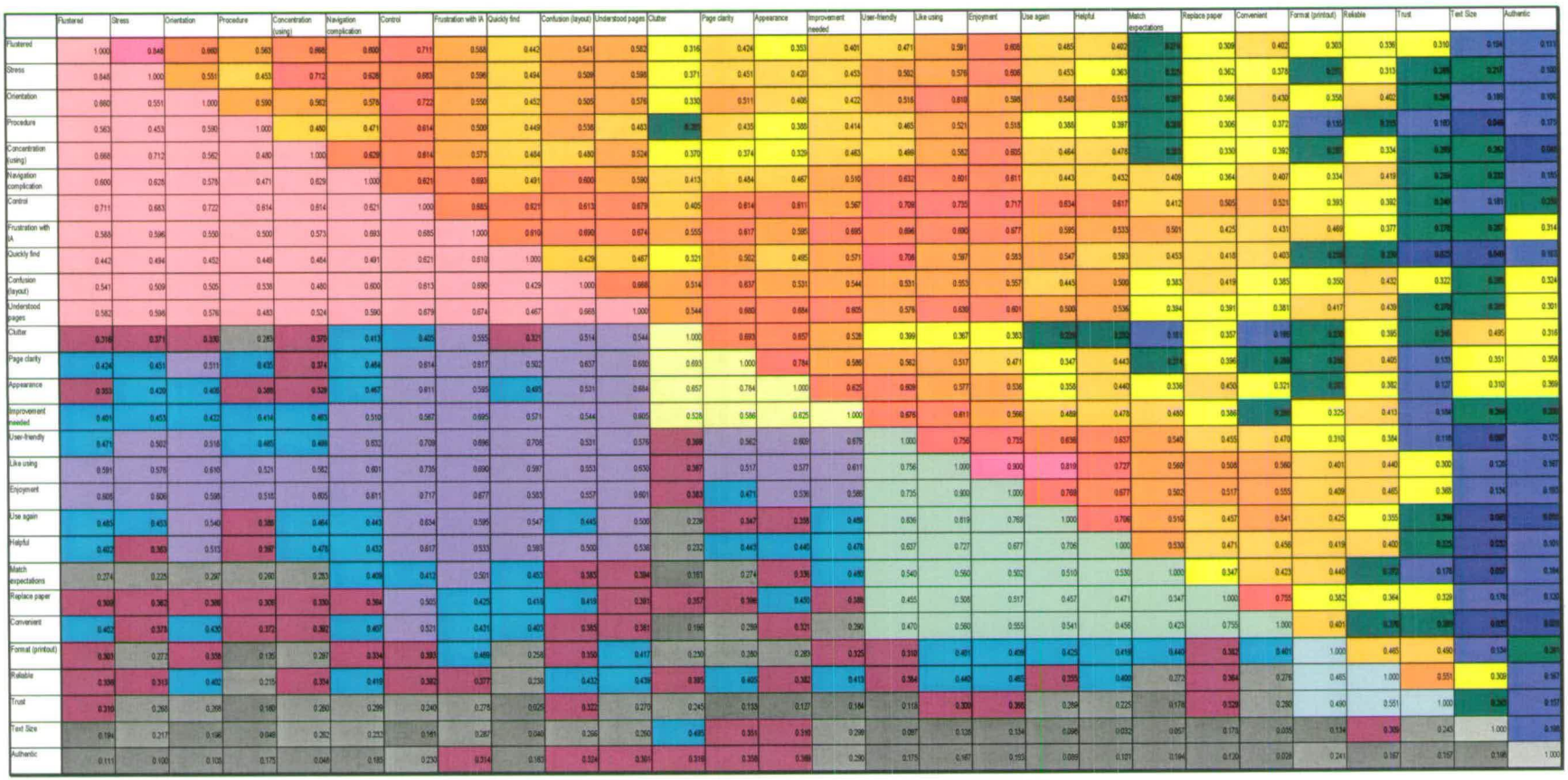
Table G7.4. Proposed Group – Integrity

Correlations all significant at $p < .001$

G7.1: Full Correlation Matrices

An annotated, full correlation matrix for the usability questionnaire items for the Advanced Search Interface is shown in Table G7.5.

Similarly, an annotated correlation matrix for the mean of the usability questionnaire items for the three eStatement Interfaces when they are all combined is shown in Table G7.6.



