MEASURING SATISFACTION IN USABILITY TESTS: A COMPARISON OF QUESTIONNAIRE ADMINISTRATION METHODS AND AN INVESTIGATION INTO USERS' RATIONALES FOR SATISFACTION

By

Todd William Zazelenchuk

Submitted to the faculty of the Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Instructional Systems Technology

Indiana University, Bloomington

April, 2002

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee:

Theodore W. Frick, Ph.D.

Elizabeth Boling

Kenneth J. Luterbach, Ph.D.

Jonathan A. Plucker, Ph.D.

April 10, 2002

ACKNOWLEDGEMENTS

Few efforts of any magnitude can ever be accurately attributed to an individual person. This dissertation is no exception. While the pages, words, and punctuation of this work are my own, the ideas and preparation of thought that went into them are the result of my good fortune in having crossed paths with several fine folks along the way. To these people, I am grateful. Their inspiration, encouragement, and assistance were most appreciated.

Dr. Theodore Frick

Elizabeth Boling

Dr. Kenneth Luterbach

Dr. Jonathan Plucker

Dr. Andrew Dillon

Dr. Richard Schwier

Dr. Earl Misanchuk

Peter Ditchburn

Abigail Gonzales

And finally, I am indebted to my parents for all of their love and support through the years. Thanks to them, my obstacles in life have been few and far between.

Maxine and William Zazelenchuk

ABSTRACT

Todd W. Zazelenchuk

MEASURING SATISFACTION IN USABILITY TESTS:
A COMPARISON OF QUESTIONNAIRE ADMINISTRATION METHODS AND AN INVESTIGATION INTO USERS' RATIONALES FOR SATISFACTION

This study investigated the measurement of users' subjective satisfaction in usability testing. Three different methods of administering satisfaction measures were compared to determine whether priming techniques are capable of influencing users' satisfaction ratings. Additionally, users' rationales for their satisfaction ratings were examined to help researchers understand why users sometimes rate systems in a manner that is inconsistent with their observed performances. Finally, the relationships between the primary attributes of usability – efficiency, effectiveness, and satisfaction – and users' rationales for their satisfaction were investigated. Forty-five undergraduate education students from a large mid-western university were randomly assigned to the three test conditions and participated in 90-minute usability sessions using a web-based enterprise portal system.

No significant difference was found between each group's post-test satisfaction ratings, suggesting that priming has no discernible impact on users' actual rating behaviors. Many users did perceive benefits from priming, however. These included an increased awareness of the expectations for the test prior to beginning the usability session, and assistance with remembering their experience throughout the test as a result of their post-task satisfaction ratings. Users expressed thirteen common rationales for

their satisfaction ratings, including *Utility and usefulness are key; Just give me some time; Clear and helpful; It's my fault, not the system's; Time equals satisfaction; Everything in its place; The paradox of information quantity; I can see its potential; Feedback is important; If someone would just tell me; Confusing terminology; Getting it right…finally*; and *I wouldn't do that task anyway.* Significant negative correlations were discovered between users' satisfaction and time spent per task, and between time spent per task and users' effectiveness. A significant positive correlation was discovered between users' satisfaction and users' effectiveness. The findings from this study should help usability researchers better understand and interpret their efforts at measuring users' satisfaction. Additionally, they inform the developers of web-based portals of some of the critical design attributes required for their users' satisfaction.

Abstract approved and accepted by:

Theodore W. Frick, Ph.D.

Elizabeth Boling

Kenneth J. Luterbach, Ph.D.

Jonathan A. Plucker, Ph.D.

April 10, 2002

TABLE OF CONTENTS

List of Tables

List of Figures

CHAPTER I.

INTRODUCTION

The usability of computer software has become an increasingly popular topic for discussion and research over the past decade. The excitement of the Internet and the rapid rise (and fall) of the e-commerce industry have seen the term *usability* gain a strong foothold in our vocabulary. Despite its popularity, however, there is much that remains to be understood about software usability. How usability varies within different contexts, what methodologies are most effective, and how best to measure and understand the various attributes that make up usability, are all questions inviting further research. This dissertation considers the issue of usability testing methodology as it examines our current practice for measuring users' subjective satisfaction with computer systems.

1. Satisfaction as an Attribute of Usability

A critical component of any popular definition of usability is users' subjective satisfaction (Nielsen, 1993; Rubin, 1994; Shackel, 1986; Shneiderman, 1998). Bevan (1995) defines satisfaction as consisting of two parts: *comfort* and *acceptability of use*. While comfort refers to one's overall physiological or emotional response to using the system (feels good vs. tense and uncomfortable), acceptability of use reflects whether users feel in command, whether they find the system to be helpful and easy to learn, and whether they feel the system supports their tasks. Together with measures of effectiveness and efficiency, users' ratings of satisfaction with a given system are considered to reveal much about the quality of that system's design (Billingsley, 1993). Unfortunately, while objective measures of efficiency (e.g. time per task) and

effectiveness (e.g. error rate or quality of solution) are relatively easy to operationalize, record, and interpret, subjective satisfaction is inherently more complex.

For the past two decades, researchers in Human-Computer Interaction (HCI) have investigated what makes users satisfied (Nielsen, 1989; Rivard & Huff, 1988; Rushinek & Rushinek, 1986), what the relationships are between satisfaction and actual usage (Baroudi, Olson, & Ives, 1986), and how to collect user satisfaction data for a given computer information system (Chin, Diehl, & Norman, 1988; Doll & Torkzadeh, 1988; Kirakowski & Corbett, 1993; Lewis, 1991). While previous research adequately addressed a number of issues involved in measuring users' satisfaction at the time, continued research into users' satisfaction with computer systems is now overdue. As computer users, we are living in a far different world from twenty, or even ten years ago. The increasing delivery of software applications via the World Wide Web and the expanding population of users who have grown up surfing the Internet in recent years are a testament to this fact. Innovations in technology continually raise our awareness of what is possible and how our past and current systems have often failed to deliver. Accompanying these technological "advancements", it requires only a little imagination to think that users' expectations and criteria for satisfaction with today's computer applications may be changing as well.

2. Measuring Satisfaction

One of the greatest challenges involved with measuring users' satisfaction has been defining the construct itself. For this reason, many usability evaluations elect to measure and discuss users' *preferences* rather than satisfaction (Andre & Wickens, 1995; Bailey, 1995; Wildemuth, Friedman, & Downs, 1998). Meanwhile, other studies use the

terms *preference* and *satisfaction* interchangeably (Nielsen & Levy, 1994; Tanin et al., 2000). Approaching the task from an Information Systems perspective, Melone (1990) expresses her desire for a relationship between satisfaction and effectiveness to be supported by research, thereby linking users' satisfaction with output-oriented measures and strengthening the business case for well-designed systems. Unfortunately, research into the relationships between satisfaction, effectiveness and efficiency has been lacking and has thus far raised more questions than it has provided answers. While some have suggested that a modest relationship exists (Nielsen & Levy, 1994), more recent studies have found little support for the idea that the three attributes of usability are correlated (Frøkjær, Hertzum, & Hornbaek, 2000; Walker, Fromer, Di Fabbrizio, Mestel, & Hindle, 1998).

A second major challenge in measuring subjective satisfaction lies with the instruments used for collecting the data. Although reliable and validated questionnaires such as the Software Usability Measurement Inventory – SUMI (Kirakowski & Corbett, 1993), the Questionnaire for User Interface Satisfaction – QUIS (Norman & Shneiderman, 1989), the End-User Computer Satisfaction Instrument – EUCS (Doll & Torkzadeh, 1988), and the IBM combination of the After-Scenario Questionnaire – ASQ and the Post-Study System Usability Questionnaire – PSSUQ (Lewis, 1995) have been developed and tested specifically for this task, custom-made questionnaires developed on an ad-hoc basis remain the norm for many usability evaluations (Omoigui, He, Gupta, Grudin, & Sanocki, 1999; Tanin et al., 2000; Walker et al., 1998). While the latter are both convenient and seductive given their customized, on-the-spot nature, the validity

and reliability of such instruments are rarely tested, making the already challenging task of measuring subjective satisfaction even more precarious (Oppenheim, 1992).

Regardless of which satisfaction instrument is used, there is also the question of how it might best be administered in a usability evaluation. While the norm has become to apply some form of post-test questionnaire following the completion of a test, there is speculation that this retrospective approach to measuring users' satisfaction produces inflated ratings (Nielsen & Levy, 1994; Oppenheim, 1992; Teague, De Jesus, & Nunes-Ueno, 2001). Some believe that the inconsistencies often observed between users' satisfaction and their observed performances may be attributed to the unfair demands that traditional post-test questionnaires place on a user's memory (Borrie & Roggenbuck, 1998; Csikszentmihalyi & Csikszentmihalyi, 1988; Nisbett & Wilson, 1977; Root & Draper, 1983). The core of their concern is the time lag between one's experience and the assessment of that experience. Given the limitations of human memory, the greater this separation, the less accurate one's assessment is likely to be.

To overcome the problem of time lag, some researchers have experimented with obtaining more frequent measures of satisfaction. A limited study by Raghavan and Perlman (2000) found that users' post-test (retrospective) subjective ratings of system performance were poor indicators of users' actual performances, while users' post-task (concurrent) ratings were better. A similar study, but this time specifically comparing interrupted versus post-task satisfaction ratings in a usability test, found that ratings gathered during task performances produced lower satisfaction scores than those ratings collected upon task completion (Teague et al., 2001). In both cases, the results of these

studies suggest that users' satisfaction ratings may be subject to manipulation depending on the method used to gather them.

Related research at IBM Corporation considered the challenge of measuring users' satisfaction more frequently and developed a specialized questionnaire for the task. The After-Scenario Questionnaire (ASQ) consists of three brief questions dealing with three important aspects of user satisfaction with a computer system: ease of task completion, time to complete a task, and adequacy of support information (Lewis, 1995). The questionnaire's brevity makes it appropriate for gathering ratings after each scenario or group of related tasks in a test. A psychometric evaluation of the ASQ instrument has also shown it to have "acceptable estimates of reliability, validity, and sensitivity" (Lewis, 1995, p.73). What has yet to be investigated is whether the more frequent rating of satisfaction using the ASQ instrument has any effect on the way users rate their overall satisfaction with a system.

3. Priming Users to Consider Satisfaction

A concept originating from social information processing theory that warrants consideration by usability researchers is that of *priming* (Salancik & Pfeffer, 1978). Priming occurs when information is made more salient or prominent to an individual as a result of some prior exposure or operationalization of that information (Tulving & Schacter, 1990). Studies investigating the effect of priming have found differences in users' attitudes (Salancik & Conway, 1975) and in the way people evaluate advertised products (Yi, 1993). In one study that examined university students' self-reporting of perceived stressors and strains during the school year, priming was implemented to increase students' awareness of the different types of stressors and strains that exist

(Moss & Lawrence, 1997). When these students were later asked to report on the level of stressors and strains they were experiencing at that point in the semester, the students who received the primer reported higher levels of stressors and strains than those students who had received no primer.

Root and Draper (1983) speculate that some form of priming may have a similar impact on users' satisfaction ratings in a usability test. They suggest that providing users with "a prior set toward thinking of the system with a view toward evaluating it" may help users remember the relevant aspects of their experience with the system and consequently, produce more accurate ratings of their satisfaction (p.85). Despite their call for attention to this matter, little research has been performed to investigate the idea, leaving usability practitioners to wonder whether their particular methods of administering satisfaction questionnaires may be influencing the data they collect.

4. Importance of Study

Though many variables are involved in users accepting and using new innovations (Rogers, 1995), it is not hard to imagine that the success of many software applications, e-learning programs, and personal productivity tools depends heavily on users' satisfaction with the product. Indeed, Raskin (1994) notes that several magazine-related usability labs consider satisfaction to represent 50% of the usability of their products. Unfortunately, measuring the satisfaction component of usability has proven to be a challenge. Practitioners have long recognized that users' satisfaction ratings often seem to contradict their observed performances, with many users rating systems highly even when significant problems were encountered during the test (Nielsen, 1993; Teague et al., 2001). While usability testing offers software developers the opportunity to gauge

users' satisfaction with a system prior to its market release, such inconsistencies between users' performances and their stated satisfaction levels can make it difficult for software developers to interpret the results. In some cases, organizations set satisfaction goals in advance of testing and base their future development decisions on the findings (R.Teague, personal communication, April, 2001). If satisfaction scores are capable of being influenced by the method of questionnaire administration, such approaches may lead to less than accurate decision-making.

In order for usability practitioners to be able to plan and conduct tests with confidence, they need to be aware of any impact that priming techniques may have on users' post-test satisfaction ratings. Understanding this relationship, along with the rationales that users provide for their satisfaction ratings, will serve to help us further refine user-centered design methodology and ensure a quality user experience for the products we create.

5. Overview of Study

Based on a combination of the researcher's professional experience and observations reported in the literature, which suggest that users' post test satisfaction ratings are often inconsistent with their observed performances (Nielsen & Levy, 1994; Teague et al., 2001), this study begins by examining the question of whether or not the administration methods used to gather satisfaction data have an impact on users' satisfaction ratings. It proceeds to explore users' rationales for rating a given web-based system the way they do and investigates whether or not those rationales are related to users' actual performances with regards to the time spent completing tasks and the number of tasks completed successfully.

5.1. Research Questions

Specifically, the three research questions for this dissertation are as follows:

1. Does the method (no primer, advance primer, post-task primers) of administering user satisfaction measures in a usability test affect users' post-test satisfaction ratings?
2. What rationales do users provide for their post-test satisfaction ratings?
3. How do users' rationales for their satisfaction ratings relate to their objective performances of efficiency (time spent per task) and effectiveness (success rate)?

5.2. Design of Study

The research questions for this study were addressed using a combination of experimental and qualitative methods. The first question, examining the potential effect of administration methods on users' post-test satisfaction ratings, was addressed by an experimental comparison of three different administrative approaches applied during the usability testing of a web-based portal application at Indiana University. The second and third research questions were addressed using a combination of qualitative methods, including verbal protocol analysis (Ericsson & Simon, 1993), semi-structured interviews (Fontana & Frey, 2000), and the constant comparative method of analysis (Gall, Borg, & Gall, 1996). Forty-five undergraduate students with introductory computer experience in the School of Education participated in the study, completing a series of personalization and information retrieval tasks with the portal (see Appendix A) and rating their satisfaction with the system using a method determined by their test group membership for the study. Complete details of the study design and procedures are presented in Chapter Three of this dissertation.

6. Overview of Dissertation

The remainder of this dissertation consists of four chapters. Chapter Two provides a review of the relevant literature on the topics of Human-Computer Interaction, usability, usability testing, satisfaction, priming, questionnaires, and portals. Chapter Three describes the methodology of the study, including the study's design, a description of the system tested, tasks, subjects, procedures for the study, and the method of data analysis. Chapter Four presents the results found for each of the research questions. Chapter Five concludes this dissertation with a discussion of the results, including an acknowledgement of the limitations of the study as well as suggestions for future research.

CHAPTER II.

LITERATURE REVIEW

1. The Foundations of Usability

1.1. Human-Computer Interaction

Although attention to the interaction between human beings and computers can be traced back to the military in the 1940's, it was the development of time-sharing and networking technologies in the 1960's that encouraged researchers to begin studying human behaviors with computers in earnest (Baecker, Grudin, Buxton, & Greenberg, 1995). Now that computers could do more than simply process jobs in a serial fashion, users had the opportunity to think and make decisions while the computer either waited or completed other tasks. This development had a profound impact on the way computer systems were designed. People began to recognize that a system designed in an iterative fashion with its intended users in mind noticeably improved the users' experience and performance with the system (Bailey, 1993; Gould, Boies, Levy, Richards, & Schoonard, 1987).

The impact of the personal computer in the 1980's saw another fundamental shift occur. Prior to that time, computer software was largely designed by engineers for engineers. This meant that users were already quite skilled in using computers and there was relatively little pressure to make systems easy to use. The introduction of the personal computer, however, saw a dramatic increase in the number of non-specialist users being required to use computers in their daily work (Shackel, 1997). This made it increasingly important for software applications to be easily learned and operated. It was in response to these needs that the field of Human-Computer Interaction (HCI) emerged.

Over the past 25 years, the field of HCI has matured significantly, spawning journals (e.g. *Behaviour and Information Technology*, *International Journal of Man-Machine Studies*, *Human-Computer Interaction*, *International Journal of Human-Computer Interaction*), professional organizations (e.g., Usability Professionals Association, Human Factors and Ergonomics Society), conferences (e.g. Association of Computing Machinery's annual CHI conference), research centers (e.g. HUSAT, Xerox PARC) and graduate programs (e.g. Carnegie Mellon, Georgia Tech, Indiana University) dedicated to the study of understanding the interaction between human beings and computers. With the recent phenomena of the Internet and the World Wide Web, the rapid increase in new computer users around the globe has only fueled the importance of HCI issues. This will almost certainly continue as new technologies such as virtual reality environments, wireless communication systems, and information appliances (Norman, 1998) are developed and adopted.

1.2. User-Centered Design

From the field of Human-Computer Interaction, a philosophy of design emerged that had at its heart the question, "what is the experience like for the user?" (Norman & Draper, 1986). This philosophy, known as *user-centered design* (UCD) has come to be "the prevalent technology design philosophy in response to an increased awareness that interactive computer-based systems often fail to achieve the goals of their designers (especially in relation to user requirements and user satisfaction)" (McKnight, Dillon, & Richardson, 1996, p.630). Collections of industry case studies provide further evidence of UCD's growing importance in their descriptions of the experiences, benefits and

challenges associated with introducing a user-centered approach to design into the traditional software development process (Bias & Mayhew, 1994; Wiklund, 1994).

In practice, UCD involves the design of usable systems with the user at the center of the process (Rubin, 1994). It is typically defined by various procedures and techniques that allow such a focus on users to take place. Some of the major approaches representative of UCD practice include participatory design (Shneiderman, 1998), contextual design (Beyer & Holtzblatt, 1998), task analysis (Hackos & Redish, 1998), usability testing (Dumas & Redish, 1993), paper prototyping (Datz-Kauffold & Henry, 2000), and verbal protocol analysis (Ericsson & Simon, 1993).

In their article on designing for usability, Gould and Lewis (1985) identified three principles that have come to largely define the process of user-centered design.

1. An *early and continuing focus* on users emphasizes the importance of directly observing those who will use the intended system, identifying their characteristics, and determining what their goals, experiences, and work contexts are like.

2. *Empirical measurement* stresses the need for testing versions of the interface design with actual users in order to measure how well users can perform with it and how much they enjoy using it.

3. *Iterative design* calls for a cycle of design, test, redesign, retest and so on until the system meets the intended goals for ease of use.

A few years later, a fourth principle was added. The principle of *integrated design* suggests that, “all aspects of usability (e.g. user interface, help system, training plan, documentation) should evolve in parallel, rather than be defined sequentially, and should be under one management” (Gould, Boies, & Lewis, 1991).

2. Usability Deconstructed

In the 1980’s, the term *user-friendly* became a popular phrase for describing computer technologies that were easy to learn and use. Unfortunately, the phrase was not

easily operationalized, making it difficult to measure and compare competing products and systems. For this reason, the term was "banished" from the index of Norman and Draper's (1986) seminal work on user-centered design.

Today, the term *usability* appears to have successfully replaced *user-friendly* in our vocabulary. In one of the simpler definitions, Preece (1993) describes the relationship between HCI and usability; "The goals of HCI are to develop and improve systems that include computers so that users can carry out their tasks: safely, effectively, efficiently and enjoyably. These aspects are collectively known as usability" (p.14).

A usable product represents attention to detail and a level of concern and respect for its users. It is hardly surprising, therefore, to find that usability has become a major marketing strategy for many companies. Mandel (1997) describes an advertising campaign by Intergraph Corporation claiming that its 3-D CAD program allows users to be more than twice as efficient as its competitors. Other corporations have discovered that an increased focus on usability can play an important role in reducing the costs associated with training, customer support and late design changes (Mayhew & Mantei, 1994). With so many positive attributes, it is easy to see why usability has drawn attention. Yet, the concept of usability remains plagued by many of the same problems as its predecessor, *user-friendly;* it is highly contextual, hard to define with precision, and consequently, difficult to measure.