	Flustered	Stress	Orientation	Procedure	Concentration (using)	Navigation complication	Control	Frustration with IA	Quickly find	Confusion (layout)	Understood pages	Clutter	Page clarity	Appearance	Improvement needed	User-friendly	Like using	Enjoyment	Use again	Helpful	Match expectations	Replace paper	Convenient	Format (printout)	Reliable	Trust	Text Size	Authentic
Flustered	1.000	0.848	0.660	0.563	0.668	0.600	0.711	0.588	0.442	0.541	0.582	0.316	0.424	0.353	0.401	0.471	0.591	0.605	0.485	0.402	0.274	0.309	0.402	0.303	0.336	0.310	0.194	0.111
Stress	0.848	1.000	0.551	0.453	0.712	0.628	0.663	0.596	0.494	0.509	0.596	0.371	0.451	0.420	0.453	0.502	0.576	0.606	0.453	0.363	0.225	0.362	0.378	0.291	0.313	0.269	0.217	0.100
Orientation	0.660	0.551	1.000	0.590	0.562	0.578	0.722	0.550	0.452	0.505	0.576	0.330	0.511	0.406	0.422	0.516	0.610	0.598	0.549	0.513	0.287	0.366	0.432	0.358	0.402	0.299	0.195	0.106
Procedure	0.563	0.453	0.590	1.000	0.480	0.471	0.614	0.500	0.449	0.536	0.483	0.293	0.435	0.360	0.414	0.465	0.521	0.518	0.388	0.397	0.288	0.306	0.372	0.192	0.315	0.183	0.048	0.175
Concentration (using)	0.668	0.712	0.562	0.480	1.000	0.628	0.614	0.573	0.484	0.480	0.524	0.379	0.374	0.329	0.483	0.466	0.582	0.605	0.464	0.478	0.288	0.330	0.392	0.280	0.334	0.269	0.362	0.040
Navigation complication	0.600	0.628	0.578	0.471	0.629	1.000	0.621	0.693	0.491	0.600	0.590	0.413	0.484	0.467	0.510	0.632	0.601	0.611	0.443	0.432	0.409	0.364	0.407	0.334	0.419	0.294	0.222	0.153
Control	0.711	0.663	0.722	0.614	0.614	0.621	1.000	0.685	0.621	0.613	0.679	0.405	0.614	0.611	0.567	0.706	0.735	0.717	0.634	0.617	0.412	0.505	0.521	0.393	0.392	0.349	0.191	0.259
Frustration with IA	0.588	0.596	0.550	0.500	0.573	0.693	0.685	1.000	0.610	0.690	0.674	0.555	0.617	0.595	0.695	0.696	0.690	0.677	0.595	0.533	0.501	0.425	0.431	0.469	0.377	0.376	0.287	0.314
Quickly find	0.442	0.494	0.452	0.449	0.484	0.491	0.621	0.610	1.000	0.429	0.467	0.321	0.502	0.495	0.571	0.706	0.597	0.553	0.547	0.593	0.453	0.410	0.403	0.258	0.336	0.825	0.548	0.163
Confusion (layout)	0.541	0.509	0.505	0.536	0.480	0.600	0.613	0.690	0.429	1.000	0.666	0.514	0.637	0.531	0.544	0.531	0.553	0.557	0.445	0.500	0.383	0.418	0.385	0.350	0.432	0.322	0.285	0.324
Understood pages	0.582	0.596	0.576	0.483	0.524	0.590	0.679	0.674	0.467	0.666	1.000	0.544	0.680	0.684	0.605	0.576	0.636	0.601	0.500	0.536	0.394	0.391	0.381	0.417	0.429	0.376	0.260	0.301
Clutter	0.316	0.371	0.330	0.293	0.379	0.413	0.405	0.555	0.321	0.514	0.544	1.000	0.693	0.657	0.528	0.399	0.367	0.363	0.239	0.232	0.161	0.357	0.186	0.235	0.395	0.346	0.495	0.316
Page clarity	0.424	0.451	0.511	0.435	0.374	0.484	0.614	0.617	0.502	0.637	0.680	0.693	1.000	0.784	0.586	0.582	0.517	0.471	0.347	0.443	0.314	0.396	0.286	0.260	0.405	0.155	0.351	0.358