In their effort to operationalize the term, the International Standards Organization (ISO) has come to define usability as, "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO #9241-11). The ISO definition goes on to define each of

the three primary attributes, although as this review of the literature demonstrates, these measures are often broadly interpreted:

> *Effectiveness* – the accuracy and completeness with which users achieve certain goals. Indicators of effectiveness typically include quality of solution and error rates.
>
> *Efficiency* – the relation between (1) the accuracy and completeness with which users achieve certain goals and (2) the resources expended in achieving them. Indicators of efficiency typically include task completion time and learning time.
>
> *Satisfaction* – the users' comfort with and positive attitudes towards the use of the system. Users' satisfaction is typically measured using attitude-rating scales such as SUMI – Software Usability Measurement Inventory (Kirakowski & Corbett, 1993) or QUIS – the Questionnaire for User Interface Satisfaction (Chin et al., 1988).

Other definitions vary slightly from the ISO definition, including characteristics such as learnability and memorability (Nielsen, 1993), and flexibility (Shackel, 1986), but the primary attributes of effectiveness, efficiency and satisfaction remain the core of most definitions. Shneiderman (1998) offers a twist by shunning the label of *usability* and opting instead for his list of *five measurable human factors central to the evaluation of human factors goals*. Practically speaking, however, the differences between the various definitions are mostly semantic as revealed by Table 2.1.

Table 2.1. Three popular definitions of usability
(van Welie, van der Veer, & Eliëns, 1999)

ISO 9241-11	**Nielsen (1993)**	**Shneiderman (1998)**
Efficiency	Efficiency Learnability	Speed of Performance Time to Learn
Effectiveness	Memorability Errors/Safety	Retention over Time Rate of Errors by Users
Satisfaction	Satisfaction	Subjective Satisfaction

Usability alone does not determine whether users end up actually accepting a new system, however. In addition to the usability and cost issues associated with a new system, Shackel (1991) notes that *utility* is as a major contributing factor to users' acceptability. Will the system do what is needed from a functional standpoint? Does it allow users to do something they couldn't do before? As critical as usability may be to the successful acceptance and use of a new system, he reminds us that it remains but one of several factors that must be continually weighed.

3. Measuring Usability

With the concept of usability came the need for some method of evaluating it. Over the past 20 years, various techniques have been developed and compared (Gray & Salzman, 1998). Of these, two basic categories of evaluation methods are commonly practiced today: usability inspection methods and empirical testing with target users.

3.1. Inspection Methods

Usability inspection methods are those evaluation methods that combine HCI experts with guidelines (heuristics), task scenarios, or some combination of the two to facilitate an expert's inspection of an interface for usability problems. Originally touted as "discount usability methods" for their low-cost and low time requirement (Nielsen, 1994), these methods include heuristic evaluation (Nielsen, 1994), heuristic walkthrough (Sears, 1997), cognitive walkthrough (Wharton, Rieman, Lewis, & Polson, 1994), and pluralistic walkthrough (Bias, 1994). While inspection methods have indeed helped to speed up the evaluation process, research has shown that they too have their limitations. Heuristic evaluations commonly generate more false problems than empirical usability testing and they tend to require multiple, qualified evaluators making it less of a discount

approach than originally intended (Jeffries, Miller, Wharton, & Uyeda, 1991). Similarly, cognitive walkthroughs have been criticized for their tedious approach and the need for evaluators with expertise in cognitive science (Rowley & Rhoades, 1992). Today, usability inspection methods are generally considered most valuable when used very early in the design process to identify major usability problems prior to enlisting the participation of actual users.

3.2. Empirical Methods

Empirical methods represent the second category of usability evaluation and are generally considered the standard against which other methods are compared (Jeffries et al., 1991; Nielsen & Phillips, 1993). Also referred to as *usability testing* (Rubin, 1994) or *user testing* (Gray & Salzman, 1998), this method involves directly observing users as they interact with a system in order to evaluate the system's ease of use. Dumas and Redish (1993) identify five characteristics that define usability testing:

1. The primary goal is to improve the usability of a product
2. The participants represent real users
3. The participants perform real tasks
4. You observe and record what participants do and say
5. You analyze the data, diagnose the real problems and make recommended changes to fix those problems. (p.22)

Usability testing has several benefits. From the user's perspective, it allows users to actually experience a system so they can more objectively rate their satisfaction with it (Root & Draper, 1983). From the perspective of the researcher and designer, the fact that usability testing employs actual users, tends to eliminate the false problems commonly identified by more speculative inspection methods (Jeffries et al., 1991). There are also

benefits for software developers. When user testing is performed early in the design process, it has been found to encourage among developers, a cyclical process of design-test-redesign, a more "consistent presence" of the eventual user, and greater consideration of the effects that design decisions will have on their users (Sugar & Boling, 1995).

Usability testing is not without its shortcomings, however. Two of these are the time and cost involved in incorporating usability into the development process (Rosenbaum, Rohn, & Humburg, 2000). Considering four of Dumas and Redish's (1993) five characteristics, it is easy to see how these factors come into play. The need for participants who represent real users demands an expense of time, effort, and cost for recruitment and scheduling. The need for real tasks requires some form of contextual analysis in order to ensure that the tasks used for the test are truly representative of those that the system must support. The need for the observation and recording of users' behavior and comments during a session requires some type of facility, whether it is a formal usability lab, a portable version, a simple video camera or basic note taking. And finally, as with any research, the data analysis and recommendation phase requires time; the fact that much of the data gathered from a usability test is qualitative only adds to this requirement.

Efforts toward reducing the time and costs associated with usability testing are well documented (Nielsen, 1994; Rosenbaum et al., 2000). In addition to the practice of inspection methods as an attempt to reduce these associated costs, remote usability evaluations have also become increasingly popular. By connecting users and evaluators via network technologies and videoconferencing tools such as Microsoft Netmeeting™, usability evaluators are able to observe and communicate with test participants from a

distance over the Internet, saving travel time and expenses for one or even both of the parties involved (Hartson, Castillo, & Kelso, 1996).

4. Subjective Satisfaction: A Critical Attribute of Usability

If one's only goal in using a given software product is to complete a given task as quickly as possible or to avoid making an error at all costs, then it would seem entirely plausible that the respective measures of efficiency or effectiveness would equate directly to the usability of that product. However, speed and accuracy are not always the principal characteristics for software applications. For several usability labs run by major magazine publishers, satisfaction plays a critical role, with software quality defined by 50% satisfaction, 30% productivity, and 20% intuitiveness (Raskin, 1994).

Consider for a moment the population of World Wide Web users that have had to make a decision between the popular browsers of Netscape Navigator™ and Microsoft's Internet Explorer™. With efficiency determined more by the common limitations of the available bandwidth connection than by the application itself, this attribute of usability no longer plays a discriminating role. And with each browser's main purpose being that of presenting basic hyperlink navigation to users, differences in effectiveness play a lesser role than they normally would for more sophisticated applications. Even cost is removed from the picture as both browsers are available as free downloads from the World Wide Web. Thus, with speed and accuracy representing relative constants in the equation, the deciding factor in choosing a Web browser becomes more subjective; people make their choice based on which application they *like* the most.

Further examples may be found among the multitude of e-commerce companies offering customers similar or even identical products at competitive prices. With so many

qualities being equal, the one variable that can tip the scales in favor of one e-commerce site over another is users' satisfaction with the quality of the online shopping experience (Bouch, Kuchinsky & Bhatti, 2000).

4.1. Defining Subjective Satisfaction

Defining users' subjective satisfaction with a computer system is, in itself, a very subjective process. The popular definitions have tended to equate satisfaction with pleasantness (Nielsen, 1993) or how much the user *likes* the system in question (Shneiderman, 1998). Shackel (1986) proposes the notion of *attitude*, suggesting that a usable system should result in "acceptable levels of human cost in terms of tiredness, discomfort, frustration and personal effort" (p.53). A more recent effort by Bevan (1995) describes satisfaction as the combination of *comfort* and *acceptability of use*:

> Comfort refers to overall physiological or emotional response to use of the system (whether the user feels good, warm, and pleased, or tense and uncomfortable). Acceptability of use may measure overall attitude towards the system, or the user's perception of specific aspects such as whether the user feels that the system supports the way they carry out their tasks, do they feel in command of the system, is the system helpful and easy to learn. (p.124)

On the periphery of usability literature, the concepts of *flow* (Csikszentmihalyi, 1990), and *engagement* (Jacques, Preece, & Carey, 1995) have blurred the lines of what constitutes users' satisfaction with a computer system. While usability methodology has traditionally proceeded with the assumption that an easy-to-use system will also be the most satisfying for the user, research in psychology and instructional design have shown users' satisfaction to be much more complex.

In his seminal work on the psychology of flow, Csikszentmihalyi (1990) identifies a paradox for those in the field of usability. He points out that in many cases, situations in

which users must overcome obstacles result in the most satisfying or optimal experiences for those users. Extending this to the usability of computer systems leads one to conclude that a system, which poses some degree of challenge for a user may, in certain cases, be responsible for higher satisfaction ratings than one that demands little or no effort.

A number of research studies have demonstrated that users often prefer systems that they perform least well on (Andre & Wickens, 1995; Bailey, 1995; Kissel, 1995). Very few studies, if any, have attempted to explore the reasons why. One possible contributing factor is the impact of having a certain degree of challenge embedded in a system. In a series of studies that examined the salient elements of educational multimedia responsible for a user's sense of engagement, the results revealed that the level of challenge associated with tasks and the amount of media control provided to users were highly important (Jacques et al., 1995). Participants in these studies tended to appreciate those media that provided them with the greatest level of control (e.g. text-based media for searching) and they enjoyed the task of "closed searching" over browsing for information because "it offered more of a challenge" (p.54). Such findings serve as a reminder that multiple aspects of a computer system contribute to users' overall satisfaction ratings and that isolating the impact of usability from that of content, tasks, media, and sense of control is difficult at best.

4.2. Satisfaction vs. Preference

When HCI research permits the luxury of direct comparison between two or more interfaces, there is the opportunity to have users work with each interface and subsequently identify the system they most prefer. Intuitively, a user's preference would seem to be the single best indicator of the more usable system – why would people prefer

a system that caused them to make errors or to take extra time to perform given tasks? One might also assume that users' preferences would relate directly to their levels of satisfaction with the system. According to a number of studies, however, neither assumption appears terribly safe.

Research on the relationship between users' preferences and their actual levels of performance on a system has been extensive and, thus far, inconclusive. On the one hand, some research has demonstrated a relationship between the factors. Neilsen and Levy (1994) suggest that we can tell much about users' expected performances based on their expressed opinions of a system. Following their meta-analysis of 57 usability studies that collected both subjective and objective data, the authors concluded that, "there is a strong positive association between users' average task performance and their average subjective satisfaction, and one has a reasonably large chance of success if one chooses between interfaces based solely on users' opinions" (p.75).

On the other hand, users' preferences have been shown to conflict with their actual performances with a system. Even in the Neilsen & Levy (1994) study, despite their claim of a "strong positive association", the authors cautioned that users frequently express preferences that conflict with their optimal performances. They supported this caution with evidence from their meta-analysis indicating that 25% of the users involved stated preferences for systems they performed least well on.

Conflicting results were also the norm in a comparison of traditional versus pull-down menus with novice and experienced users (Carey, Mizzi, & Lindstrom, 1996). In this study, the authors discovered that novice users preferred the interface that they were most efficient and effective with, while experienced users expressed a definite preference

for the system that they performed least well on. Speculation regarding the experienced users' choices included their desire to appear more sophisticated by using pull-down menus and the possibility that pull-down menus support an internal locus of control, which is important to experienced users.

Bailey (1995) describes an experiment in which users expressed their preferences for four different interfaces both before and after use of the systems. In neither case did users select the system that elicited their most efficient performances. Hertzum and Frøkjær (1996) discovered similar results in their study of a text-based retrieval system. In this study, after initial experience with four different methods of text retrieval – browsing, conventional Boolean, Venn-based Boolean, or a combination of all three – users were free to choose their preferred method to complete a final task. The result saw an overwhelming 98% of users select the combination of methods even though browsing had proven to be the fastest method and was responsible for the fewest errors.

Though several studies have used the concepts of satisfaction and preference interchangeably (Nielsen & Levy, 1994; Tanin et al., 2000; Walker et al., 1998), one study refutes the notion that the two concepts are equivalent (Foster et al., 1998). In this study of users' preferences for an automated telephone data entry interface, subjects completed an attitude questionnaire and indicated their system preference after experiencing three different interfaces. Users completed tasks for each of the three interfaces, followed by a 30-item questionnaire after using each interface (subjective rank ordering). Once all three interfaces had been used, subjects were asked to select their most preferred interface for a fourth activity (objective rank ordering). A comparison of the results found that the subjective rank order (questionnaire ratings) predicted the

objective rank order (preference of interface for the fourth task) with 72% accuracy. Interestingly, though a strong relationship was found, users' levels of satisfaction did not always predict their ultimate preferences.

This incomplete overlap between users' preferences and satisfaction ratings suggests that we are dealing with two related, but distinct concepts. A preference requires an opportunity for comparison. It affords users the luxury of considering all the variables involved and making a final decision based on that knowledge. By doing so, users are able to be more objective. Even so, it is clear from the research that users' *objective preferences* frequently contradict their *objective performances* – those measures of effectiveness and efficiency that we frequently use to evaluate the usability of a system.

In contrast to users' preferences, satisfaction ratings are typically gathered when an opportunity for comparison is unavailable. In these cases, objectivity is limited by the fact that users are able to experience one system only, forcing them to speculate how much they may like or dislike the given system. Nielsen (1993) suggests, however, that objectivity is not necessarily forfeited by such an approach.

> …subjective satisfaction may be measured by simply *asking* the users for their subjective opinion. From the perspective of any single user, the replies to such a question are subjective, but when replies from multiple users are averaged together, the result is an objective measure of the system's pleasantness. (p.34)

In other words, provided that users are granted the opportunity to experience using the system firsthand, the larger the sample size, the more objective the findings. Certainly, one may still argue that relying on satisfaction ratings collected in this manner is less desirable than providing users with an opportunity to compare systems, the reality is that opportunities for valid comparisons do not always exist.

5. Measuring Satisfaction

5.1. Are Satisfaction, Effectiveness and Efficiency Correlated?

As previously mentioned, one of the greatest challenges in measuring satisfaction is in defining a construct of satisfaction that is specific to usability. For the past 30 years, the user satisfaction construct has been a major focus of research in the field of information systems, where substantial efforts have been made to link user satisfaction to *output-oriented* measures of effectiveness (Melone, 1990). Unfortunately, there have been few studies that have examined the relationship between these two attributes of usability. In a comprehensive literature review for their meta-analysis of 57 experimental HCI studies, Neilsen and Levy (1994) were unable to find a single study that compared indicators of satisfaction and effectiveness. Such studies have since been conducted, but they remain surprisingly few in number (Frøkjær et al., 2000). Particularly rare are studies that have examined the relationship between all three components of usability. A recent review of 19 studies from CHI'97 to CHI'99 (Frøkjær et al., 2000) produced only eight cases where measures for each of efficiency, effectiveness, and satisfaction were collected. Of these eight studies, only one proceeded to examine the correlations between each of the three attributes (Walker et al., 1998).

Walker et al. (1998) examined users' satisfaction along with their objective performances of efficiency and effectiveness in their study of two different automated speech recognition (ASR) interfaces for an email system. The authors hypothesized that that users would be most satisfied with the interface that provided the greatest efficiency and most accurate speech recognition. What they found, however, was that users preferred the system that they performed least well on. This led the authors to conclude

that, “qualitative behaviors associated with poor ASR performance, the predictability of the system, and the ability of users to acquire a model of system performance, are more important than the commonly assumed performance factors of efficiency and task success” (p.589).

In a second study examining the relationship between the three attributes of usability, the usability of an information retrieval system was tested (Frøkjær et al., 2000). Similar to the Walker et al. (1998) study, this effort revealed that effectiveness, efficiency and satisfaction were weakly correlated at best. Based on this lack of predictive relationship, the authors concluded with the recommendation that evaluators strive to collect all three measures of usability whenever they conduct a study.

Finally, in a study that examined objective vs. subjective perceptions of quality of service on the World Wide Web, Bouch et al. (2000) discovered that contextual factors such as the length of time the person has been interacting with the system and the number of hours spent using the Web were more important predictors of satisfaction than either efficiency or task success. Together, the results of these studies make a rather strong case that efficiency, effectiveness, and satisfaction are unable to serve as predictors of one another.

5.2. Other Challenges in Measuring Satisfaction

Another major challenge in measuring users’ satisfaction has to do with users’ general reluctance to be critical of a system (Oppenheim, 1992). Norman (1988) wrote of this phenomenon as well, chastising users for their unwillingness to blame designers for unusable products. Regarding usability testing specifically, users “tend to be polite and give fairly high ratings unless they are really upset with an interface” (Nielsen & Levy,

1994, p.71). This may be especially true when only a single system is being tested. In a study comparing two interfaces for a collaborative groupware system, users' perceived effort was measured by a repeated questionnaire given after each task (Gutwin & Greenberg, 1998). Results showed no difference in users' perceived effort despite the fact that many more errors were observed for one of the interfaces. The authors speculated that for between-subject studies, uses had difficulty with subjective ratings because they had nothing to compare the system to. In such cases, Oppenheim (1992) notes that attitude scores become even more subjective.

A further consideration in measuring user satisfaction, yet one that often goes unmentioned, has to do with the technique for actually collecting users' satisfaction ratings. A study by Reeves and Nass (1996) investigated the effect of having users complete an electronic post-test evaluation on the same computer that they tested the system with versus completing the evaluation on a different computer. The authors discovered that users tended to give higher (more positive) ratings when they submitted their scores on the test computer, suggesting that a pseudo-social relationship may exist between people and computers that influences people's behaviors. Though print-based questionnaires are probably the most commonly used approach for recording users' satisfaction ratings, online questionnaires are becoming increasingly popular with the advent of the Internet. The work of Reeves and Nass has important implications for usability testing and the gathering of satisfaction data, especially for remote evaluations (Hartson et al., 1996) where satisfaction data are commonly collected online.

6. The Satisfaction Questionnaire

6.1. A Popular Measurement Tool

In the previous sections, we have already seen that users' satisfaction can be difficult to define, measure and interpret. Unlike measures of effectiveness and efficiency, one cannot simply count the number of errors or record the time spent completing a task to know how well the system is meeting users' needs. In some cases, attempts at measuring users' satisfaction are made indirectly, using measures such as the ratio of positive to negative comments during the test, the rate of absenteeism on the job, or health problem reports (Bevan & Macleod, 1994). In other cases, satisfaction has been equated with users' self-reported measures of actual use (Case et al. as cited in Ettema, 1985). In the majority of cases, however, users' satisfaction in usability tests is measured using some form of attitude questionnaire.

User satisfaction questionnaires are typically very short and are applied as part of a debriefing session after a user test (Nielsen, 1993). Longer questionnaires have been developed for use in more detailed studies (Kirakowski & Corbett, 1993; Lin, Choong, & Salvendy, 1997). In some cases, questionnaires have been applied during the course of a test in an attempt to gather satisfaction ratings that are more closely related to actual performance scores (Lewis, 1995; Raghavan & Perlman, 2000; Teague et al., 2001).

6.2. Reliable Instruments for Measuring Satisfaction

Several instruments have been developed over the years specifically for the purpose of measuring users' satisfaction with computer software systems (Chin et al., 1988; Doll & Torkzadeh, 1988; Ives, Olson, & Baroudi, 1983; Kirakowski & Corbett, 1993; Lewis, 1995). While not all of them have proven to be reliable instruments for the

task (Ives et al., 1983), the majority of them are considered to be valid tools for measuring users' satisfaction.

The End-User Computing Satisfaction Instrument – EUCS (Doll & Torkzadeh, 1988) is a 12-item instrument consisting of five subscales: content, accuracy, format, ease of use, and timeliness. The emphasis on quality and presentation of information is evidence of the fact that EUCS was developed in the field of Information Systems. Only two of the 12 items in the instrument have to do with ease of use: 1) is the system user-friendly, and 2) is the system easy to use? While the inclusion of "user-friendly" in the first item makes the instrument outdated by today's standards, a test-retest of EUCS found it to be both consistent and reliable in its measurement of users' satisfaction (Torkzadeh & Doll, 1991).

The Software Usability Measurement Inventory – SUMI is a 50-item questionnaire designed in accordance with psychometric practice (Kirakowski & Corbett, 1993). Of all the questionnaires mentioned in this dissertation, SUMI is the most closely aligned with the ISO 9241-11 definition of usability. It is designed for use with a minimum of ten users and with a working version of the software product. Perhaps the aspect that most sets it apart from other instruments, however, is an accompanying standardization database containing usability profiles of more than 200 different computer applications. This database allows evaluators to apply SUMI to a single application and assess users' overall satisfaction ratings by comparing them to the means for similar applications contained in the database.

The Questionnaire for User Interface Satisfaction – QUIS 5.0 (Chin et al., 1988) is a 27-item questionnaire that includes main categories for overall reaction, screen,

terminology and system information, learning, and system capabilities. While earlier versions of the instrument had many more items, successive versions have been able to reduce the number of items and maintain a high level of reliability. In their study of different commercial software applications known in advance to be liked or disliked by professional and hobbyist users, Chin et al. found that QUIS 5.0 discriminated well between the applications.

The After-Scenario Questionnaire – ASQ (Lewis, 1991) and the Post-Study System Usability Questionnaire – PSSUQ (Lewis, 1995) represent two examples of the contribution to HCI research made by IBM Corporation over the years. Similar to the other satisfaction questionnaires mentioned thus far, the PSSUQ is a relatively short (19 item) post-test instrument designed to measure overall satisfaction with a system (see Appendix B). The ASQ, however, is a brief, 3-item questionnaire that is administered following the completion of each task in a test (see Appendix C). Psychometric evaluations of both the PSSUQ and ASQ have shown acceptable estimates of validity, reliability and sensitivity (Lewis, 1995).

6.3. Challenges and Criticisms of Questionnaires

The challenges and criticisms that accompany the use of questionnaires in social science research have their origin in the differences between quantitative (positivist) and qualitative (postpositivist or interpretive) research.

> Positivist research is grounded in the assumption that features of the social environment constitute an independent reality and are relatively constant across time and settings. Positivist researchers develop knowledge by collecting numerical data on observable behaviors of samples and then subjecting these data to numerical analysis. Postpositivist research is grounded in the assumption that features of the social environment are constructed as interpretations by individuals and that these interpretations

> tend to be transitory and situational. Postpositivist researchers develop knowledge by collecting primarily verbal data through the intensive study of cases and then subjecting these data to analytic induction (Gall et al., 1996, p.28).

In general, positivist research emphasizes the collection and interpretation of objective data, while interpretive research regularly deals with subjective data. Most usability evaluators collect both objective and subjective data depending on the purpose of the evaluation. For example, when increased user productivity is the goal, objective performance measures may be deemed most important. Meanwhile, when there is an expressed need for high user satisfaction with a system, greater emphasis may be placed on subjective measures (Lewis, 1995). While quantitative research may be criticized for its (mis)application of statistical methods or inadequate samplings of the population, qualitative research is most often scrutinized for the validity and reliability of its constructs and instruments. One favorite target of this scrutiny is the attitude questionnaire, the primary instrument for measuring users' satisfaction with a computer system.

The development of valid and reliable questionnaires is not a mystery. Much research has been done to guide us in our design of questionnaires, and psychometric principles are available for helping us test their accuracy (Chin et al., 1988; Lewis, 1995; Oppenheim, 1992; Torkzadeh & Doll, 1991). Unfortunately, as Oppenheim (1992) points out, it is much easier to simply 'throw together' a quick survey than it is to carefully design and test one.

> Too often, surveys are carried out on the basis of insufficient design and planning or on the basis of no design at all. 'Fact-gathering' can be an exciting and tempting activity to which a questionnaire opens a quick and seemingly easy avenue; the weaknesses in design are frequently not recognized until the results have to be interpreted – if then! (p.7)

Nevertheless, the use of ad-hoc questionnaires in usability testing remains common with many researchers electing to measure users' satisfaction in their own way (Omoigui et al., 1999; Tanin et al., 2000; Walker et al., 1998). Unfortunately, the issue of instrument reliability is often overlooked in the ad-hoc development of questionnaires, making it difficult to defend any conclusions made from the data.

Another source of concern with questionnaires has to do with their length. Neilsen (1993) advises that questionnaires be kept relatively short to maximize their return rate. As Oppenheim (1992) acknowledges, however, questionnaires that are too short, may jeopardize the instrument's reliability.

> The assumption underlying these procedures is that there is such a thing as a 'true' attitude, which is also relatively stable, just as in the case of factual questions there are 'true' facts or events. However, since an attitude is more complex than, say, a respondent's method of traveling to work, it is unlikely that a single question will reflect it adequately. Also, the chances are that too much will depend on the actual question form and wording, on context, emphasis and mood of the moment, so that the results will be a compound of the (relatively stable) attitude and of these other (momentary) determinants – hence the poor reliability of the single-attitude question. (p.147)

Ives, Olson and Baroudi (1983) support Oppenheim's concern about one-item instruments, adding that they gather little information from users about what is wrong with a system.

The use of Likert Scales in satisfaction questionnaires represents another source of concern for many researchers. In their favor, Likert Scales are easy to construct, they are more precise than simply asking respondents to agree or disagree with a statement, and they are able to include items that are indirectly related to the attitude in question, thereby enabling "subtler and deeper ramifications of an attitude to be explored" (Oppenheim, 1992, p.200). The indictments of Likert Scales are equally numerous, however. Given that the same total score can be obtained in many different ways, Likert Scales are often criticized for their lack of reproducibility (Oppenheim, 1992). In the case of usability testing, this emphasizes the need to examine individual patterns of scoring for each item, rather than simply the overall average score. Additionally, Oppenheim points out that Likert Scales offer no metric or interval measures and they lack a neutral point, making it difficult to know where scores in the middle range change from mildly positive to mildly negative. Support for this criticism can be found in a meta-analysis of subjective preferences for 127 different systems where the actual mean was found to be 3.55 (+-0.12) on a 1-5 scale (4.82 +-0.19 on a 1-7 scale) rather than the "ostensibly neutral" point of 3 on 1-5 scale and 4 on 1-7 scale (Nielsen & Levy, 1994).

7. Administering User Satisfaction Questionnaires

7.1. Post-Test Ratings

Very little research has been conducted that examines the impact of when questionnaires are applied in a usability test. Traditionally, they are implemented at the end of a test session, after the user has completed all of the scheduled tasks. In this manner, the user is asked to retrospectively rate their overall satisfaction with the system. Satisfaction questionnaires such as EUCS (Doll & Torkzadeh, 1988), SUMI (Kirakowski

& Corbett, 1993), QUIS (Chin et al., 1988) and PUSSQ (Lewis, 1995) are all examples of instruments designed to gather post-test measures of users' satisfaction. The majority of ad-hoc questionnaires developed specifically for individual studies are administered in a post-test manner as well (Omoigui et al., 1999; Tanin et al., 2000; Walker et al., 1998).

7.2. Post-Task Ratings

A few efforts have been made to gather users' satisfaction ratings more frequently during the course of a usability test (Lewis, 1995; Raghavan & Perlman, 2000). Post-task ratings are those gathered from users at the end of each task in a test. At IBM Corporation, an interest in task-specific satisfaction led to the development of the After-Scenario Questionnaire or ASQ (Lewis, 1995). In a study designed to test the reliability of the ASQ with previously developed post-test questionnaires at IBM, Lewis found the 3-item ASQ to be surprisingly reliable ($\alpha = .93$) for such a short questionnaire. It also proved to be highly sensitive to differentiating between scenarios ($F(7, 126) = 8.92$, $p < .0001$) and the interaction between scenarios and systems ($F(14, 126) = 1.75$, $p = .05$). Lewis did not attempt to investigate the correlation between the ASQ's ratings and users' objective measures of performance.

Raghavan and Perlman (2000) reported an experiment that looked at the differences between users' concurrent and retrospective (post-test) perceptions of how accurate a system was. Regrettably, the lack of detail regarding how they operationalized their definitions makes it difficult to know for certain whether their concurrent measures were really within-task measures or post-task measures. In the end, they found that users' retrospective ratings of their perceived accuracy of the system were poorer indicators of

the users' actual performances than were the more frequently gathered concurrent measures.

7.3. Concurrent Ratings

A third type of questionnaire administration may be referred to as concurrent or within-task measures that are recorded as a task is in the midst of being completed. A study by Teague et al. (2001) gathered concurrent ratings on multiple occasions during the course of a test by periodically interrupting users as they worked on tasks. At 30 and 120-second intervals, users were interrupted and asked to verbally answer two questions 1) how easy is this task, and 2) how enjoyable is this task. At the end of each task, the concurrent subjects were also asked to provide post-task ratings by answering the same two questions one more time. Meanwhile, a second group of subjects performed the identical tasks with the identical system, but were only asked to provide post-task ratings for each task. Teague et al. examined the between group differences for users' post-task ratings as well as the within group differences for users' concurrent and post-task ratings and found that the more frequently users were asked to rate the system, the lower their satisfaction scores.

7.4. The Limitations of Retrospective Satisfaction Measures

Problems with asking people to provide retrospective, perceptual measures of their experiences have been well documented. In their review of the literature and based on their own experiments, Nisbett and Wilson (1977) found that subjects who were directly involved in a situation were no better than external observers at reporting the causes of their own behavior. They suggest that people will often base their reports on *a priori causal theories* supplied by their culture, rather than apply an introspective view

based on their direct experience. Ettema (1985) found similar results when examining self-reports of usage of a computerized system for retrieving farm management data. In this case, users' subjective reports correlated poorly with their actual use as measured by objective data gathered automatically by the system.

Nisbett and Wilson (1977) identified five circumstances that may compromise the accuracy of self-reports. In each case, these circumstances hold implications for the self-reporting of satisfaction measures in usability evaluations.

1. *Removal in time*. The longer the period of time between the actual event and the time of reporting, the lower the accuracy of the self report.

2. *Mechanics of judgment.* Factors such as serial order effects, position effects and contrast effects may influence a user's satisfaction rating without the user consciously realizing it.

3. *Context.* When questions are asked about an object, contextual cues are often ignored. However, when a question about context is asked directly, users may report contextual factors as influential when they are not.

4. *Nonevents.* Often people will judge or evaluate an event based on what is missing as much as what is present.

5. *Discrepancy between the magnitudes of cause and effect.* There is a tendency to assume that large causes produce large effects and small causes produce small effects. The result is that small causes may be overlooked entirely as contributors to a user's satisfaction level.

Csikszentmihalyi & Csikszentmihalyi (1988) identify an additional three reasons to question the validity of retrospective measures:

1. Memory is not exact and recall may not be complete.

2. By reconstructing the events and consciousness through recall, subjects may not be able to regain the same situational context where the events actually occurred.

3. Subjects may not be able to separate between actual events as they occurred in a situational context and the personal wishes or social expectancy that may influence its retelling.

To reduce the difficulty with users reporting their prior experiences, the general advice is to keep the delay between recall and report to a minimum (Borrie & Roggenbuck, 1998). This is supported in Root and Draper's (1983) work examining the differences between immediate (hot) and delayed (cold) satisfaction ratings. While the results failed to show a change in users' rank orderings of problems with the interface, more informative details were associated with the immediate condition, leading the researchers to conclude that questionnaire administration should be "arranged as much as possible toward 'hot' conditions, where the user has as fresh experience of using the system as possible" (Root & Draper, 1983, p.87). A similar experiment, which had users rate their perceptions of system accuracy both during and after their use of the system, found that more frequent (inter-task) ratings were slightly better indicators of actual speed and accuracy measures than were retrospective (post-test) ratings (Raghavan & Perlman, 2000).

Despite our understanding about the difficulties with retrospective measures, our current usability methodology continues to encourage evaluators to apply post-test questionnaires to measure users' satisfaction with a system. While a few examples can be found where satisfaction measures have been collected more frequently (Lewis, 1995; Teague et al., 2001), more research is necessary to determine what the associated advantages and disadvantages might be.

7.5. Increasing the Frequency of Collecting Satisfaction Data

Gathering satisfaction data more frequently during the course of a usability test may have potential for increasing the correlation between satisfaction, efficiency and effectiveness (Raghavan & Perlman, 2000; Teague et al., 2001). One possible technique

for accomplishing this goal may be found in past research from the field of psychology. The Experience Sampling Method – ESM (Csikszentmihalyi & Larson, 1987) was originally developed to improve the quality of subjective self-reports produced by subjects participating in social science research. The ESM required subjects to carry with them an electronic pager and questionnaire booklet as they went about their normal routines. At various intervals during the project, the pager would signal subjects to stop what they were doing and fill out a page in the questionnaire booklet. While the ESM makes it possible to gather data frequently, it also introduces the variable of interrupting subjects from their activity so that they may record their subjective perceptions. The impact of such interruptions on the subjects' questionnaire ratings is not well understood.

In a study that examined the internal experiences of people using the World Wide Web, Chen and Nilan (1998) employed an online version of the ESM that randomly presented users with a popup window asking them to respond to certain questionnaire items. Although their approach was able to avoid the problems of subjects' recall and the constant intrusion of a think-aloud protocol (Ericsson & Simon, 1993), the authors noted a couple of negative aspects associated with their method:

> First, some subjects complained that the questionnaire popping up was very annoying and too intrusive. Even though our application had been programmed to minimize the possible intrusion, some subjects felt uncomfortable when they had to stop web surfing to answer the questionnaire. Second, we also received some negative feedback about the length of questionnaire and the time they needed to spend on answering questions (Chen & Nilan, 1998).

In their study involving the usability of various websites, Teague et al. (2001) were careful to consider the issues of interruption and time required for completing within-task (concurrent) satisfaction ratings. To keep the rating activity to a minimum,

the researchers asked only two questions: 1) how easy is this task; and 2) how enjoyable is this task. Subjects were asked these two questions at various intervals throughout the course of each task and again at the end of each task. They simply replied out loud with their responses based on a given Likert Scale. Although the researchers were unable to state with certainty that the interruptions themselves did not influence the overall lower ratings of satisfaction associated with the concurrent scores, their post-test conversations with subjects suggested that interruptions did not contribute to more negative ratings. In fact, discussions between the researchers and subjects following the tests found that many of the subjects actually felt the interruptions helped them accurately evaluate what was happening during the session.

8. Priming

8.1. Priming Defined

The concept of priming has its origins in research on social information processing theory (Salancik & Pfeffer, 1978). In contrast to need-satisfaction and expectancy models (Maslow, 1973), which emphasize individual dispositional explanations for behavior over situational factors, social information processing theory emphasizes the effects of context and the consequences of past choices (Salancik & Pfeffer, 1978). According to this perspective, the salience and relevance of information play an important role in influencing a person's behavior and attitude. The concept of priming was conceived as one method to increase the saliency of information for individuals.

In basic terms, priming is a nonconscious form of human memory based on the idea that an individual's memory is subconsciously triggered by whatever information is

available at the time (Tulving & Schacter, 1990). In practice, priming occurs when exposure to a given stimulus, often in advance of the time that a response is invited, produces an observable effect on the given response. Such effects from priming are abundant in the literature.

In a study of individuals asked to complete a popular leadership questionnaire, Head (1991) examined whether the presence of certain constructs embedded in the questionnaire's items could be responsible for altering subject's responses. By comparing two groups of people in which the first group completed the questionnaire followed by an open-response form and the second group completed the open form followed by the questionnaire, Head observed a priming effect to occur for individuals in the first group. This effect manifested in the higher frequency of certain items appearing in the open-response form for the group that completed the questionnaire first. Head concluded that these items were made salient by the questionnaire and that a priming effect had been realized.

A pair of similar experiments in which priming was observed to have an effect involved the manipulation of subjects' responses to an attitude questionnaire by means of a linguistic device (Salancik & Conway, 1975). In the first study, the endorsement of pro and anti-religious behaviors was manipulated. This resulted in subjects for whom pro-religious behaviors were made more salient, expressing more favorable responses to being religious. Meanwhile, their counterparts for whom anti-religious behaviors were made salient, tended to express less favorable responses to being religious. In the second study, a similar approach was taken, only this time endorsing pro and anti-course behaviors between two groups of university students. In addition to realizing a priming

effect, this study found that subjects' attitudes were also influenced by how relevant the course was to their program of study. When the course being discussed was a part of the student's major area of study, the salient positive or negative behaviors endorsed by the questionnaire tended to be reflected in the student's attitude toward the course. When the course was not part of the student's major program of study, the student's attitude tended to be a function of the grade received.

The effect of priming has been demonstrated in the field of corporate advertising as well. In a study by Yi (1993), university business students read a printed article that primed them to consider specific attributes of a car, including its size, fuel economy, and safety features. The students were then shown different print advertisements of various cars and were asked to express their attitudes toward each one. Depending on the attribute activated by the initial article, students' attitudes toward the car's brand and their intentions to purchase were very different. Attitudes tended to be stronger when the attribute of safety was made salient in the preceding article than when the attribute of fuel economy was emphasized.

Finally, in a study that examined university students' self-reporting of perceived stressors and strains, Moss and Lawrence (1997) found that priming university students to make them more aware of different types of stressors and strains, resulted in those individuals reporting higher levels of stressors and strains than individuals who received no such primer. While Moss and Lawrence's results demonstrated a modest impact only, they posited that repeated exposure to the priming information could potentially elicit stronger results.

8.2. Priming Users for Rating Satisfaction in a Usability Test

Traditionally, users' satisfaction in a usability test is measured in a post-hoc manner following what may be a lengthy test session of one or more hours. Such retrospective ratings are subject to scrutiny, given the limitations of human memory and the separation in time between the actual experience and the rating activity (Csikszentmihalyi & Csikszentmihalyi, 1988; Nisbett & Wilson, 1977). Root and Draper (1983) identify similar challenges within the context of usability testing:

> …without a prior set towards thinking of the system with a view to evaluating it, people may not categorize and remember their experiences from this point of view. Thus when you later ask them to do so, they may be unable to recall much of relevance. (p.85)

The impact of priming as demonstrated by other research (Moss & Lawrence, 1997; Salancik & Conway, 1975; Yi, 1993) suggests that priming might also have a similar effect on users' post-test satisfaction ratings following a usability test. By exposing subjects to the post-test instrument in advance of the test, the information on the questionnaire may be made more salient and serve to assist subjects' recall of relevant details from their experience. If Moss and Lawrence's (1997) speculation is correct that repeated exposure to a primer might also increase the effect on people's subsequent responses, then a primer in the form of the post-task ASQ satisfaction ratings (Lewis, 1991) may serve to provide that repetition.

If priming users with the post-test satisfaction criteria either before or during a usability test is shown to have an effect on their ultimate ratings, usability practitioners need to be aware of this impact. Should a difference exist, the challenge will then be to determine which approach produces the more valid satisfaction ratings. This may indeed

be a challenge as the intuitive correlation one might expect between subjective satisfaction ratings and users' objective performances may not necessarily exist (Andre & Wickens, 1995; Bailey, 1995). Nonetheless, simply knowing whether such an effect exists will help to improve our understanding of usability methodology and perhaps inform the course of future research in this area.

9. Web-Based Portal Applications

9.1. Portals Defined

According to a recent Gartner Group survey, portals are currently the second-most hyped item in the Information Technology world, right after e-business (Phifer, 2000). Despite this popularity, defining what a portal is and what it is not can be a challenge. In its simplest form, a portal provides its users with a single, "one-stop" web page where they can view and use information that is relevant and important to them without feeling overwhelmed or lost on the Web (Looney & Lyman, 2000). With that said, various "flavors" of portals can be described.

Consumer portals provide users with easy, personalized access to vast amounts of information (Looney & Lyman, 2000). Yahoo.com (see Figure 2.1), represents one such example of a consumer portal as it attempts to provide consumers with access to everything they need on the World Wide Web.

Figure 2.1. Consumer portal (http://www.yahoo.com)

Community portals link together groups of people with similar interests or goals (Looney & Lyman, 2000). The finance page for Lycos.com (see Figure 2.2) is just one of many portals designed to serve the financial community, providing investors with Web access to research, discussion boards, and real-time market performance data.

Figure 2.2. Community portal (http://www.finance.lycos.com)

A third category of web-portal is that of *vertical portals*, designed to offer users a specialized business theme or service (Looney & Lyman, 2000). Ebay.com (see Figure 2.3) represents one example of a vertical portal in that it offers its users the specialized business service of connecting together buyers and sellers of new and used products.