Appearance	0.353	0.420	0.406	0.360	0.329	0.467	0.611	0.595	0.495	0.531	0.684	0.657	0.784	1.000	0.625	0.609	0.577	0.536	0.358	0.440	0.336	0.455	0.321	0.287	0.382	0.122	0.310	0.359
Improvement needed	0.401	0.453	0.422	0.414	0.483	0.510	0.567	0.695	0.571	0.544	0.605	0.528	0.586	0.625	1.000	0.676	0.611	0.566	0.488	0.470	0.460	0.386	0.298	0.325	0.413	0.184	0.294	0.200
User-friendly	0.471	0.502	0.516	0.465	0.466	0.632	0.709	0.696	0.706	0.531	0.576	0.399	0.582	0.609	0.676	1.000	0.756	0.735	0.636	0.637	0.540	0.455	0.470	0.310	0.364	0.116	0.087	0.175
Like using	0.591	0.576	0.610	0.521	0.582	0.601	0.735	0.690	0.597	0.553	0.630	0.367	0.517	0.577	0.611	0.756	1.000	0.900	0.819	0.727	0.560	0.506	0.560	0.401	0.440	0.300	0.126	0.167
Enjoyment	0.605	0.606	0.598	0.518	0.605	0.611	0.717	0.677	0.583	0.557	0.601	0.363	0.471	0.536	0.588	0.735	0.900	1.000	0.768	0.677	0.502	0.517	0.555	0.409	0.465	0.369	0.174	0.193
Use again	0.485	0.453	0.549	0.388	0.464	0.443	0.634	0.595	0.547	0.445	0.505	0.239	0.347	0.358	0.488	0.636	0.819	0.769	1.000	0.706	0.510	0.457	0.541	0.425	0.355	0.299	0.080	0.050
Helpful	0.402	0.363	0.513	0.397	0.478	0.432	0.617	0.533	0.383	0.500	0.536	0.232	0.443	0.446	0.478	0.637	0.727	0.677	0.706	1.000	0.530	0.471	0.456	0.419	0.400	0.225	0.052	0.101
Match expectations	0.274	0.225	0.287	0.260	0.283	0.409	0.412	0.501	0.453	0.585	0.384	0.161	0.274	0.336	0.480	0.540	0.560	0.502	0.510	0.530	1.000	0.347	0.423	0.440	0.372	0.178	0.057	0.164
Replace paper	0.309	0.362	0.366	0.306	0.330	0.364	0.505	0.425	0.410	0.418	0.391	0.357	0.396	0.430	0.386	0.455	0.506	0.517	0.457	0.471	0.347	1.000	0.755	0.382	0.364	0.329	0.178	0.133
Convenient	0.402	0.378	0.430	0.372	0.392	0.407	0.521	0.431	0.402	0.385	0.381	0.186	0.286	0.321	0.290	0.470	0.560	0.555	0.541	0.456	0.423	0.755	1.000	0.401	0.376	0.289	0.055	0.059
Format (printout)	0.303	0.273	0.358	0.135	0.287	0.334	0.393	0.469	0.258	0.350	0.417	0.230	0.380	0.289	0.325	0.310	0.481	0.439	0.425	0.418	0.440	0.382	0.401	1.000	0.465	0.490	0.154	0.291
Reliable	0.336	0.313	0.402	0.219	0.354	0.419	0.392	0.377	0.238	0.432	0.439	0.395	0.405	0.382	0.413	0.364	0.440	0.405	0.355	0.400	0.272	0.364	0.276	0.465	1.000	0.551	0.309	0.167
Trust	0.310	0.269	0.299	0.160	0.269	0.299	0.240	0.275	0.025	0.322	0.270	0.245	0.159	0.127	0.184	0.118	0.309	0.390	0.388	0.225	0.178	0.329	0.282	0.490	0.551	1.000	0.265	0.157
Text Size	0.194	0.217	0.196	0.048	0.262	0.253	0.161	0.267	0.040	0.285	0.260	0.495	0.351	0.310	0.299	0.087	0.126	0.134	0.096	0.032	0.057	0.178	0.035	0.134	0.309	0.243	1.000	0.186
Authentic	0.111	0.100	0.106	0.175	0.040	0.185	0.230	0.314	0.163	0.324	0.301	0.316	0.358	0.359	0.290	0.175	0.167	0.153	0.089	0.101	0.164	0.120	0.028	0.241	0.167	0.157	0.186	1.000