Figure 2.3. Vertical portal (http://www.ebay.com)

While consumer, community, and vertical portals represent the majority of web-based portals in operation today, the type that has received the most attention recently by corporations and educational institutions is the *enterprise portal* (Frazee, 2001; Looney & Lyman, 2000; Steinbrenner, 2001). According to Looney and Lyman, "enterprise portals are intended to assist employees to be more efficient and productive by centralizing access to needed data services – for example, competitive information, manufacturing and accounting data, 401K information, and other human relations data" (p.30). An enterprise portal is more than just an access point for centralized data, however. Enterprise portals remove the need for multiple logins to various applications, they allow users to perform individualized or *self-service* processes that could previously be handled only by dedicated staff, and they permit organizations to target their users for individualized services and information based on their identities (Steinbrenner, 2001). The Gartner Group identifies a further five characteristics necessary to define an enterprise portal framework (Phifer, 2000, p.1):

1. Robust search across all structured and unstructured repositories
2. Taxonomy support
3. Content management and aggregation
4. Personalization
5. Application integration and development

While the integration of content, taxonomies, and multiple applications are important benefits of any enterprise portal, it is the personalization aspect that is generally seen as the main drawing card in attracting and retaining users. According to executives at Excite.com, users of portals like MyExcite and MyYahoo (see Figure 2.4)

are five times more likely to return to those sites due to the fact that they can personalize them to meet their individual needs (Looney & Lyman, 2000). As these personalization features continue to improve to include easy, consolidated, and secure access to one's private information, it is not unreasonable to think that portals may become even more attractive to users.

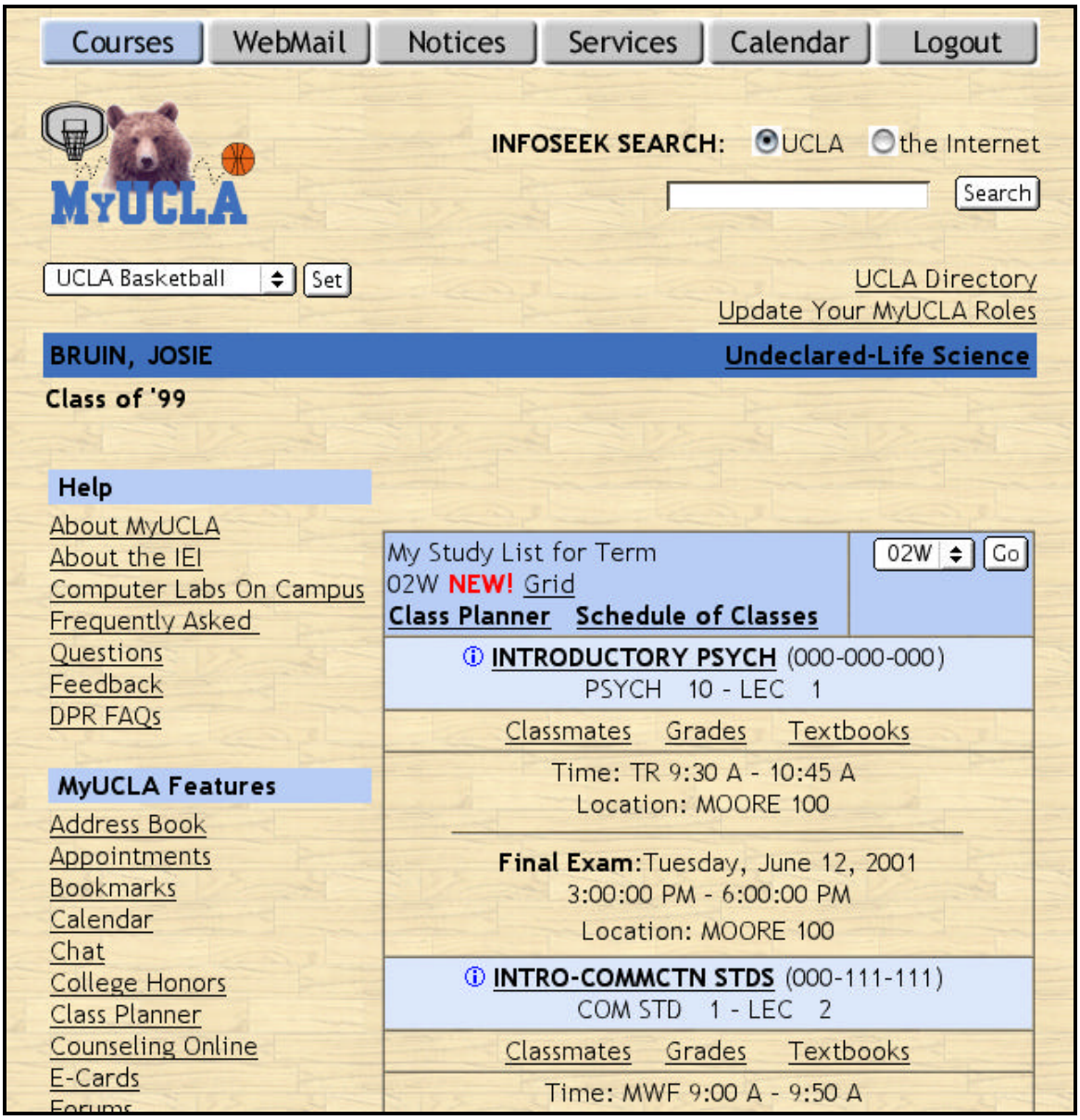
Figure 2.4. Commercial enterprise portal (http://my.yahoo.com)

Despite the recent downturn in the economy and the self-correction of the e-commerce industry, the excitement surrounding enterprise portals remains high. Improvements in employee efficiency, empowerment of people through the individualization of tasks, and the promotion of communities are just some of the benefits seen to accompany the successful implementation of an enterprise portal in industry. It is hardly surprising then to find that enterprise portals have also become one of the hottest topics in the world of higher education.

9.2. Portals in Higher Education

While higher education was arguably the first to embrace the World Wide Web following its creation by Tim Berners-Lee (1998), the effective development and implementation of enterprise portals began in industry, with universities following suit once they recognized their potential impact (Looney & Lyman, 2000). Early pioneers of the portal development effort in higher education included UCLA (see Figure 2.5), the University of Delaware and the University of Buffalo (Frazee, 2001). Other success stories include the University of Minnesota (Kvavik & Handberg, 2000), Villanova University (Connolly, 2000) and Louisiana State University (Ethridge, Hadden, & Smith, 2000). The further development of enterprise portals across the nation's universities will likely only increase as institutions continue to search for ways to attract students to their campus and away from their competitors.

Figure 2.5. University enterprise portal (http://my.ucla.edu)



For institutions of higher education, the campus version of the enterprise portal provides benefits for both the institution and its constituents. From the user's perspective, portals offer two primary advantages (Frazee, 2001):

1. The portal serves as a central gateway into university database systems, resources, and web-supported courses using a single user name and password.

2. The portal lets users customize the interface to meet their needs (p.43).

From the institution's perspective, the power of the portal lies in its ability to span all campus information and provide a consolidated view, resolving the differences between disparate systems (Jacobson, 2000). For those institutions that are able to overcome the hurdles involved in pulling all the pieces together, the potential rewards are significant.

9.3. Indiana University's OneStart Portal

Indiana University's OneStart portal project began in May, 1998 with the publication of the university's Information Technology Strategic Plan. This guiding document identified a number of goals for the institution's eight campuses, including the development of a single "front door" to all administrative applications and an improved level of service offered by the institution's administrative offices and service providers. The characteristics identified as necessary to meet these goals were synonymous with the traits associated with a web-based, enterprise portal application – the ability to sign-on to all applications at once, continuous (24 hour x 7 day) availability and remote access to the institution's applications, role-based presentation of activities, automated workflow capability, and a personalized desktop. (J. Thomas, personal communication, Indiana University, November, 2001). In addition to these requirements, Indiana University desired to make these features available to its entire constituency, including current students, staff, faculty, prospective students, alumni, and service vendors.

Integral to the IT Strategic Plan was the requirement of a user-centered design approach. This included the gathering of input and feedback from numerous steering committees and focus groups early on in the project. It also resulted in an iterative approach to the system's design that included three rounds of usability testing prior to the round associated with this study. Unlike the majority of other campus portals, the

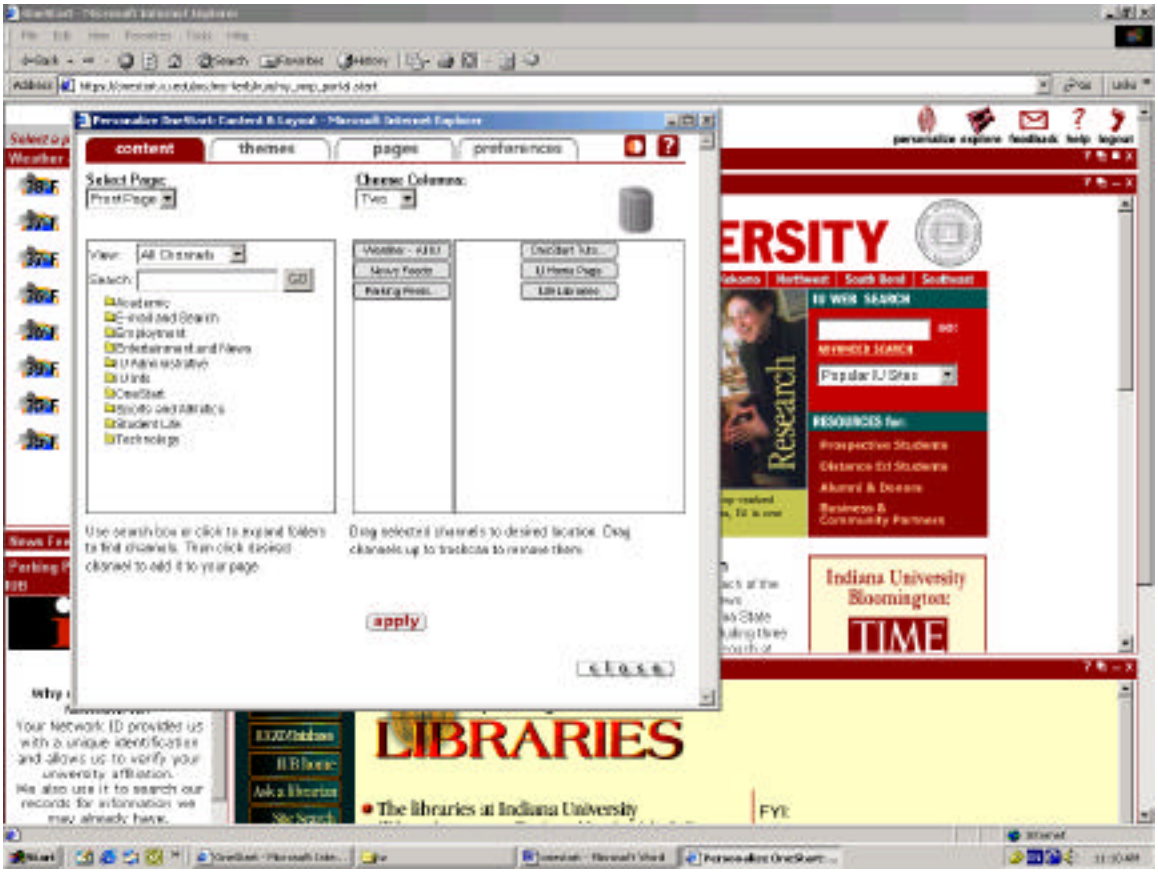
OneStart portal design includes I-frame technology designed to allow users to manipulate multiple web pages on their screen at one time (see Figure 2.6). This design choice continues to be the topic of some debate given that it poses significant usability and accessibility issues, several of which became further evident during the course of this study. As the OneStart project moves forward to its next release, the majority of the usability problems identified during this study are being considered for implementation.

Figure 2.6. Indiana University's OneStart portal (http://onestart.iu.edu)

Certain features were added to the OneStart interface just prior to this study and were tested for their first time. These features included the drag and drop content

management function (see Figure 2.7), the *Tutorial* channel, the *My Bookmarks* channel, *My Custom Channels,* and the *Accessible Option* for selecting and arranging portal content. These new features formed the basis for several of the tasks performed in this study (see Appendix A).

Figure 2.7. The personalization window of the OneStart portal



9.4. Users' Satisfaction with Enterprise Portals

In the mid-1980's, well before the World Wide Web came to people's desktops, Rushinek and Rushinek (1986) asked the question, "what makes users happy?" They surveyed subscribers to a popular computing magazine and identified a number of factors

that correlated highly with users' overall satisfaction rating (see Table 2.2). From their results, we find that certain items (e.g. 4, 6, 8, 13, 15, 16, 17) are less applicable today due to the distributed network advantages offered by the World Wide Web. However, issues surrounding response time and the extent to which systems meet users' expectations would seem to remain highly relevant to users' satisfaction with an enterprise portal.

Table 2.2. Variables that influence user satisfaction with computer systems (Rushinek & Rushinek, 1986)

#	**Description**
1	Satisfaction with response time
2	Efficiency and Effectiveness of database language
3	The extent to which systems are meeting user expectations
4	Promptness of equipment delivery
5	Cost-effectiveness of productivity aids
6	Percentage of systems that are mainframes
7	Compatibility of peripherals
8	Percentage of systems that are microcomputers
9	Compatibility of programs
10	Power-energy efficiency of system
11	System cost
12	Number of users
13	Number of systems
14	Average system life in months
15	Promptness of software delivery
16	Percentage of systems that are minicomputers
17	Expandability of system

The success of a web-based enterprise portal in higher education hinges on many factors. From the perspective of information technology, there is the centralization of vast amounts of organizational data that must occur in order for a portal to successfully merge multiple services into a single application. From the business administration perspective, there is almost always the need to redefine business practices to allow departments and business units to update and maintain their information in the new environment. But what about those the portal is intended to serve? What are the necessary ingredients and characteristics that determine users' satisfaction and ultimate acceptance of an enterprise portal in higher education? While much has been written about meeting the demands of the information technologists and administrators (Connolly, 2000; Ethridge et al., 2000; Frazee, 2001; Jacobson, 2000; Kvavik & Handberg, 2000; Phifer, 2000; Steinbrenner, 2001), a lack of published research exists that identifies users' satisfaction criteria for these new campus applications (Jafari, 2000). Although industry professionals claim that the personalization features of a web-based portal are enough to attract users to return (Looney & Lyman, 2000), additional research remains to be done in order to determine what other factors of an enterprise portal in higher education contribute to users' satisfaction.

CHAPTER III.

METHODOLOGY

1. Design of Study

Answering the research questions for this dissertation required a combination of experimental and qualitative methodologies. The following sections describe the specific approach to answering each of the three research questions.

1.1. Research Questions

The three research questions for this study were:

1. Does the method (no primer, advance primer, post-task primers) of administering user satisfaction measures in a usability test affect users' post-test satisfaction ratings?
2. What rationales do users provide for their post-test satisfaction ratings?
3. How do users' rationales for their satisfaction ratings relate to their objective performances of efficiency (time spent per task) and effectiveness (success rate)?

Answering the first research question required three hypotheses to be defined, one to address each pairing of the three test conditions in the study. These hypotheses, along with the null hypothesis to be rejected in each case are listed below:

H1: Users who experience no priming technique will rate their satisfaction with the system higher than users who experience advance priming.

$H1_0$: There is no difference in the satisfaction ratings of users who experience no priming and those who receive advance priming.

H2: Users who experience no priming technique will rate their satisfaction with the system higher than users who experience post-task (concurrent) priming.

$H2_0$: There is no difference in the satisfaction ratings of users who experience no priming and those who receive post-task priming.

H3: Users who experience advance priming will rate their satisfaction with the system higher than users who experience post-task (concurrent) priming.

$H3_0$: There is no difference in the satisfaction ratings of users who experience advance priming and those who receive post-task priming.

1.2. Experimental Comparison

To address the first research question, an experimental comparison was conducted using a between-subjects one-variable multiple-conditions design (Gall et al., 1996). The dependent variable was the mean post-test satisfaction rating for each test group as measured by the mean PSSUQ ratings. The independent variable was the method of administering the post-test satisfaction questionnaire and consisted of three levels:

1. Control group (Group A)
2. Advance primer (Group B)
3. Concurrent primer (Group C)

The control group (Group A) was used to represent current industry practice for administering satisfaction measures. This group experienced no priming technique during the course of its usability session. After receiving the introductory protocol to the study, subjects in the control group began completing the series of tasks defined for the study. When they had completed the final task, they were then asked to rate their satisfaction using the PSSUQ questionnaire. Table 3.1 illustrates the steps and their sequence for each of the three test groups.

Table 3.1. Experimental comparison groups

Group	Pre-Test	During Test	Post-Test
A	-	-	PSSUQ
B	Advance PSSUQ	-	PSSUQ
C	-	Post-task ASQ	PSSUQ

The advance primer group (Group B protocol) was designed to have subjects become acquainted with the post-test PSSUQ questionnaire prior to the usability test. Subjects in this group began their sessions by using the PSSUQ questionnaire to rate a computer application of their choice that they felt proficient with. Following this activity, subjects proceeded to complete the series of tasks defined for the study. Following their completion of the last task, they completed the PSSUQ questionnaire.

The concurrent primer group (Group C protocol) was designed to have subjects rate their satisfaction with the system more frequently by having them apply the three-item ASQ questionnaire after each task in the session. Following their ASQ rating of the final task in the test, subjects in this group also completed the PSSUQ questionnaire.

1.3. Verbal Protocol

As a part of the introductory protocol for every session, subjects were asked to perform a “think-aloud” or verbal protocol (Ericsson & Simon, 1993) both during their task performances and while completing the post-test PSSUQ questionnaire. Verbal protocol data provides a source of hypotheses about subject’s cognitive processes and predictions about their non-verbal behavior (Bainbridge & Sanderson, 1995; Ericsson & Simon, 1993). While anecdotal notes were taken of the verbal protocol during each

subject's task performances, a complete transcription of this material was not performed due to the total length and number of sessions involved. Full transcriptions were produced of the data produced by each subject during the PSSUQ activity and the post-test interview.

1.4. Semi-Structured Interview

Following subjects' completion of the post-test PSSUQ questionnaire, semi-structured interviews were conducted to determine users' rationales for their post-test satisfaction ratings. Unlike structured interviews that employ predefined questions only, semi-structured interviews are merely guided by pre-written questions. This increased flexibility allows the researcher to pursue promising avenues of importance that emerge from the participants' comments (Fontana & Frey, 2000). Subjects' qualitative responses in the post-test interview also served as a validity check of their questionnaire ratings (Schuman, 1970).

For this study, the semi-structured format allowed the researcher to ask all subjects a standard set of questions (see Table 3.2) and then probe further on an individual basis to explore particular behaviors. A common opportunity for such probing occurred when subjects were observed to struggle with a particular task or aspect of the system and yet continued to rate their satisfaction with that aspect highly on the questionnaire. When this occurred, the inconsistency was pointed out to the subject and an inquiry was made about his/her particular rationale for the rating.

Table 3.2. Post-test interview questions

Group Receiving Questions	Question
A,B,C	What did you like most about the OneStart portal?
A,B,C	What did you like least about the OneStart portal?
A,B,C	Will you use OneStart in the future? Why or why not?
A,B,C	Describe how you decided on your rating for Item #19 (your overall satisfaction with the system)? What factors did you consider? What was most important to you?
B	How did you feel about the activity of filling out the questionnaire beforehand?
C	How did you feel about the activity of rating your satisfaction for each task throughout the session?

For subjects in Groups B and C, the interview included a question about their feelings regarding the particular priming technique applied (see Table 3.2). Although the research on priming suggests that its influence is generally on a subconscious level (Moss & Lawrence, 1997; Tulving & Schacter, 1990), previous usability research involving concurrent satisfaction measures found that subjects actually perceived the more frequent measures to help them accurately evaluate what was happening (Teague et al., 2001). Understanding whether the priming techniques used in this study were consciously perceived and whether the effect was positive or negative was deemed important to fully understanding the impact of implementing primers to help measure users' satisfaction.

All interviews took place at the end of the usability test after the subjects had completed their tasks and had given their ratings on the PSSUQ. The length of the interviews ranged from 10-20 minutes depending on the amount of additional probing

and the depth of response by the subjects. The interviews were recorded onto videotape and later transcribed for data analysis.