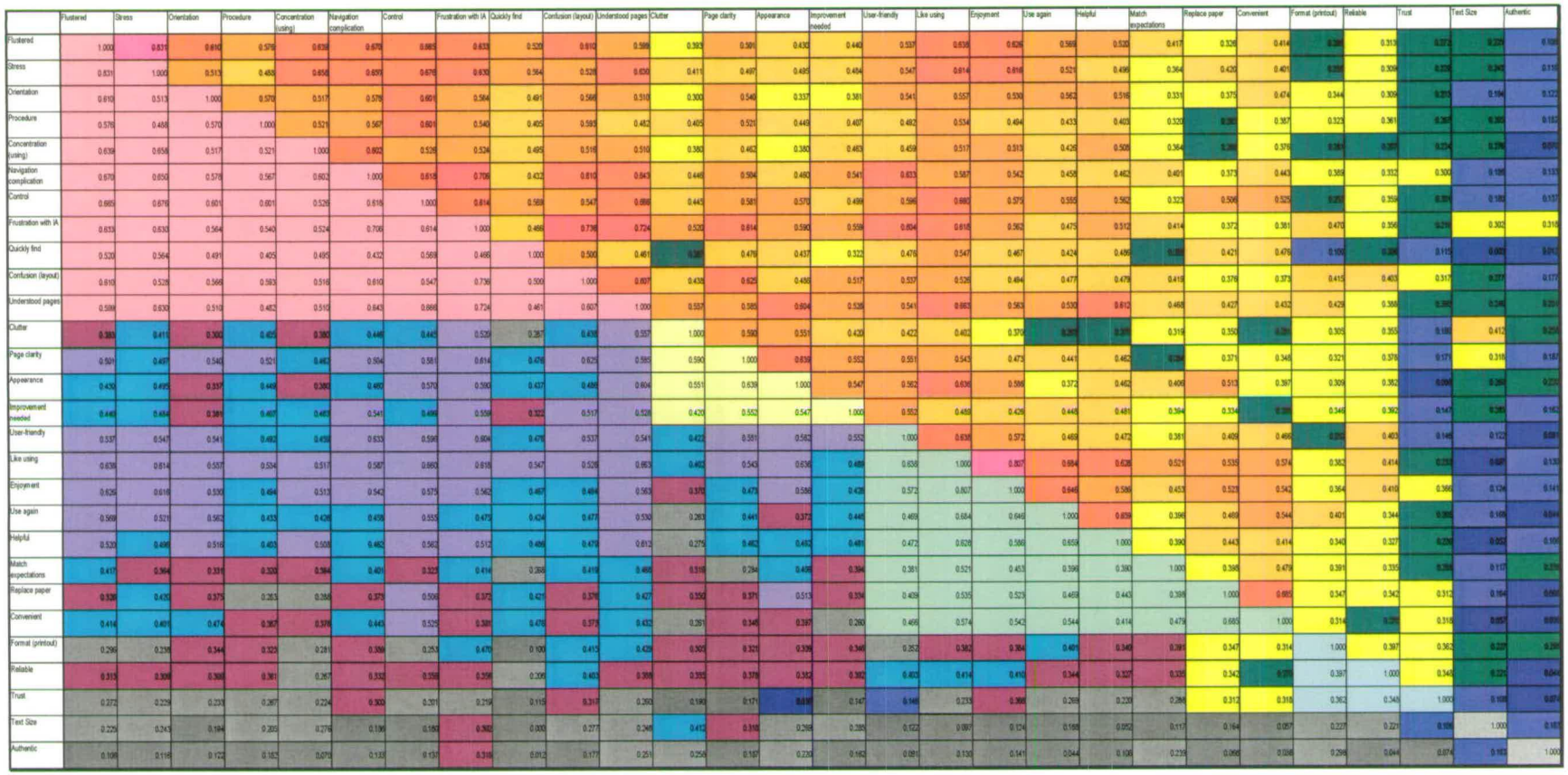
Table G7.5. Advanced Search Correlation Matrix

KEY

Correlation (*r*) magnitude

Top-Right - workings	> .8	>.6	>.5	>.4	>.3	>.2	>.1	≤.1
Bottom-Left - residuals			>.5	>.4	>.3	>.2	≤.2	

Proposed Groups	Interaction	Fulfilment	Visual Design	Integrity	NOT Grouped



	Flustered	Stress	Orientation	Procedure	Concentration (using)	Navigation complication	Control	Frustration with IA	Quickly find	Confusion (layout)	Understood pages	Clutter	Page clarity	Appearance	Improvement needed	User-friendly	Like using	Enjoyment	Use again	Helpful	Match expectations	Replace paper	Convenient	Format (printout)	Reliable	Trust	Text Size	Authentic
Flustered	1.000	0.831	0.610	0.576	0.639	0.670	0.665	0.633	0.520	0.610	0.599	0.383	0.501	0.435	0.446	0.537	0.638	0.626	0.569	0.520	0.417	0.326	0.414	0.286	0.313	0.272	0.225	0.108
Stress	0.831	1.000	0.513	0.488	0.656	0.650	0.676	0.630	0.564	0.526	0.630	0.411	0.497	0.495	0.484	0.547	0.614	0.616	0.521	0.496	0.364	0.420	0.401	0.259	0.309	0.229	0.243	0.116
Orientation	0.610	0.513	1.000	0.570	0.517	0.578	0.601	0.564	0.491	0.566	0.510	0.300	0.546	0.337	0.301	0.541	0.557	0.530	0.562	0.516	0.331	0.375	0.474	0.344	0.309	0.273	0.194	0.122
Procedure	0.576	0.488	0.570	1.000	0.521	0.567	0.601	0.540	0.405	0.585	0.482	0.405	0.521	0.449	0.407	0.482	0.534	0.494	0.433	0.403	0.320	0.381	0.367	0.323	0.361	0.267	0.355	0.182
Concentration (using)	0.639	0.656	0.517	0.521	1.000	0.602	0.526	0.524	0.495	0.516	0.510	0.382	0.462	0.393	0.483	0.459	0.517	0.513	0.426	0.505	0.364	0.381	0.378	0.383	0.357	0.224	0.276	0.072
Navigation complication	0.670	0.650	0.578	0.567	0.602	1.000	0.618	0.709	0.432	0.610	0.543	0.446	0.504	0.460	0.541	0.633	0.587	0.542	0.456	0.462	0.401	0.373	0.443	0.359	0.332	0.300	0.199	0.133
Control	0.665	0.676	0.601	0.601	0.526	0.618	1.000	0.614	0.569	0.547	0.666	0.445	0.581	0.570	0.499	0.596	0.680	0.575	0.555	0.562	0.323	0.506	0.529	0.351	0.359	0.351	0.182	0.137
Frustration with IA	0.633	0.630	0.564	0.540	0.524	0.709	0.614	1.000	0.466	0.736	0.724	0.526	0.614	0.590	0.559	0.804	0.618	0.562	0.475	0.512	0.414	0.372	0.381	0.470	0.356	0.276	0.302	0.318
Quickly find	0.520	0.564	0.491	0.405	0.495	0.432	0.569	0.466	1.000	0.500	0.461	0.383	0.476	0.437	0.322	0.476	0.547	0.467	0.424	0.406	0.333	0.421	0.476	0.150	0.336	0.115	0.003	0.012
Confusion (layout)	0.610	0.526	0.566	0.583	0.516	0.610	0.547	0.736	0.500	1.000	0.607	0.438	0.625	0.466	0.517	0.537	0.526	0.494	0.477	0.478	0.418	0.376	0.373	0.415	0.403	0.317	0.277	0.177
Understood pages	0.599	0.630	0.510	0.482	0.510	0.643	0.666	0.724	0.461	0.607	1.000	0.557	0.585	0.604	0.536	0.541	0.663	0.563	0.530	0.612	0.468	0.427	0.432	0.429	0.366	0.286	0.246	0.251
Clutter	0.383	0.411	0.300	0.405	0.382	0.446	0.445	0.526	0.287	0.438	0.557	1.000	0.590	0.551	0.420	0.422	0.402	0.370	0.269	0.319	0.319	0.350	0.291	0.305	0.355	0.188	0.412	0.259