2. Materials

2.1. OneStart Portal

The OneStart self-service portal at Indiana University (www.onestart.iu.edu) is a web-based application designed to serve students, staff and faculty by providing a single sign-on, customizable environment in which users are able to select activities of personal relevance and arrange them according to their individual preferences. In late August, 2001, the first version of the OneStart portal was released to the university audience after a three-month period of pilot testing and revisions. The August release represented a system that had already benefited greatly from an iterative design approach and an early and continual focus on users (Gould & Lewis, 1985). However, there remained several known usability issues with the design that were identified as target areas for the study and the August release included some new features (e.g. Tutorial, Bookmarks, My Custom Channels) that had not yet experienced any usability testing.

Due to enterprise-related constraints and dependencies, the August version of the OneStart portal did not have all of its intended features in place for users to experience during this study. The single sign-on feature had not yet been implemented, nor was there a full complement of *channels* (i.e. websites and related activities) for users to select from.

2.2. ASQ and PSSUQ Satisfaction Instruments

The questionnaire materials used in the study consisted of the After-Scenario Questionnaire or ASQ (Lewis, 1991) and the Post-Study System Usability Questionnaire

or PSSUQ (Lewis, 1995), both developed and tested by IBM Corporation. Previous research had demonstrated that both the ASQ (see Appendix C) and PSSUQ (see Appendix B) were highly reliable with coefficient alphas ranging from .90 to .96 for the ASQ and an overall coefficient alpha of .97 for the PSSUQ (Lewis, 1995). Consisting of only three items, the ASQ also offered brevity, making it practical to administer the instrument multiple times throughout the course of a test.

2.3. Tasks

The tasks for each subject included a combination of information retrieval tasks and interface personalization tasks. Information retrieval tasks consisted of locating *channels* of content to be added to the subject's portal page. Personalization tasks required the user to change the look and organization of their interface (e.g. screen color, layout, content, etc.). As part of the pilot sessions conducted prior to the study, the tasks were tested and refined to ensure that variance in efficiency and effectiveness would be realized. A list of the finalized set of tasks performed by subjects in this study is included in Appendix A.

During the data analysis phase, it was determined that certain tasks should be separated into multiple components for the purpose of coding effectiveness or success rates. One such example was Task C, which required subjects to perform a related series of activities as they customized their portal. While the task was a realistic one in terms of requiring the subject to perform a series of sub-tasks along the way, it was possible for a subject to succeed at some of the parts and not others. For this reason, Task C, E, F, and G were divided into two or three parts to allow for more accurate coding of success rates.

3. Subjects

A total of 45 subjects participated in the study. All subjects were either current members or recent graduates of *W200: Using Computers in Education*, an undergraduate course on computing in education at Indiana University.

Two methods were used to invite subjects' participation. In the first method, the researcher visited twelve W200 classrooms in the first week of the fall semester, 2001-'02 and gave a short presentation on the details of the study. In the second method, an email invitation was sent to approximately 300 recent graduates of W200 from the 2000-'01 semester, explaining to them the details of the study and inviting their participation. In both cases, subjects were asked to indicate whether they had any previous experience with the OneStart portal. No subjects indicated having any prior experience with the system.

Subjects were randomly assigned to one of the three test groups using Microsoft Excel. Each person's name was entered into an Excel spreadsheet in the order that people confirmed their willingness to participate. The random numbers (RAND) function was then applied to generate a random number between one and three for each name. Subjects were then assigned accordingly to one of the three test groups described at the beginning of this chapter (Group A =1, Group B = 2, Group C = 3).

In return for their participation in the study, subjects were paid with two gifts, a 150-minute Sprint long-distance telephone card and an IUWare CD-ROM of software (combined value of $20.00). Subjects' participation was purely voluntary and no course credit was provided for their participation.

4. Facilities

All test sessions and interviews were performed at the usability lab operated by University Information Technology Services (UITS) at Indiana University. Located at the Wrubel Computing Center in Bloomington, Indiana, this lab consists of a standard two-room environment with a test room separated from an observation room by a one-way mirrored window. Subjects used a Pentium II computer with a T1 network connection and Microsoft Explorer 5.5 to complete the test sessions.

The researcher facilitated each session by orienting the subject to the system during completion of the first task and then observing the remainder of the test from the observation room. While in the observation room, the researcher was able to view the subject's computer screen on a second monitor and listen to the individual's think-aloud protocol. Verbal communication between the researcher and the subject was possible via an intercom system. All sessions were recorded directly onto VHS tape for later data analysis.

5. Pilot Tests

Prior to the regular test sessions being performed, a total of seven pilot tests were conducted over a period of two months to help refine the tasks and the study methodology. Two pilot tests were conducted early on to evaluate the first draft of the protocol, the clarity and appropriateness of the tasks, the methods for collecting post-task ASQ ratings, and the types of data recorded. From these tests, an estimate was made for the length of the anticipated sessions and the test protocol was refined. These changes included having subjects record their ASQ ratings on a paper form as well as verbalize them out loud, making the tasks independent of each other so that they could be

randomized, and presenting two of the more complex customization scenarios in the form of a *target screen* that users were asked to replicate. The latter decision involved having users view a color print out of a target screen and then use the portal to make the necessary changes so that their personal screen matched the target screen's characteristics. This was intended to avoid cueing subjects through the explicit wording of tasks (e.g., change the theme, delete a channel, etc.).

Five additional pilot tests were conducted in the two weeks prior to the regular sessions. From these sessions, the wording of certain tasks was clarified and a couple of the easier tasks were substituted with others to increase the likelihood of obtaining variance between the performances of different subjects. The sequence of activities in the test protocol were optimized, the recording quality of both the think-aloud and the interview portions of the test was improved, and the semi-structured questions in the post-test interviews were refined to more accurately address the second and third research questions.

6. Procedures

6.1. Scheduling Activity

Following the recruitment activity, subjects' preferred times for participation were reviewed and a tentative schedule for the test sessions was completed. Subjects were then contacted with confirmation of their first or second choice of session. Sessions were scheduled at two-hour intervals over a period of one month including evenings and weekends.

6.2. Test Activity

At the beginning of each session, subjects were asked to read and sign the informed consent statement describing the nature and purpose of the study and the rules guiding their participation. By signing the statement, subjects agreed to having the session videotaped for data analysis purposes. Each subject also completed a one-page demographic form in order to provide information regarding his/her computer and web-related experience (see Appendix D).

Once the informed consent form had been signed, subjects were read an introductory protocol (see Appendix E). This protocol included an explanation for the session, a brief description of the OneStart system, and a comment that at the end of the session, subjects would be asked to rate their satisfaction with the system. The protocols for each group were identical with the exception of an additional short passage for Groups B and C that explained the rules regarding the priming activity for the respective group (see Appendix F for Group B's protocol; see Appendix G for Group C's protocol). As part of the protocol, subjects were asked to think aloud during the session and to try and complete each task on their own without my assistance. Each subject began the first task of logging into the system (Task A) with the researcher present in the test room. After this task had been completed, the researcher left to observe the remainder of the session from the adjacent room. The different protocols for each of the three test groups are described in the following sections.

6.3. Group A

After logging into the system (Task A) and completing the online tutorial (Task B), subjects in Group A were asked to select randomly from a series of 10 task cards and

complete each selected task before moving on to the next one. After the final task in the test was completed, subjects were asked to complete a paper version of the PSSUQ for the OneStart system. Following their completion of the PSSUQ, each subject was asked a series of pre-written interview questions intended to explore users' rationales for their satisfaction ratings. Depending on the subject's observed performance and the degree of consistency among their ratings, the researcher probed subjects for additional comments and elaborations.

6.4. Group B

After completing the introductory protocol requirements, but prior to logging into the system, subjects were presented with a blank PSSUQ and asked to complete it for a computer application of their choice that they felt proficient with (e.g. IU WebMail, MS Word, etc.). Following their completion of the PSSUQ for that system, subjects followed the identical procedures as for Group A, including completing a paper version of the PSSUQ for the OneStart system and participating in the post-test interview.

6.5. Group C

After completing the introductory protocol requirements, subjects in this group followed the identical procedures as Group A with the exception that after each task, they were asked to rate their satisfaction with that task using the ASQ. Subjects performed this rating by referring to the 3-item ASQ appearing at the bottom of each task sheet. In addition to recording their rating for each item on the task sheet, the subject verbally stated their rating, allowing the researcher to record the ratings along the way. After completing the ASQ rating for the final task in the test, each subject completed a paper version of the PSSUQ for the OneStart system and participated in the post-test interview.

During each of the 45 sessions conducted in the study, the researcher observed subjects' actions, recording their performances for efficiency (time spent on each task) and effectiveness (success or failure of completing each task).

7. Data Analysis

The methods of data analysis for this study were determined by the nature of the individual research questions. The first research question was addressed using a combination of reliability analysis and ANOVA statistical methods. The second question required a qualitative data analysis method known as the constant comparative method (Gall et al., 1996) to discover common themes. Answering the third question involved an analysis of the quantitative data for correlations and a review of the data with respect to each individual's objective performances.

7.1. Reliability Analysis and One-Way ANOVA

One of the common problems noted in the design of experimental studies is a failure to confirm the reliability of the measurement techniques used (Gall et al., 1996). Although both the ASQ and PSSUQ instruments had been shown to be reliable in earlier research studies (Lewis, 1991, 1995), a further reliability analysis was performed of both instruments and of the tasks used in the study. Following this analysis, a one-way Analysis of Variance (ANOVA) was conducted to determine whether the method of administering user satisfaction measures had an impact on users' post-test ratings. In this case, the ANOVA statistic compared the means of users' post-test (PSSUQ) ratings between the three groups. All instances of the *Not Applicable* response were coded as missing values within SPSS.

To counter one of the common criticisms of comparing the means between questionnaires comprised of Likert Scale items (Oppenheim, 1992), a factor analysis was conducted on the 19 items of the PSSUQ and a second ANOVA was performed on the primary factors that were identified. This was completed as an alternative to performing the ANOVA for each item in the PSSUQ.

7.2. Constant Comparative Method

The second research question in this study dealt with users' rationales for rating their satisfaction with the system. This question was analyzed using the constant comparative method (Gall et al., 1996). In its simplest form, this method involves reviewing the qualitative data and identifying common patterns or themes. It is generally performed by reading through the data and underlining or highlighting those passages which the researcher feels address the given research questions. Whenever evidence of a particular theme is identified, that section of the data is highlighted and a tally for that theme is recorded on a separate log sheet. Additional themes are highlighted and recorded in a similar manner, adding to the comprehensive list of themes. When evidence of an existing theme is found within the data, the tally for that theme is simply incremented on the comprehensive log sheet.

An important criterion for any research is that of *reliability*, or the extent to which other researchers would arrive at the same result using the same methodology and procedures (Gall et al., 1996). In qualitative research, an important consideration in determining the reliability of one's findings is the amount of *observer agreement* that exists (Frick & Semmel, 1978). While reliability is often mistakenly assumed to be synonymous with observer agreement, Frick and Semmel caution that this is not always

the case; “observers can agree nearly perfectly, yet they can collect very unreliable data if the behaviors of the observed teachers/pupils differ little, or if behaviors are truly unstable from occasion to occasion” (p.159). To calculate observer agreement, one typically compares the observation records of two or more observers with each other or with some criterion when coding the same events or records. The amount of agreement may then be calculated as a simple percentage based on the frequency of agreement between observers (Frick & Semmel, 1978, p.164).

To demonstrate observer agreement of the qualitative analysis for this study, two external raters were employed to conduct independent reviews of the data. The first individual held a Master’s degree in English and had previous experience with qualitative research, although no expertise in the domain of Human-Computer Interaction. This rater was provided with a copy of the themes and a copy of the entire set of subjects’ transcripts. He was then instructed to review the transcripts and identify all instances of the given themes. Whenever the rater identified evidence of a theme, he recorded that theme’s number in the margin of the transcript. Upon completing each subject’s transcript, the rater tallied up the frequency of instances recorded for that individual and summarized the tally at the top of the page

The second individual held a Master’s degree in Instructional Technology, including training and work-related experience in Human-Computer Interaction. Prior to completing the same rating activity as the first external rater, the second rater participated in a short training exercise with the researcher designed to ensure that she could clearly identify the various themes. This exercise involved rating 15 excerpts taken from the 45 transcripts. These excerpts included comments representing each of the sample themes as

well as with a couple of comments that represented no theme. The researcher reviewed the rater's performance and discussed any misperceptions with the rater prior to her beginning the actual rating activity.

Following each external rating activity, the two external reviewers' ratings were compared to the primary researcher's ratings and the percentage of observer agreement was calculated. In cases where both the external rater and primary researcher identified the presence or absence of a given theme for a given subject's transcript, agreement was noted. Whenever an external rater or the primary researcher recorded a theme that the other did not, the disagreement was noted. A kappa coefficient was computed to take into consideration the percentage of chance agreement.

7.3. Correlational Analysis and Descriptive Statistics

The third research question examined whether any relationships existed between users' rationales for their satisfaction ratings and their objective performances of efficiency and effectiveness. In other words, when subjects performed well (few errors, low time per task), were they more or less likely to identify certain themes to explain their satisfaction than when they performed poorly (many errors, high time per task)?

The analysis for this question began with a consideration of the quantitative results for effectiveness, efficiency, and satisfaction. A correlational analysis was conducted to determine whether any relationships existed between subjects' satisfaction scores, the time they spent completing tasks (efficiency), and their success in completing tasks (effectiveness). The value for each subject's satisfaction score was determined by calculating the mean of the 19 items in the post-test PSSUQ. Efficiency and effectiveness scores for each subject were recorded according to common industry practice (Bevan,

1995; Neilsen, 2001) with efficiency represented by the mean time spent completing each task, and effectiveness represented by the number of tasks completed successfully.

To allow for easier interpretation of the positive and negative correlations calculated in response to this question, the original scale of the PSSUQ was reversed within SPSS. Where the original Likert Scale had seven divisions with two anchors (1=strongly agree; 7=strongly disagree), the reversed scale was just the opposite (1=strongly disagree; 7=strongly agree).

Following the correlational analysis, descriptive statistics for users' effectiveness, efficiency, and satisfaction, along with their frequencies of expressing the various rationales were analyzed to determine whether any relationships were present. Additional inferential statistics were not performed on these data to avoid violating the assumption of independence of observations and increasing the probability of a Type I error being committed (Kirk, 1995).

CHAPTER IV.

RESULTS

1. Overview

This chapter presents the results for the three research questions in this study. It begins by presenting the primary descriptive statistics for the study including frequencies, means, and standard deviations. It then proceeds to describe the results of the reliability analyses and the ANOVA performed to answer the first research question of whether different administration methods affect users' satisfaction ratings. The chapter continues with a description of the qualitative rationales for users' satisfaction ratings that emerged in response to the second research question. Finally, the chapter concludes with the results of the correlations and descriptive statistic analysis designed to address the third research question about the relationship between users' rationales and the usability attributes of efficiency, effectiveness and satisfaction.

2. Descriptive Statistics

Requisite for generalizing the results from any experimental study is the random selection and assignment of subjects. While true random selection is rarely attainable and was not attempted in this study, the random assignment of subjects proved to be highly effective based on the evenly distributed characteristics and performances of the subjects (see Table 4.1). It is important to note at this time that the original PSSUQ Likert Scale of 1-7 with anchor points of 1 = highly agree (satisfied) and 7 = highly disagree (dissatisfied) was reversed for all data analyses in order to more easily convey the positive and negative relationships obtained in the correlational analyses. This is to say that the reader should interpret all references to satisfaction data using the reversed

PSSUQ scale where values approaching seven indicate higher satisfaction and values approaching zero indicate lower satisfaction.

Table 4.1. Descriptive statistics

Distribution of Subjects by Gender	**Group**	**Female**	**Male**
Number of Subjects (N)	A	9	6
	B	13	2
	C	11	4
	Total	33	12
Other Demographic Variables	**Group**	**Mean**	**SD**
Age	A	21.47	6.05
	B	20.53	5.50
	C	20.67	5.72
	Total	20.89	5.64
Years at University	A	2.20	1.15
	B	2.00	1.25
	C	2.40	0.91
	Total	2.20	1.10
Web Activities	A	4.40	1.35
	B	4.87	1.41
	C	4.67	1.50
	Total	4.64	1.40
Software Products Proficient in Using	A	6.07	1.53
	B	6.53	2.33
	C	6.67	2.29
	Total	6.42	2.05
Efficiency (Mean time per task in seconds)	A	483.86	118.70
	B	479.03	194.37
	C	439.35	113.73
	Total	467.41	145.00
Effectiveness (Mean percentage of tasks performed correctly)	A	.733	.234
	B	.749	.215
	C	.710	.276
	Total	.731	.238
Satisfaction (Mean PSSUQ rating excluding Item #9) 1 = strongly disagree (dissatisfied) 7 = strongly agree (satisfied)	A	4.989	1.055
	B	5.166	1.196
	C	5.191	1.121
	Total	5.115	1.103

A total of 45 subjects (F = 33, M = 12) participated in this study. All subjects were undergraduate students enrolled in the School of Education at Indiana University with the exception of one subject who was enrolled in the School of Business. Their average age was between 20 and 21 years of age (M = 20.89, SD = 5.64) with the full range extending from 17 to 41 years of age. The majority of subjects indicated having spent slightly over two years at university (M = 2.20, SD = 1.10). Subjects indicated using the World Wide Web to perform an average of more than four different types of activities (M = 4.64, SD = 1.40) including email, online banking, online purchases, and general research among others. They also indicated an average of more than six different types of software applications that they were proficient in using (M = 6.42, SD = 2.05), including such products as Microsoft Word, Microsoft Excel, IU WebMail, Internet Explorer, and Netscape Navigator.

The mean time per task (efficiency) for subjects to complete Tasks C through I in this study was 467 seconds or slightly less than eight minutes (M = 467.41, SD = 145.00). On average, subjects completed the tasks with approximately 73% accuracy (M = .731, SD = .238). Overall, subjects were quite satisfied with the system used in the study with ratings above the midpoint of 3.5 on the reversed PSSUQ scale (M = 5.115, SD = 1.103) with anchor points of 1 = strongly disagree (not satisfied) and 7 = strongly agree (satisfied).

3. The Impact of Administration Methods on Satisfaction Ratings

Answering the question of whether or not the administration method for measuring satisfaction has any impact on users' post-test satisfaction ratings was completed in two steps. The necessary first step involved checking the reliability of the

ASQ and PSSUQ instruments and of the tasks performed by participants to ensure internal consistency (Gall et al., 1996). The second step involved performing a one-way Analysis of Variance (ANOVA) to determine if users' satisfaction ratings varied with their group membership. The results of each step are presented in the following sections.

3.1. Reliability Analysis

The reliability of the ASQ instrument was performed based on data from the 15 subjects in Group C. The results of this analysis showed the ASQ to be highly reliable (Cronbach's $\alpha = .91$), suggesting a high internal consistency among the questionnaire's three items. This was encouraging given that the reliability of measurement scales generally demands a larger number of items (Nunnally, 1970). This strong finding was not entirely unexpected however, as previous research using the ASQ instrument had also shown it to be very reliable with a coefficient alpha in the .90 to .96 range (Lewis, 1995).

The reliability analysis of the PSSUQ instrument was performed using 18 of the 19 items contained in the questionnaire. The ninth item on the questionnaire, which asked subjects whether "the system gave error messages that clearly told them how to fix problems" was removed for the ANOVA calculation due to the large number of subjects who responded that this item was *Not Applicable (N/A)* in their experience. The results of this analysis showed the PSSUQ to be highly reliable ($\alpha = .94$). Again, this was consistent with the results from previous research performed by IBM where the coefficient alpha for the PSSUQ was .97 (Lewis, 1995).

A final reliability analysis was conducted to assess the internal consistency of the tasks that users performed in this study. With Tasks C, E, F and G divided into their component parts, the complete set of Tasks A through I numbered 14 in total. Of this

complete set, Tasks A and B were excluded from the reliability analysis. Task A was removed on the basis that all subjects logged in successfully, resulting in no variance among the sample. Task B was excluded on the basis that, while using the tutorial to learn about the system was deemed a necessary and worthwhile task for subjects to complete, it was difficult to consistently rate it as a pass or fail in terms of users' performances. It was therefore treated more as a training activity designed to allow subjects to learn about the OneStart system than it was as a performance task.