Page clarity	0.501	0.497	0.540	0.521	0.462	0.504	0.581	0.614	0.476	0.625	0.585	0.590	1.000	0.638	0.552	0.551	0.543	0.473	0.441	0.462	0.394	0.371	0.348	0.321	0.378	0.171	0.319	0.187
Appearance	0.435	0.495	0.337	0.449	0.393	0.460	0.570	0.590	0.437	0.466	0.604	0.551	0.638	1.000	0.547	0.562	0.636	0.596	0.372	0.462	0.406	0.513	0.397	0.309	0.382	0.094	0.268	0.232
Improvement needed	0.446	0.484	0.381	0.407	0.483	0.541	0.496	0.559	0.322	0.517	0.526	0.420	0.552	0.547	1.000	0.552	0.468	0.426	0.445	0.481	0.394	0.334	0.289	0.346	0.392	0.147	0.285	0.162
User-friendly	0.537	0.547	0.541	0.482	0.459	0.533	0.596	0.804	0.476	0.537	0.541	0.422	0.581	0.582	0.552	1.000	0.638	0.573	0.468	0.472	0.381	0.408	0.466	0.252	0.403	0.146	0.122	0.081
Like using	0.638	0.614	0.557	0.534	0.517	0.587	0.680	0.618	0.547	0.526	0.663	0.462	0.543	0.636	0.488	0.638	1.000	0.807	0.684	0.626	0.521	0.535	0.574	0.382	0.414	0.231	0.087	0.130
Enjoyment	0.626	0.616	0.530	0.494	0.513	0.542	0.575	0.562	0.467	0.494	0.563	0.373	0.473	0.596	0.426	0.572	0.807	1.000	0.645	0.586	0.453	0.523	0.542	0.364	0.410	0.366	0.124	0.141
Use again	0.569	0.521	0.562	0.433	0.426	0.459	0.555	0.472	0.424	0.477	0.530	0.263	0.441	0.372	0.445	0.469	0.684	0.645	1.000	0.839	0.396	0.468	0.544	0.401	0.344	0.388	0.168	0.044
Helpful	0.520	0.496	0.516	0.403	0.503	0.462	0.562	0.512	0.486	0.475	0.612	0.275	0.462	0.462	0.481	0.472	0.626	0.586	0.656	1.000	0.390	0.443	0.414	0.340	0.327	0.226	0.052	0.056
Match expectations	0.417	0.364	0.331	0.320	0.364	0.401	0.323	0.414	0.288	0.418	0.468	0.318	0.294	0.406	0.394	0.381	0.521	0.453	0.396	0.390	1.000	0.398	0.479	0.391	0.335	0.284	0.137	0.076
Replace paper	0.326	0.420	0.375	0.283	0.358	0.375	0.506	0.372	0.421	0.376	0.427	0.350	0.371	0.513	0.334	0.409	0.535	0.523	0.468	0.443	0.398	1.000	0.685	0.347	0.342	0.312	0.164	0.060
Convenient	0.414	0.401	0.474	0.367	0.378	0.443	0.525	0.501	0.476	0.573	0.432	0.291	0.346	0.397	0.280	0.466	0.574	0.542	0.544	0.414	0.479	0.685	1.000	0.314	0.295	0.316	0.057	0.019
Format (printout)	0.286	0.259	0.344	0.325	0.281	0.359	0.253	0.470	0.100	0.415	0.429	0.505	0.321	0.336	0.340	0.252	0.582	0.384	0.401	0.340	0.391	0.347	0.314	1.000	0.397	0.362	0.227	0.286
Reliable	0.313	0.309	0.309	0.361	0.357	0.332	0.359	0.356	0.206	0.403	0.398	0.332	0.378	0.382	0.382	0.403	0.414	0.410	0.344	0.327	0.331	0.342	0.295	0.397	1.000	0.348	0.225	0.044
Trust	0.272	0.229	0.233	0.267	0.224	0.300	0.351	0.276	0.115	0.317	0.260	0.190	0.171	0.618	0.147	0.148	0.233	0.388	0.268	0.233	0.284	0.312	0.316	0.362	0.348	1.000	0.188	0.074
Text Size	0.225	0.243	0.194	0.203	0.276	0.199	0.182	0.302	0.000	0.277	0.248	0.412	0.319	0.266	0.285	0.122	0.097	0.124	0.168	0.052	0.117	0.164	0.057	0.227	0.221	0.188	1.000	0.183
Authentic	0.108	0.116	0.122	0.182	0.072	0.133	0.137	0.318	0.012	0.177	0.251	0.259	0.187	0.220	0.162	0.081	0.130	0.141	0.044	0.106	0.239	0.060	0.086	0.296	0.044	0.074	0.183	1.000