Similar to the findings for the PSSUQ instrument, the analysis of the tasks developed for this study showed them to have acceptable reliability ($\alpha = .82$). Together with the two previous reliability analyses, this finding confirmed the internal consistency of the instruments and tasks used in the study and allowed the researcher to proceed with additional confidence in comparing the satisfaction ratings between the test groups.

3.2. One-Way ANOVA

To determine if subjects rated their satisfaction with the OneStart portal differently depending on their group membership, a one-way ANOVA was performed to compare the three groups. Each subject's mean satisfaction rating was calculated for 18 of the 19 items in the PSSUQ. As per the reliability analysis, the ninth item in the questionnaire was excluded due to the large number of subjects who indicated that the presence of error messages was *Not Applicable (N/A)* to their experience in this study. The means and standard deviations for each group were calculated (please refer to Table 4.1 presented earlier in this chapter) and then compared using the ANOVA statistic. The result of this comparison revealed no significant difference between the three test groups $F(2,44)=.144$, $p = .866$. Therefore, this study was unable to reject the null hypotheses

identified for the first research question and concludes that the priming techniques used in this study do not influence users' satisfaction rating behaviors in a usability test.

To counter the common criticism of Likert Scale questionnaires that the mean score can be obtained through a multitude of different responses (Oppenheim, 1992), an image factor analysis with varimax rotation was performed and a second ANOVA was conducted to compare the three test groups based on the factors found. From the factor analysis, three main factors were identified as being responsible for 68.6% of the common variance (see Table 4.2). A subsequent ANOVA performed to compare the three factors revealed no significant difference between them (see Table 4.2). This finding further supported the results from the earlier ANOVA in which no significant difference was found between the overall mean satisfaction ratings of users from each of the three test groups.

Table 4.2. Three common factors identified from image factor analysis of PSSUQ

Factor	**Description**	**PSSUQ Items**	α	***F***	***p***
1	The experience was good.	1, 2, 3, 4, 5, 6, 7, 19	.9451	0.222	.640
2	The quality of information and feedback was good.	8, 10, 11, 13, 14, 15, 18	.8734	0.016	.900
3	The interface was good.	16, 17	.8042	2.336	.134

Two final ANOVAs were conducted to ensure that any differences between the mean efficiency and effectiveness of the three test groups were not responsible for masking any actual differences in users' satisfaction. These results revealed no differences between the three groups for either their efficiency $F(2,44)=.414$, $p=.664$ or their effectiveness in using the system $F(2,44)=.099$, $p=.906$.

3.3. Users' Reactions to the Priming Techniques Used

As part of the post-test interview, subjects from the two priming groups (B and C) were asked to comment on their feelings about the particular priming technique they experienced. For both groups, these questions were asked at the very end of the interview after all other questions had been addressed. Subjects in Group B were asked how they felt about filling out the PSSUQ questionnaire in advance of the usability test. Subjects in Group C were asked how they felt about the activity of rating their satisfaction with the ASQ instrument for each task, and whether they felt this activity influenced their ratings on the final PSSUQ questionnaire.

The results were mixed for both Groups B and C, but in general, subjects reacted favorably to the priming techniques used. For Group B, a few subjects believed that the advance exposure to the PSSUQ questionnaire may have influenced their assessment of the system. The majority of Group B subjects, however, felt that its main benefit was in making them more aware of the expectations for rating the system before they began using it. Some common reactions to the advance primer implemented in Group B included:

- It's good just to know what things you are going to evaluate the system on to see the questions beforehand and have something to compare it against the whole time.
- It wasn't too bad. It kind of got me an idea about what I was going to look for here (Onestart) to know how I was going to answer the questionnaire. It kept me thinking about whether I liked it or not and how I would rate it.
- I thought that it did give me kind of what I was looking for when it came to the questions and stuff. Then when I was doing this, I was kind of keeping in mind the questions and stuff…does this do that? I didn't really remember the 'organize' one, but other things like that. What kind of problems am I having with this? And I understand that Microsoft Word is just one program and it's a little bit simpler, plus I've used it a lot. So that was really good cause I kept it in mind what I was looking for in programs.

- It's kind of like reading the questions at the end of the book before you read it just to know what you are looking for to pay special attention. I wasn't thinking about rating it, but I was thinking in my head about what I liked and didn't like.

Subjects in Group C who experienced the post-task ASQ primer were slightly more divided on the perceived value of their activity. While a number of subjects believed that rating individual tasks helped them to later recall their experiences, a few disagreed, believing that the post-test PSSUQ questionnaire was different than the simple sum of one's individual task experiences. Some common reactions to the post-task priming activity implemented in Group C included:

- I felt that the rating system was great. If I had rated it afterwards, I think I would have probably been influenced by other things that happened as I was going through the experience. Just having it done right after, you are able to get done with that idea and that task instead of having them all intertwine.
- I think it's a good idea because I would have forgotten how I felt at the time. You asked me since these were more harsh, it kind of makes me see. I don't know, I think it's a good idea just because you know right then what you are feeling.
- No [I don't think it was helpful] because I would give things either a 1-7 and that was how I was feeling right then but at the end it was like the overall…I gave a lot of things 4-5-6 (low satisfaction on original PSSUQ), but at the end it's a 2 (high satisfaction). I think I'm going to use it because I save time and that's really key to me. It's the overall picture that was at 2.
- If I would have just had to do one in the end, it would have been too many things to think about. I think they are good because you get to be more specific about the stuff you liked or didn't like.
- I wasn't looking through the individual tasks. I might have been more inclined to mark things individual ways. It didn't really necessarily impact me because I was thinking overall when I answered the final questionnaire.
- I think the questions are good because they keep you mindful of the things you are looking for. You know, with the amount of ease, and if you are comfortable with it, and time is a big factor cause we are always trying to go faster. Yeah, I think they are good questions. I think they work.
- It helps rate the whole system at the end too. Just because I can go through and think about the ratings that I gave each individual task and it helps when I rate the entire system…well, I rated that really well, but there is that one thing that drove me insane and I didn't rate it that good so we'll kind of combine that and give it a 3 or something like that.

4. Identifying Users' Rationales for Satisfaction

Answering the second research question required an analysis of the transcripts of each subject's PSSUQ activity and interview. This section describes the procedures that were followed and the common themes or rationales that emerged as users described their level of satisfaction with the OneStart portal.

4.1. Common Rationales for Users' Satisfaction

During the initial analysis of the qualitative data, patterns emerged from the material representing users' rationales for their satisfaction or dissatisfaction with their OneStart portal experience. Approximately 40 patterns were initially identified as appearing in two or more of the 45 individual transcripts. This number was then reduced to include only those themes that occurred with a frequency greater than 25% (≥12 out of 45 subjects). Themes that were identified in 12 or more transcripts were tentatively considered to be common themes. Those themes identified in fewer than 12 transcripts were discarded. This conservative rating of 25% was intended to ensure that possible common themes were not eliminated prematurely prior to external validation. The 25% limit resulted in the original number of 40 themes being distilled down to thirteen.

Following a pair of external rating activities, performed to validate the selection of the 13 themes, the results confirmed that certain themes appeared more frequently than others. Despite some low agreed upon frequencies, observer agreement data to be presented later in this chapter warranted retaining all 13 themes for reporting and discussion. Table 4.3 presents the 13 themes according to their observed frequencies with the most frequent themes presented first.

Table 4.3. Common rationales for users' satisfaction ratings

1	**Utility and usefulness are key** - users appreciated the convenience and efficiency of being able to arrange multiple websites of their choice on a single screen.
2	**Just give me some time** – users seemed to be willing to grant systems a grace period. It takes time and practice to get used to any new computer program. Getting things right the first time isn't absolutely necessary.
3	**Clear and helpful** - when a system is self-explanatory and when help screens actually help, users are satisfied. When these things are missing, they are less satisfied.
4	**It's my fault, not the system's** - any problems with learning and using the system are perceived by users as their own lack of skills and abilities rather than the system's flawed design. Users believe that other people will probably find it easy to use, just not them.
5	**Time equals satisfaction** - too much time spent or too many tries to complete a task reduce a users' satisfaction. Conversely, when tasks are able to be completed quickly and easily, the system is more satisfying.
6	**Everything in its place** - users find an organized, consistent interface to be more satisfying than a cluttered one.
7	**The paradox of information quantity** - systems that provide a lot of information are perceived as more valuable, but they are also more intimidating to learn and use.
8	**I can see its potential** - even though a system may not be perfect now, users often imagine the future potential of the system to believe they will be satisfied with it.
9	**Feedback is important** - when a system responds to users' actions and helps guide them through the process, users are more confident that they are doing the right thing.
10	**If someone would just tell me** - users are confident that they could learn and use the system if someone could just 'walk them through it'.
11	**Confusing terminology** - having to learn specific system terminology (e.g. page, channel, portal) can be confusing and tends to reduce users' satisfaction with the system.

12	**Getting it right…finally!** - even when a given task is difficult to complete using the system, succeeding in the end is very satisfying. There is a sense of overcoming the system – of triumph.
13	**I wouldn't do that anyway** - when users consider a task to be unimportant, they are not overly concerned when they cannot complete it. Conversely, when tasks are seen as very important but cannot be completed easily, satisfaction is reduced.

Evidence for each of the thirteen rationales in Table 4.3 was found in subjects' responses to the post-test questionnaire and interview. In an effort to illuminate each rationale for the reader, the following sections present exemplars from the qualitative transcripts.

4.1.1 Utility and Usefulness are Key

The rationale that users most frequently gave to explain their satisfaction was that the system provided them with some additional utility over what they were previously used to. Users cited the characteristics of convenience, efficiency, and the ability to customize their own information as being most valuable and appreciated. Sample exemplary statements from users included the following:

- I would probably use it because I could upload all my websites for my classes and stuff as opposed to putting them into favorites in Internet Explorer. It would save me a lot of time. Like I could go to just one place and check all my class's website and have access to search engines and all the other link things that I use regularly through the internet.
- I liked how you can personalize the channels. I think that's great. I use MyYahoo and it reminded me a lot of that. This is lots better than MyYahoo cause the page is already pulled up rather than having to link to that site. I really, really like that part of it.
- I'd use it. Definitely. Because I can set up the different channels and have them all here. I think it's a pain to jump between the IU Webmail, IU website, Insite and all that. I think it would be great to have it all right there so you can click to this window and click to that window and have it all. I like the weather as well. You can't do that with anything else I've seen, you have to leap to that page and that is a pain.

- So you could log into this from any computer…that would definitely be a plus.
- I don't think it's easier than what we have now, but it definitely allows you to do more things than what we have now. Like what IU uses now, I guess it's not really called anything. In the one I use now I can't go in and personalize my home page, I can't add things on my own. It just allows me to do more, so if I learn how to use it, I think I will.
- It's very convenient for dealing with a lot of different pages at once. Being at school especially, you need to use several pages at once, using Insite and Oncourse, getting different pieces of information from all of them…having them all in one place is really nice so that's why I would use it is convenience.
- It's easy, instead of going to click Internet Explorer and getting one page and then minimizing it and clicking it again and going to another page, all you have to do is go here and there would be the 3-4 pages you use daily.

4.1.2 Just Give Me Some Time

A second rationale for users' satisfaction with the OneStart portal was that it simply requires time and practice for a person to get acquainted with any new system. Ironically, while designers continually strive to design completely intuitive interfaces that allow people to walk up and use new systems, subjects in this study indicated that getting it right the first time was neither expected, nor absolutely necessary, for them to be satisfied. Some common examples of this rationale included:

- It was the first time using it so I wouldn't expect someone to completely do everything right the first time.
- I had difficulties with it, but it's kind of hard to make it easy enough to never have difficulties. I mean you are always going to have to fiddle around with a program when you are first learning how to use it.
- The thing about it was once you got going into it, it got easier… I got more comfortable. A couple of more hours, more days more experience, I think I could become more comfortable with it.
- The fact that it seemed a little complicated. It might have to do with the fact that I've never had to do anything like this. Everything I've ever downloaded it was all there for me. If I had more time…if you could have given me something to read up on this before I came in, I would have felt a lot more comfortable and I'm sure I would have gotten through this a lot quicker.

- Anything with computers, you got to practice it. You gotta get lots of practice in before you decide if you like something so that kind of makes sense anyway. That's the way it's been with anything I've ever tried to do on a computer.
- It was pretty much what I know about computers. It will eventually happen, you just have to go through everything.
- I thought it was easy to use once you got your way around it. and with anything new, I just kind of expected that it wasn't going to be that easy. I was satisfied with it because I completed most of the tasks and even though it did take me awhile, I kind of expected that I wouldn't be able to get them on the first try.
- It's easy to use and then again it's not. But it's something new. Like when I started using the Internet, it wasn't the easiest thing to use…like it was pretty easy, you just go in and type whatever you want to look for and press the button and it brings up all the possible websites, but working with it more I think it would become easier. The first time, things are going to be a little bit hard.

4.1.3 Clear and Helpful

A large number of comments were centered on the general aspect of clear instructions and adequate, effective help for when users needed assistance. When users found these elements of the system to be present, they commented positively about the system and rated their satisfaction highly. Conversely, when these elements were found to be absent or of low quality, users expressed noticeable dissatisfaction with the system. Both types of comments were counted as evidence of this theme. Some exemplars include:

- I struggled to find the information to help me through the process of finishing the task. When I needed the information it wasn't there plain as day so I couldn't do it.
- It was very confusing to use. You can get lost real easily in this system. The lack of information…how to do certain tasks wasn't clear.
- There was the one that I tried to click on personalize and it said please wait…and there was another when, I can't remember. But error messages never tell you how to fix a problem, at least not in my experience. Whenever I get an error message, it tells you blah, blah…that's typical of error messages.
- It's not really self-explanatory. You really have to know something to get into it. I know something about computers where I know if I keep looking and looking the

answer will come. But for someone who has never used a computer this would be very difficult I think, or never used MyYahoo, this would be very difficult.

- I could go to the question box, but that still didn't tell me exactly what I wanted. Sometimes I used the help with other programs and I could type in exactly what was wrong and this didn't have that.
- The tutorial I don't think was very…not that it didn't capture my attention, but I didn't think it was a very easy read. I don't know how to describe it. Did it describe in there how to create channels?
- The help button I needed didn't answer my question. I would have had to email them… There wasn't much information on how exactly to do it, I just had to play around and find it…like creating my own channel and when I wanted to create my own page.
- I think that a lot of it was very self-explanatory. Some of the things I felt I was just able to know what I needed to go to.
- Most of the time the instructions were there to tell me what to do. It just seemed to logically fit together pretty well. I could think of where to look. Even if I didn't read, I could usually figure out where stuff was.

4.1.4 It's My Fault, Not the System's

There was a common tendency for users to be critical and even derisive of their own abilities and performances when describing the difficulties they encountered with the system. Frequently, users believed that others would be able to use the system without problems, but that they themselves lacked either the ability or experience (or both) to use it without making errors. Examples of this common rationale included:

- It kind of took me awhile to figure out what was going on. But that's normal for me though too.
- Some things were more challenging. I found them challenging, most people wouldn't.
- I wasn't too efficient in the beginning because it took me so long to understand. But I think other people would have a much easier understanding knowing the lingo and the basic concept of computers.
- It was pretty clear except that since I was designing it, I made it unclear… I bet the system is perfect and it's just…cause I have problems like this anyway with other programs. For me, it seems like I needed more baby steps to get to the point where it was.

- As I was in there trying to create the pages. especially my Custom Channels, that took me several times to figure it out. Because when I was on the page, I needed that button, but I couldn't figure out how to get it to the page. It was me. I wasn't on the right spot, so that was definitely me.
- It was hard for me. In the beginning, reading about the channels and everything and what everything was that really helped. I think I should have probably read it and paid more attention to it.
- I just think I wasn't the ideal user for it…I'm only in W200 and it's an easy class for me, but as technology goes, I'm more of a technotard.
- There was nothing that told me what I was doing wrong, it was just wrong. I couldn't do it over just because I was blind and didn't know or didn't read the tutorial carefully enough.
- It's pretty easy for everybody else, but for me it was little bit harder. My roommates could have done this in a half hour, but for me, it took way longer.

4.1.5 Time Equals Satisfaction

Users commonly described a relationship between the time spent completing tasks and their level of satisfaction with the system. In general, users tended to be most satisfied with the system when the system responded quickly and they were able to perform the given tasks in short order. On the converse, those users who experienced either slow system performances or who required multiple attempts to complete given tasks, tended to express dissatisfaction with the system. Examples of this rationale included:

- I saw myself trying over and over for things and making the same mistakes. I felt that if you made one mistake you had to start all the way over.
- I did complete them all although it did take me longer than I was happy with I guess.
- Is this connected to Ethernet? It was a little slow, but there is a lot of stuff so I can understand that. I guess I would be concerned if you were using it without Ethernet, how slow it would be.
- If it was something at home, I usually try about 3-4 times and then get frustrated and just say forget it, but this one was good because it seemed like after the second or third try, I got it.

- It just took me so long to figure everything out, it left a bad taste in my mouth, I took up so much time. I was scanning everywhere. I wanted it to be quicker and understand it better.
- Once you get the hang of it, it's very easy to work. It's fast…it was very fast, you pressed apply, you pressed close and it was right there…it was very fast. Once you get it all set up and go through the basic stuff, it will make life faster and easier.
- First of all, the speed with which I was able to learn all of this stuff was exceptional. That's one thing I can't fault it on.
- Probably the complication of doing simple things [was what I liked least]. It just seems a little overwhelming. Like color and printing. Little jobs that should take two seconds and you are going all over the place trying to figure out how to do it.
- I didn't like that it took me long to figure it out. It's understandable I guess. I'm the type of person, I'm very simple. To have all this stuff, it's very nice but if I hadn't gone through this session, I probably wouldn't use it. It takes too long to understand and figure out for me.

4.1.6 Everything in its Place

Another commonly cited rationale was the degree to which users perceived the system to be organized and well laid out. Users expressed greater satisfaction when they found the system to be well organized and consistent in its appearance. Users were less satisfied when they found the system to be inconsistent and unorganized or cluttered. Examples of this rationale included:

- I liked the appearance. I'm big on information being out of the way, but yet still being available. I could just zip up there and go to it rather than a huge toolbar. I liked how you could change the colors and stuff.
- I don't like small things on the side. It's harder to read and you have to scroll.
- I can't explain why, but just kind of the layout and everything kind of bugged me. I don't know why. Like I said before, when you create MyYahoo, it's real simple to maneuver and the layout is real easy. This [Onestart] with the windows popping up. You saw me at the beginning, if you see a window you instinctively click on it to make it go away like ads. Maybe trying to do too much in some ways. The nice thing about Yahoo, the list was very clear about the objects you wanted on the page rather than having to move them. I think there was a separate screen…just much simpler.

- Also, the look of the presented page. Right now, it looks great. But when you have multiple ones open and you are not able to change the column diameter, it looks awkward…I think that is the right word.
- I personally don't like to look at eight things at once. I like to look at one and do whatever I need on that one. It's too much. I like to look at one, check my email or whatever I'm doing. And when I do have more than one open, I minimize them all and open them back up when I need to.
- It was really cluttered.
- I definitely do like the way the layout allows you to have different columns. Different channels that allow you to maximize and minimize.
- For my own personal eye, when things are separated and different, it's a lot easier for people to follow along. You also learn that people's eyes scan from the top left all the way down. I like the fact that this is smaller over here and it has more things on it and this is more of a bigger screen over there.
- It was just very clear. To have all these sections, everything is easy to organize. It's very clear what each section is, differentiating between this one and this one and this one…and finding those different ones. Very clean organization so it's easy to remember and keep things straight.

4.1.7 The Paradox of Information Quantity

Another interesting rationale for users' satisfaction had to do with the amount of information provided by the system. Based on users' comments, a challenging paradox exists for web-based portal designers. Users perceive systems to be more valuable when they contain large amounts of information. However, the more information a system contains, the more intimidating and difficult they find it to learn and operate. Users' satisfaction ratings were frequently rationalized in both of these directions throughout this study. Some examples of users' comments included:

- It was kind of too busy. When I first sat down, it was like wow, there was a lot of information to look at… I thought it was too busy. Too much information at once…[The thing I liked most about it was] just all the information it has. Once I actually understand how to use it, it would be a very helpful tool to have right at my fingertips…very resourceful.
- It's definitely appealing and there's so much information there, but again it's just new to me.