Table G7.6. Mean of both Search Designs - Correlation Matrix

KEY

Correlation (*r*) magnitude

Top-Right - workings	> .8	>.6	>.5	>.4	>.3	>.2	>.1	≤.1
Bottom-Left - residuals			>.5	>.4	>.3	>.2	≤.2	
Proposed Groups	Interaction	Fulfilment	Visual Design	Integrity	NOT Grouped			

Reliability was then computed on each subscale for various interfaces studied, and combinations of those:

Attributes	GROUPS			
	INTERACTION	***FULFILMENT***	***VISUAL DESIGN***	***INTEGRITY***
Usability of eStatements	Flustered Stress Orientation Procedure Concentration (using) Navigation complication In control Frustration with IA Quickly find Understood pages Confusion (layout)	Match expectations Helpful Replace paper Convenience Like using Enjoyment Use again User-friendly	Clutter Page clarity Appearance Improvement needed	Trust Reliable Format (printout)
Cronbach's alpha for mean of all three interfaces	.946	.917	.878	.748
Cronbach's alpha for mean of both searches	.938	.921	.863	.730
Cronbach's alpha for Advanced Search only	.936	.896	.825	.623
Cronbach's alpha for Simple Search only	.940	.937	.830	.704

Table G7.7. Reliability of Sub-Scales

Appendix H - Appendix to the Discussion

H1: Comparison of Questionnaires

Banking Portals (Pilot A & B)	eBanking Transactions (Experiments 1 & 2)	eStatements (Experiment 3)
I felt that the Web site was helpful	I felt that the web site was helpful	I felt that the online Statement Service was helpful
The pages on this Web site were attractive	The pages on this web site were attractive	I liked the appearance of the online Statement Service
When using this Web site I didn't always know what to do next	When using this web site I didn't always know what to do next	When using the online Statement Service I didn't always know what to do next
I could quickly find what I wanted on this Web site	I could quickly find what I wanted on this web site	I could find what I wanted quickly with the online Statement Service
I found the page layout on this Web site confusing	I found the page layout on this web site confusing	The layout of the online Statement Service was confusing
I found the layout of the pages on this Web site very clear	I found the layout of the pages on this web site very clear	I found the layout of the pages on the online Statement Service very clear
I always knew where I was on this Web site	I always knew where I was on this web site	I always knew where I was on the online Statement Service
Moving about this Web site was too complicated	Moving about this web site was too complicated	Moving through the pages of the online Statement Service was too complicated
I felt this Web site was reliable	I felt this web site was reliable	I felt the online Statement Service was reliable
I got flustered when using this Web site	I got flustered when using this web site	I felt flustered when using the online Statement Service
I found the organisation of this Web site very frustrating	I found the organisation of this web site very frustrating	I found the organisation of the online Statement Service very frustrating
The pages on this Web site were easy to understand	The pages on this web site were easy to understand	The pages on the online Statement Service were easy to understand
I did not enjoy using this Web site	I did not enjoy using this web site	I did not enjoy using the online Statement Service
The text on this Web site was too small	The text used by the web site was too small	Some of the text used by the online Statement Service was too small
I found this Web site 'user friendly'	I found this web site user friendly	I found the online Statement Service user-friendly
I liked using this Web site	I liked using this web site	I liked using the online Statement Service
I felt in control when using this Web site	I felt in control when using this web site	I felt in control when using the online Statement Service
Reading the pages on this Web site took a lot of concentration	Reading the pages on this web site took a lot of concentration	Using the online Statement Service took a lot of concentration
I would not use this Web site again	I would not use this web site again	I would be happy to use the online Statement Service again
I feel that this Web site needs a lot of improvement	I feel that this web site needs a lot of improvement	I feel that the online Statement Service needs a lot of improvement
I felt under stress while using this Web site	I felt under stress while using this web site	I felt under stress when using the online Statement Service
The pages on this Web site were very cluttered		The pages on the online Statement Service were very cluttered
The options available did not match my expectations		The options available on the online Statement Service matched my expectations
I found this Web site trustworthy		I don't trust the information from the online Statement Service
This Web site needs more graphics and pictures		
The links on this Web site provided a clear indication of their content		
The links I needed were always visible on the screen		

Table H1.1. Question wording comparisons

Banking Portals (Pilot A & B)	eBanking Transactions (Experiments 1 & 2)	eStatements (Experiment 3)
	It was easy to change existing arrangements using this web site (Expt. 1)	
	The words and phrases on this web site were easy to understand (Expt. 2)	
		The online Statement Service is a poor replacement for paper statements
		Using the online Statement Service was more convenient than paper statements
		The format of the statement printed out from the online Statement Service was suitable for my needs
		The printout from the online Statement Service did not look authentic

Table H1.1 (cont). Question wording comparisons