- I just think that it's really easy to use and it offers a lot. I've never even used bookmarks on regular Internet. This gives you a lot of options and things like that.
- It has anything that yahoo or any of those would have, but it's through IU so it's nice. Yahoo puts a lot of stuff on there that I don't need. This gives me an option to put what I want on here…It's got a lot of stuff, but it's not stuff that they are throwing at you, it's kind of optional screen type stuff. I think of them as folders, but they are not necessarily full and open and right in front of you. Whereas in Yahoo, everything is spread out right in front of you. This is more of an option for you if you want it.
- I really think it has too much information for me to play with.
- There's too much information coming out at once. The instructions…as I was whining about I'm reading this, but I need a visual. The ones in the tutorial at the start were too small to grasp.
- There is so much on here. I really don't know where to begin to look to find what I'm looking for… Sometimes, it's too busy, too much. But you can kind of control that too by how many columns and channels you set up. I'm just not used to that much.
- Everything was out there, it was clear and there was a lot of information on it. I like that.
- How much it can let me do and I feel that this system has a lot of stuff that I didn't even see, but stuff that I would probably use.
- While it has everything I want, it kind of makes it very 'out there' at first. Like before I became familiar with the program at all, it was a lot of information thrown at me and was kind of confusing and almost a little intimidating. Once I became comfortable, I obviously liked that, the fact that it has the options for anything. So by the same token, it would be a like and a dislike.

4.1.8 I Can See its Potential

Related to the previously mentioned rationale of utility and usefulness was the common sentiment that users believed the system to have great potential. In fact, the similarity between the themes of utility and potential later proved to be a challenge for distinguishing between the two at times, a point that is reinforced by the observer agreement results presented later in this chapter. In general, users expressing this theme tended to comment in more general terms about the system's future potential, whereas the utility theme tended to identify concrete examples (e.g. personalization, accessibility,

efficiency, etc.) that users perceived to be advantageous. Examples of this rationale included:

- It looks like once you get to know how to use it, it would be a great system to use. It would be easy.
- But I do think it is a good thing and will be a good thing. I can see where it will benefit me and others.
- I think it has a lot of potential. I'm not completely satisfied after this little session, but I'm interested in learning more about what it can do for us.
- I mean obviously it's supposed to help people use their time more efficiently, would that be a good word? See not now [for me], but in the future I think it could be so I will use it in the future. I'm not sure I'll be a regular, but I will give it a try.
- The setup on the main screen was also very nice. It would be nicer if you were able to vary the column dimensions, but I'm sure that will come out sometime in the future.
- I had problems but that's the first time I worked on it and it's still getting worked on. You just look at this and you know that it's going to be great for the future. And as soon as you get all the kinks out and we figure out how to use it and manage it properly it will be great. It's just going to take everyone some time which is why I rated it so high…it's a great idea and it's going to be so fun.
- I think the program looks interesting enough that I would take the time to learn how to use it. I think that's important. I could see some people not really wanting to bother with something new that is going to take them forever to learn, then they might not be as attracted to it. *[includes evidence of Theme 5 as well]*

4.1.9 Feedback is Important

In addition to the need for clear instructions and effective help, several users expressed a need for responsive and informative feedback to help inform them how to proceed in certain situations. When such feedback was perceived as present, users' satisfaction ratings tended to be higher. When users perceived the system to be lacking in feedback, their satisfaction was reduced. Examples of this rationale include:

- It would be nice if something popped up when I closed it (the portal window) where it said, "Are you sure you want to do this?" There were a couple of times where I just lost stuff because it didn't do that. If it did, that would definitely help.

- I strongly agree [that error messages were helpful] cause when I actually read them, I figured out what they are saying. There is one that kept coming up when I would press close…it would be, "Do you really want to lose [your changes]?"…I pressed the X [button] thinking that I would minimize it and the error message came up and said, "Do you want to remove it?" so I pressed no and knew to use minimize.
- I was kind of confused on some of those. But they did pop up and they did kind of help.
- Yeah. When I was trying to do too many things, it told me to stop and wait till it loaded. That was clear.
- As well, when I add something or delete something, it's nice to get a message back saying that's done, you did it correctly. That would be real good… I grew up with Star Trek and you have this assumption that a computer will respond to tell you, "Don't go this way, go that way." It just doesn't give me any feedback as to whether I did something right.
- When I saw apply, I figured it will be there and then it wasn't. So I figured I had done something wrong or maybe it didn't listen. To get a bad rating, it would have had to be really slow and not responded to me at all.

4.1.10 If Someone Would Just Tell Me

In several cases, users expressed a certain confidence that they would be able to effectively learn and use the system if they could just have someone "walk them through it". To use it on their own, however, was perceived to be beyond their ability. Examples of this rationale included:

- I would feel more comfortable if I had someone who knew what was going on walk me through it. Once I'm walked through something, I'm fine, but starting, this would be very scary. If I didn't have a paper telling me what to do or giving me some instructions, I wouldn't know even how to start.
- Like I said it would be so much easier if someone would come and tell me how to do it. I'm the type that once I have someone tell me how, I'm fine. I can work through it again and usually not have a problem. But if I started off and logged in without someone telling me okay try to do this, I wouldn't have known what to do. I would have no idea.
- I could once I have someone show me you can do this, do that…I'm a visual learner…if you do this, do that, I could get it…not on my own, but if I had someone, I'd learn faster.

- I like to be taught personally with people to show me things…I guess I'm lazy like that I don't know so doing it on my own is more frustrating you know so I probably wouldn't'…I'd rather have someone explain it to me and I have friends who are big into computers so they just sit there and they tell me verbally and that's how I usually go.

4.1.11 Confusing Terminology

The use of confusing and unfamiliar terminology was a common source of frustration for users in this study. Not all users found this to be a problem, nor were all identified terms equally problematic, but users' confrontations with terms such as *page*, *channel*, and *portal* frequently resulted in users' negative comments about the system. Some examples of this rationale included:

- They were pretty clear, but I couldn't understand the lingo. The information that was there was clear, it just wasn't what I was looking for.
- The first thing that comes to mind is the terms used…page, channel. I see the need to be inventive, but it doesn't help me figure out where they are going in terms of the terminology they use. They really don't match up with page and channel for me.
- The tutorial. I thought the terminology in it wasn't that great. I just thought it was a little confusing with the channels and pages and the differences between the two of them. It got really wordy at times. Maybe if there were more diagrams in the tutorial that would help.
- In the My Custom Channels, I didn't know what the 'detach with toolbar' meant.
- Changing the colors I eventually found under themes…not sure that's the best word.
- At first the tutorial was like…man, this is confusing. It had small pictures and I thought, I don't really know what they are doing here, what they are talking about. I don't know a lot of computer words.
- I didn't like the terminology. The whole starting out using 'portal' and all that other stuff. There are other words out there already that would be much better if you used stuff that people were already aware of. I felt you were trying to reinvent the wheel. I thought channel was good. Portal was kind of confusing. The channels were really nice. That term didn't bother me. It was in combination with portal and all that other stuff. I would say you have your page, which you are able to put links or connections or some other …something else like that. But I've

never…I've heard portal once or twice in technology, but it's just not something that I've grasped onto yet.

- Channels I understand now. No hassles. I still don't know what portal icons are for sure. I'd put all my pages and stuff, channels. I'd be more likely to call them sites.

4.1.12 Getting it Right…Finally!

Even when users were observed to have experienced significant difficulty in completing given tasks, they commonly expressed great satisfaction when they finally completed them successfully. Upon completion, users noted a sense of triumph in overcoming the system and often echoed the second rationale of blaming oneself by suggesting that they should have been able to figure it out more quickly. Some examples of this rationale include:

- I was happy when I did it…like in chemistry in high school and you figure out the elements, it takes forever, and you do fifty trillion tests, but when you finally figure them out, you're happy. So even though it was long and frustrating it is rewarding when you do get it done.
- The fact that I was able to figure it out in the end, I managed to figure out all that I had to do… it was like 'yeah I finally figured it out' because I messed it up. Just the fact that I overcame the difficulty of screwing it up and then getting it right.
- Creating the page when I finally did it that was really exciting. Like the IUB page when I got, I don't know what I put on there. When I finally figured out how to put the new channel on there, that was fun.
- The ones that are easy, you go through it and no big deal, but when you are sitting there frustrated real bad about why can't I do this or what's going on here, and then you finally do it, it's like oh yeah! The Mystuff one [was most enjoyable] cause I struggled at first and then got it.
- I liked succeeding on my first time. It's nice. It's like I can do this. The second time, it's like, well, I just need practice. It's just kind of frustrating. There is still a lot of satisfaction there because you're like 'It's like I beat it, I gotcha'. But the first time you get it right, it's like "I'm so smart'.
- The Bookmarks one [task] was satisfying because I was able to figure it out finally. Then once the whole refresher button was figured out, it was on my screen and I was able to go to a site instantly without having to type in the address…so I guess it was satisfying that once all the jumping through barrels was done, I had everything I wanted on my page.

- In the end after I finished them, I was probably more satisfied with the Bookmarks one [task] because I finished it and got it done and figured it out by myself.
- [I enjoyed] The first one [the more difficult task] because I'm a perfectionist…sort of an 'Aha! I showed you!' I was getting a little frustrated with the first one because I couldn't find it and I wanted to figure it out.

4.1.13 I Wouldn't Do That Anyway

The final rationale expressed by users in this study suggested that satisfaction ratings may be influenced by users' perception of the importance of the task. When users consider a task to be unimportant, they tend to be less concerned when they are unable to complete it. Conversely, when a task is regarded as highly important or valuable, being unable to easily complete it reduces one's satisfaction with the system. Some examples of this rationale include:

- I had a little trouble with it at first, but when I designed where I wanted things and what I wanted on there, it made me feel like I put this together you know, it looks this way, not because I made the program, but because I arranged it this way and the channels that are on this page are because I picked them. It's the one that I think was the most beneficial anyway.
- My Customs Channels [was the thing I liked least] because I couldn't figure it out…and that really annoys me. I don't like it when I can't figure out how to use things. It seems like it would probably be something that I would want to be able to use…adding something that I couldn't search for. I would probably want to have the movie site in here somewhere. I use that one quite frequently because they don't print them in the IDS all the time. I'd probably have a page of random, like, horoscopes and all that stuff I'd put up and I don't know if they would show up in a search…so I'd need to be able to use My Custom Channels which I don't know how to do.
- I was thinking how at times, it was very easy for me to use and at times I got frustrated. The easy outweighed the frustration. There was just the one time with the colors, but to me that really wasn't important. To me, colors doesn't really matter…as far as colors I wasn't going to get stressed out about it and search high and low for it, cause it's color and it doesn't really make a difference, whereas my channels, I kept working on that even though I was frustrated, cause in the end, that would help me to have a specific channel right there.
- There are a couple of things that could have been more pronounced, or more in the foreground of the page so it would be easier so you didn't have to search so

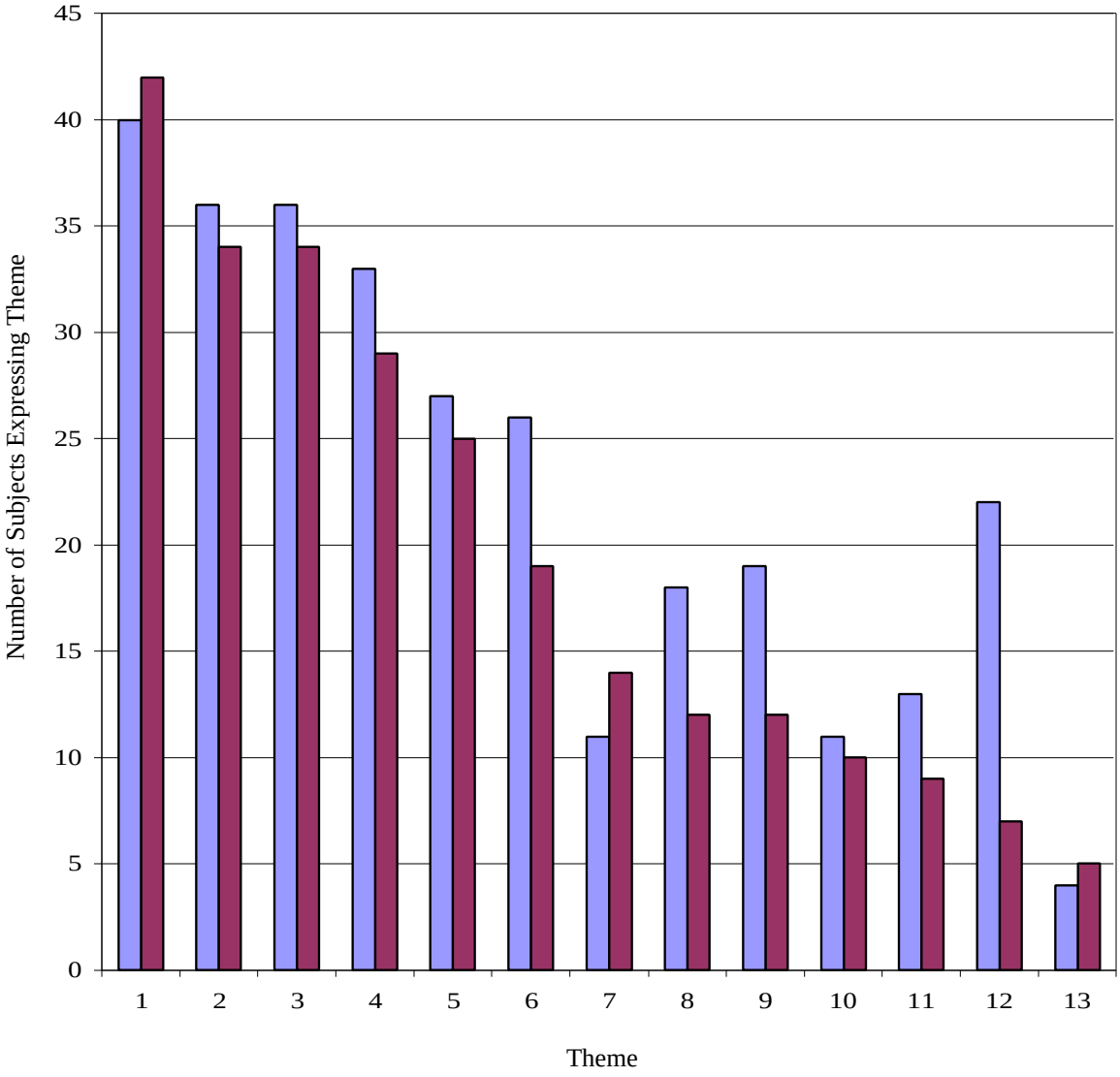
long to find them. I can't remember exactly…probably the preferences or whatever it was to change the red and gold, but those are just silly little things not necessarily important.

- The color thing…I wouldn't not use it because of that, but I would probably go to help or email feedback and ask how I would do that because I would want to get it personalized.
- I think it's a cool idea, but I don't know that I would personally ever use it. I don't think I have a need personally to have all these things open. The way I have it now, I have my favorites and I can look at one and it's all organized that way where I can go to one page and it brings it up for me.
- Putting the links on the page [was the most enjoyable task]. Those are the ones that you get to choose. The least was the bookmarks. I was getting mad at that one. Even though I didn't get the color thing, I still liked that task best. Actually putting the links on the screen are what is going to matter.

4.2. Observer Agreement Regarding Users' Rationales

In an effort to demonstrate reliability and ensure that the constructed theme categories were not merely the opinion of the primary researcher, two external raters were employed to review the qualitative data. The researcher and the external raters each independently reviewed the entire set of 45 transcripts and used the theme descriptions presented earlier in Table 4.3 to identify instances of themes for each subject. The first external rater had no prior experience in Human Computer Interaction and did not receive any training in recognizing the given themes other than the theme descriptions themselves. The second external rater had graduate level training in Human Computer Interaction and participated in a brief training activity in which she and the primary researcher each rated a sample of transcripts and compared their results prior to her rating the entire set of transcripts. The relative frequency of the 13 themes as agreed upon between the primary researcher and each external rater is presented graphically in Figure 4.1.

Figure 4.1. Relative frequency of rationales based on external rating activities



Themes

1 Utility and usefulness are key
2 Just give me some time
3 Clear and helpful
4 It's my fault, not the system's
5 Time equals satisfaction
6 Everything in its place
7 The paradox of information quantity
8 I can see its potential
9 Feedback is important
10 If someone would just tell me
11 Confusing terminology
12 Getting it right…finally!
13 I wouldn't do that anyway

For each theme in Figure 4.1, two vertical bars represent the number of subjects for which the external raters agreed with the primary researcher that the theme was present. For example, in the case of *Theme 5: Time equals satisfaction*, the first external rater agreed with the primary researcher that this theme was present for 27 of the 45 subjects in the study. Similarly, the second external rater agreed with the primary researcher that *Theme 5* was present for 25 of the 45 subjects in the study. The table shows that the primary researcher and the external raters found the first five themes to be present for more than half of all subjects in the study. The remaining eight themes were considerably less frequent with Themes 10, 11, 12 and 13 each being commonly identified in roughly ten or fewer transcripts in the study.

To gain a more complete view of the level of agreement between the primary researcher and the external raters, contingency tables were produced (see Tables 4.4 and 4.5). In addition to presenting those instances where the researcher and the external rater agreed that a given theme was present, the contingency tables reveal where raters agreed that themes were absent, and where raters disagreed.

Table 4.4. Contingency table demonstrating observer agreement for rating activity #1

Identification of Themes for Primary Researcher: Themes	Identification of Themes for External Rater 1: 1	2	3	4	5	6	7	8	9	10	11	12	13	Null
1	40 (1)													2
2		36 (4)												2
3			36 (2)											0
4				33 (2)										2
5					27 (2)									5
6						26 (8)								0
7							11 (17)							5
8								18 (6)						2
9									19 (5)					2
10										11 (32)				1
11											13 (23)			3
12												22 (8)		0
13													4 (19)	8
Null	2	3	7	8	11	11	12	19	19	1	6	15	14	-

in diagonal – number of subjects for which observers agreed that theme was present
(#) in diagonal – number of subjects for which observers agreed that theme was absent
Null column – number of subjects for which the researcher identified theme and external rater did not
Null row – number of subjects for which the external rater identified theme and the researcher did not

Themes

1 Utility and usefulness are key
2 Just give me some time
3 Clear and helpful
4 It's my fault, not the system's
5 Time equals satisfaction
6 Everything in its place
7 The paradox of information quantity
8 I can see its potential
9 Feedback is important
10 If someone would just tell me
11 Confusing terminology
12 Getting it right…finally!
13 I wouldn't do that anyway

Table 4.5. Contingency table demonstrating observer agreement for rating activity #2

Identification of Themes by Primary Researcher	Identification of Themes by External Rater 2													
Themes	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**Null**
1	42 (2)													0
2		34 (4)												4
3			34 (6)											1
4				29 (9)										6
5					25 (9)									6
6						19 (16)								7
7							14 (21)							2
8								12 (21)						7
9									12 (18)					9
10										10 (31)				2
11											9 (28)			5
12												7 (24)		14
13													5 (33)	6
Null	1	3	4	1	5	3	8	5	6	2	3	0	1	-

in diagonal – number of subjects for which observers agreed that theme was present
(#) in diagonal – number of subjects for which observers agreed that theme was absent
Null column – number of subjects for which the researcher identified theme and external rater did not
Null row – number of subjects for which the external rater identified theme and the researcher did not

Themes

1 Utility and usefulness are key
2 Just give me some time
3 Clear and helpful
4 It's my fault, not the system's
5 Time equals satisfaction
6 Everything in its place
7 The paradox of information quantity
8 I can see its potential
9 Feedback is important
10 If someone would just tell me
11 Confusing terminology
12 Getting it right…finally!
13 I wouldn't do that anyway

The contingency tables provide us with a comprehensive summary of the two rating activities and allow us to make comparisons between them. Using Table 4.5 as an example, the reader finds the external rater's record presented in the columns of the table, while the primary researcher's record is presented in the rows of the table. The non-bracketed values displayed in the diagonal of the table represent the number of subjects for which both raters agreed that the given theme was present. The bracketed values displayed in the diagonal represent the number of subjects for which both raters agreed that the theme was absent. The null row indicates those cases where the external rater identified a theme, but the researcher did not. The null column indicates those cases where the researcher identified a theme, but the external rater did not. For example, for *Theme 4: It's my fault, not the system's*, the two raters were in agreement that 29 subjects expressed this theme one or more times within their transcripts. The raters also agreed that for nine subjects there was an absence of Theme 4. The null row reveals that the external rater identified Theme 4 in one additional transcript where the researcher did not. Meanwhile, the null column shows that the researcher identified Theme 4 in six transcripts where the external rater did not.

A cursory analysis of the diagonal portions of Table 4.4 and Table 4.5 reveals that the two rating activities produced highly similar accounts of whether themes were present or absent (e.g. Themes #1, 2, 3, 4, 5, 7, 10 and 11). Certain themes were less consistently agreed upon (e.g. Themes #6, 8, 9, 12, and 13), suggesting that these themes may require further refinement or that the raters varied in their ability to accurately recognize them. An examination of the null row in Table 4.4 clearly demonstrates a tendency for the first

external rater to identify the presence of themes where the primary researcher did not. The null row and column in Table 4.5 reveal a more evenly distributed level of disagreement between the primary researcher and the second external rater. One exception to this is the substantial disagreement associated with *Theme 12: Getting it Right…Finally* in which the primary researcher identified this theme on 14 occasions where the external rater did not.

Based on these data from the contingency tables (Tables 4.4 and 4.5), the kappa statistic (Cohen, 1960) was applied to measure the level of observer agreement corrected for the probability of chance agreement. The kappa coefficient (κ) is calculated as:

$$\kappa = \frac{(P_o - P_e)}{(1 - P_e)}$$

where: P_o is the observed count of agreement (uncorrected for chance)
P_e is the expected count of agreement (i.e. chance agreement)

The kappa coefficient ranges between 0 (when the agreement is no better than chance) and 1 (when there is perfect agreement). It may also be a negative value when considerable disagreement occurs between observers and P_e ends up being greater than P_o. While there is some contention that conclusions regarding observer agreement based on kappa results fail to acknowledge the statistical dependence of raters and the need for an explicit model of rater decision-making (Uebersax, 1987), the method has a long-standing tradition of use in social sciences and medical research (Cohen, 1960; Frick & Semmel, 1978; Kraemer, 1982). The results of the kappa calculation for this study are presented in Table 4.6.

Table 4.6. Kappa coefficient for rating activities #1 and #2

		Rating activity #1			Rating activity #2		
#	**Theme**	P_o	P_e	κ	P_o	P_e	κ
1	Utility and usefulness are key	0.91	0.88	0.29	0.98	0.90	0.79
2	Just give me some time	0.89	0.75	0.55	0.84	0.72	0.44
3	Clear and helpful	0.84	0.77	0.31	0.89	0.69	0.64
4	It's my fault, not the system's	0.78	0.73	0.18	0.84	0.59	0.62
5	Time equals satisfaction	0.64	0.65	0.00	0.76	0.56	0.44
6	Everything in its place	0.76	0.55	0.46	0.78	0.50	0.56
7	The paradox of information quantity	0.62	0.50	0.25	0.78	0.50	0.55
8	I can see its potential	0.53	0.46	0.13	0.73	0.52	0.45
9	Feedback is important	0.53	0.48	0.11	0.67	0.51	0.32
10	If someone would just tell me	0.96	0.61	0.89	0.91	0.61	0.77
11	Confusing terminology	0.80	0.52	0.58	0.82	0.59	0.57
12	Getting it right…finally!	0.67	0.50	0.34	0.69	0.52	0.35
13	I wouldn't do that anyway	0.51	0.55	-0.08	0.84	0.69	0.50

From Table 4.6, we see that the observed agreement (P_o) for both rating activities was quite high (over 70% for 11 of the 13 themes in the second rating activity). Once corrected for chance or expected agreement (P_e), a moderating effect on the observed agreement was realized. For the first rating activity, the majority of themes continued to demonstrate a greater than chance level of agreement ($\kappa > 0$) between the first external rater and the primary researcher. However, several themes (#4, 5, 8, 9, 13) either approached $\kappa = 0$ or became negative in value, suggesting that observer agreement for these themes may have occurred due to chance alone. For the second rating activity, the kappa results demonstrated a much stronger level of agreement overall between the primary researcher and the second external rater, lending support to the reliability of the 13 themes.

Although the frequency counts presented earlier in this chapter (see Figure 4.1), revealed that not all themes are equally common, the level of observer agreement demonstrated by the contingency tables and the kappa results (particularly for the second external rating activity) supports the existence of the thirteen themes as valid rationales that users provide for their satisfaction and that are identifiable by reviewers. Based on these levels of observer agreement, all thirteen themes were retained for the purpose of reporting and discussion in this dissertation.

5. Examining the Relationship Between Usability and Users' Rationales for Satisfaction

The question of whether users' rationales for their subjective satisfaction ratings are related to their performances of effectiveness and efficiency was examined through a combination of a correlational analysis, and an analysis of the descriptive statistics for users' effectiveness, efficiency, and frequency of stated rationales. The following sections present the findings of each analysis.

5.1. Correlating Efficiency, Effectiveness and Satisfaction

Pearson's Product Moment correlational analysis was performed to determine whether predictive relationships existed between satisfaction, effectiveness (success), efficiency (time spent), and the various demographic factors gathered to help describe the sample of users who participated in this study. This analysis produced six different correlations significant at the .01 level and five additional correlations significant at the .05 level. A summary of this analysis is presented in Table 4.7.

Table 4.7. Bivariate correlations between usability attributes and demographics

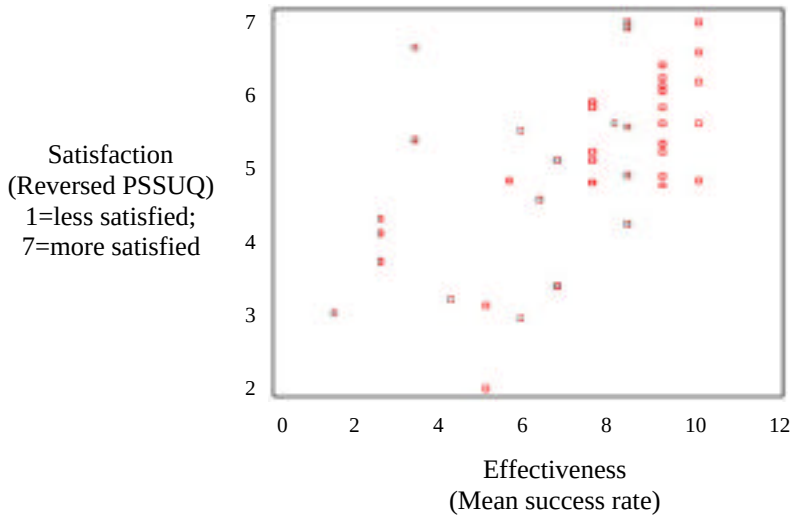
Variable	Satisf	Success	Time Spent	Age	Yrs	Soft Profic	Web Activity
Satisfaction	1.000	.593**	-.452**	-.586**	-.382**	.303*	.199
Success (effectiveness)		1.000	-.394**	-.369*	-.013	.282	.395**
Time spent (efficiency)			1.000	.327*	.088	-.302*	-.139
Age				1.000	.260	-.151	-.143
Years at University					1.000	-.058	.047
Software Proficiency						1.000	.299*
Web Activities							1.000

* Correlation is significant at the 0.05 level (2-tailed)
** Correlation is significant at the 0.01 level (2-tailed)

Among the significant correlations discovered in this analysis were strong positive and negative relationships between satisfaction, efficiency, and effectiveness. This was in contrast to recent studies that have demonstrated no significant correlation between these three attributes of usability (Frøkjær et al., 2000; Walker et al., 1998).

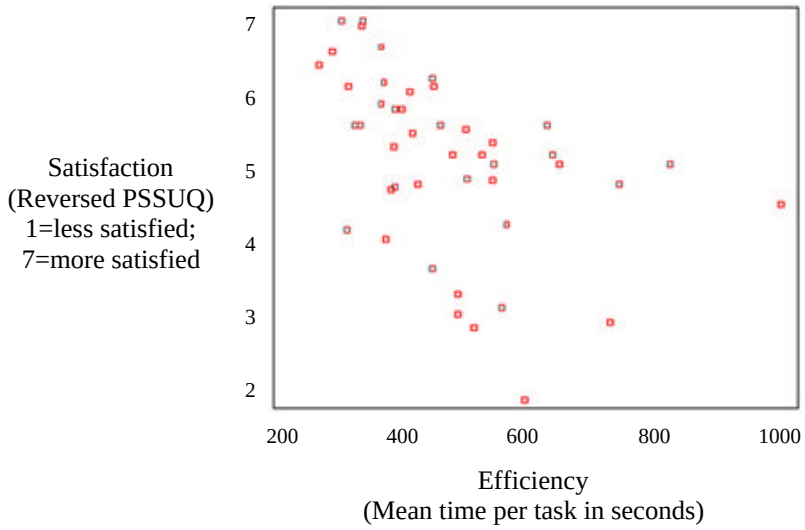
In this study, users' satisfaction ratings were positively correlated (.593, $p < .01$) with their effectiveness or level of success in completing tasks (see Figure 4.2). This correlation reveals that users who were most successful in completing tasks with the system were also the most satisfied with the system overall. Conversely, those users who failed to complete tasks were less satisfied with the system overall.

Figure 4.2. Bivariate scatterplot of satisfaction and effectiveness (r = .593)



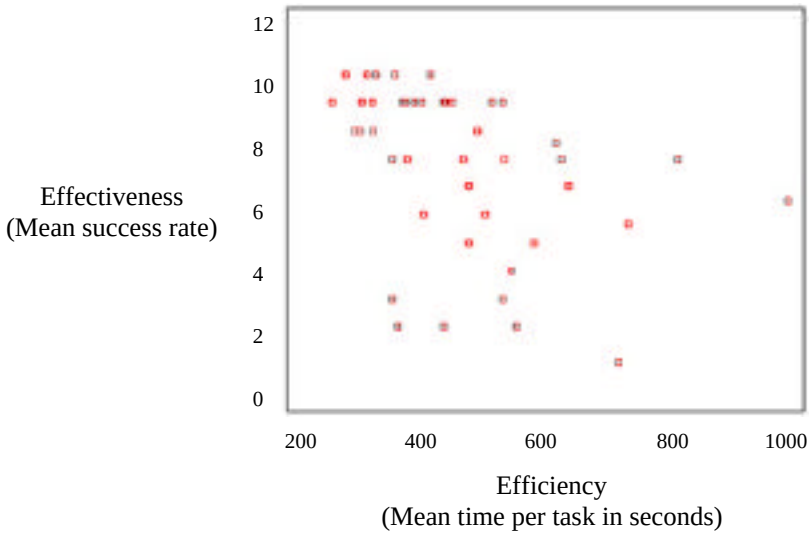
Users' satisfaction levels were negatively correlated (-.452, $p < .01$) with their efficiency or time spent completing tasks (see Figure 4.3). This correlation reveals that the more time required for a user to complete a given set of tasks using the system, the less satisfied the user was with the system overall. Conversely, the most satisfied users were those who completed the given tasks most quickly.

Figure 4.3. Bivariate scatterplot of satisfaction and efficiency (r = -.452)



A significant negative correlation (-.394, $p < .01$) was also found between users' efficiency and effectiveness (see Figure 4.4). This correlation reveals that those users who were more successful at completing tasks tended to complete them more quickly on average. Conversely, users who spent a long time performing tasks tended to be less successful at completing them.

Figure 4.4. Bivariate scatterplot of efficiency and effectiveness (r = -.394)



Additional significant correlations were also discovered between users' satisfaction and each of the following: age, software proficiency and years spent at university. Users' satisfaction was negatively correlated with their age (-.586, $p < .01$), revealing that the older a user was, the less satisfied they were with the OneStart portal. User's satisfaction and software proficiency were positively correlated (.303, $p < .05$), indicating that the more experience users had with different types of software, the more satisfied they were with the OneStart system. Finally, users' satisfaction and the number

of years spent at university were negatively correlated (-.382, $p < .01$), revealing that those with less university experience were more satisfied with the OneStart portal.

Given the highly significant correlations in this study between users' satisfaction and the variables of effectiveness, efficiency, age, and years at university, there was the chance that one or more of these strong relationships may have masked or obscured any actual differences between the test groups in terms of their overall satisfaction with the system. To investigate this possibility, a post-hoc Analysis of Covariance (ANCOVA) was performed to determine if the statistical adjustment of the satisfaction means might expose some real difference between the groups. The results of the ANCOVA revealed no significant difference between the groups, further supporting the results of the initial ANOVA that suggest users' satisfaction ratings are not directly affected by the use of priming.

5.2. Relating Users' Rationales and Objective Performances

The question of whether users' rationales for their satisfaction ratings are related to their objective performances of efficiency and effectiveness was addressed by an analysis of descriptive statistics. Inferential statistics (i.e., ANOVAs with F tests) were not used since: 1) Observations were not independent (multiple codes often occurred for each subject); 2) The relatively large number of F tests (39) would increase the probability of committing Type I errors in this study; and 3) the numbers of subjects associated with each rationale were often highly unequal, increasing the likelihood of violation of the homogeneity of variance assumption (Kirk, 1995). Given this decision, the interpretations regarding the relationships between rationales and users' performances

are intended to be descriptive only and are not proposed for generalization to the larger population.

Among the 13 rationales, two groups were identified: *non- directional* and *directional*. Non-directional rationales were characterized by a consistent response from all users who expressed that theme. These included rationales such as *Theme 1: Utility and usefulness are key* where all users who expressed this theme indicated that they felt the system's utility was important to their satisfaction, and *Theme 4: It's my fault, not the system's* where all users who expressed this theme indicated that they felt they were to blame for any difficulties they experienced. All non-directional rationales were entered into SPSS 10 based on the results of the observer agreement activity and coded as present (1) or absent (0).

Directional rationales were characterized by a mixture of positive and negative responses from users. For example, some users who expressed *Theme 3: Clear and helpful* commented that they felt the system was very self-explanatory and clear (i.e. positive), while other users stated that it was unclear and lacked sufficient instruction (i.e. negative). Similarly, *Theme 5: Time equals satisfaction,* included both negative comments about the system being too slow, and positive comments about it being pleasantly fast. In both cases, the general theme about time being important to satisfaction was clearly present. For directional rationales, data were entered into SPSS 10 as one of positive (1), absent (0), or negative (-1). The descriptive statistics demonstrating these relationships are presented in Table 4.8.

Table 4.8. Descriptive statistics for rationales, efficiency, effectiveness and satisfaction

			N	Mean	SD
Theme 1:	Utility and usefulness are key				
	Time per task	Present	42	467.901	149.004
		Absent	3	460.583	85.778
	Effectiveness	Present	42	.753	.221
		Absent	3	.417	.289
	Satisfaction	Present	42	5.173	1.116
		Absent	3	4.306	.431
Theme 2:	Just give me some time				
	Time per task	Present	34	473.486	154.927
		Absent	11	448.646	113.116
	Effectiveness	Present	34	.767	.212
		Absent	11	.619	.289
	Satisfaction	Present	34	5.243	.958
		Absent	11	4.722	1.450
Theme 3:	Clear and helpful *(directional)*				
	Time per task	Positive	11	458.357	142.133
		Absent	11	460.227	184.102
		Negative	23	475.182	131.658
	Effectiveness	Positive	11	.752	.267
		Absent	11	.785	.208
		Negative	23	.694	.241
	Satisfaction	Positive	11	5.561	.933
		Absent	11	5.812	.779
		Negative	23	4.570	1.053
Theme 4: It's my fault, not the system's					
	Time per task	Present	29	504.716	158.984
		Absent	16	399.803	83.242
	Effectiveness	Present	29	.711	.241
		Absent	16	.766	.236
	Satisfaction	Present	29	5.070	1.153
		Absent	16	5.197	1.037
Theme 5:	Time equals satisfaction *(directional)*				
	Time per task	Positive	4	366.656	111.935
		Absent	21	484.883	144.598
		Negative	20	469.222	149.088
	Effectiveness	Positive	4	.771	.300
		Absent	21	.668	.275
		Negative	20	.788	.172
	Satisfaction	Positive	4	6.108	.669
		Absent	21	5.061	1.273
		Negative	20	4.974	.900

			N	Mean	SD
Theme 6:	Everything in its place *(directional)*				
	Time per task	Positive	14	400.482	91.492
		Absent	25	493.075	163.373
		Negative	6	516.663	128.079
	Effectiveness	Positive	14	.780	.218
		Absent	25	.736	.222
		Negative	6	.593	.329
	Satisfaction	Positive	14	5.767	.732
		Absent	25	4.941	1.143
		Negative	6	4.324	.975
Theme 7:	The paradox of information quantity *(directional)*				
	Time per task	Positive	3	567.917	49.420
		Absent	34	436.707	119.848
		Negative	8	560.228	212.131
	Effectiveness	Positive	3	.500	.220
		Absent	34	.756	.231
		Negative	8	.711	.253
	Satisfaction	Positive	3	4.504	1.102
		Absent	34	5.194	1.122
		Negative	8	5.010	1.078
Theme 8:	I can see its potential				
	Time per task	Present	12	532.279	165.829
		Absent	33	443.826	131.555
	Effectiveness	Present	12	.676	.297
		Absent	33	.750	.214
	Satisfaction	Present	12	4.821	1.392
		Absent	33	5.223	.981
Theme 9:	Feedback is important *(directional)*				
	Time per task	Positive	5	415.025	78.387
		Absent	33	478.544	144.927
		Negative	7	452.363	186.661
	Effectiveness	Positive	5	.767	.253
		Absent	33	.729	.236
		Negative	7	.714	.272
	Satisfaction	Positive	5	5.800	.439
		Absent	33	4.992	.975
		Negative	7	5.206	1.800
Theme 10:	If someone would just tell me				
	Time per task	Present	10	503.848	222.049
		Absent	35	457.004	116.863
	Effectiveness	Present	10	.654	.224
		Absent	35	.752	.240
	Satisfaction	Present	10	4.853	1.529
		Absent	35	5.191	.964

			N	Mean	SD
Theme 11: Confusing terminology					
	Time per task	Present	9	490.650	199.760
		Absent	36	461.605	130.914
	Effectiveness	Present	9	.819	.164
		Absent	36	.708	.250
	Satisfaction	Present	9	4.914	1.137
		Absent	36	5.166	1.105
Theme 12: Getting it right…finally!					
	Time per task	Present	7	477.274	155.226
		Absent	38	465.597	145.173
	Effectiveness	Present	7	.845	.176
		Absent	38	.709	.244
	Satisfaction	Present	7	5.579	1.138
		Absent	38	5.030	1.090
Theme 13: I wouldn't do that anyway					
	Time per task	Present	5	365.675	46.152
		Absent	40	480.131	148.360
	Effectiveness	Present	5	.800	.162
		Absent	40	.722	.246
	Satisfaction	Present	5	5.517	.771
		Absent	40	5.065	1.135

From Table 4.8, we find consistent relationships between several themes and users' effectiveness, efficiency, and satisfaction. While the descriptive statistics alone do not allow us to suggest that these patterns hold beyond this study, they do help to support the overall finding that effectiveness, efficiency, and satisfaction are correlated and they suggest opportunities for future validation research. The following paragraphs identify some of the stronger patterns found through an analysis of Table 4.8.

5.2.1 Relating Rationales to Users' Efficiency

A pattern between the presence or absence of the thirteen rationales and users' efficiency was evident in several places. In general, the subjects who were more efficient (low time per task) tended to express more positive themes and comments, while subjects

who were less efficient (high time per task) tended to express more negative themes and comments. In some cases, the differences in mean time per task were negligible (e.g. Themes 1, 2, 3, 11, 12), while in others, the differences were substantial considering the standard deviations (e.g. Themes 4, 5, 6, 8, 13), In both examples, the unequal number of subjects makes it difficult to know whether the differences, or lack thereof, would be maintained for a larger sample. Table 4.9 illustrates the strongest relationships found between the themes and users' efficiency and indicates whether subjects' comments were negative or positive in those cases where the themes were directional.

Table 4.9 Relating users' rationales and mean efficiency

	Theme	Low efficiency (high time/task)	High efficiency (low time/task)
1	Utility and usefulness are key		
2	Just give me some time		
3	Clear and helpful		
4	It's my fault, not the system's	√	
5	Time equals satisfaction	√ (-)	√ (+)
6	Everything in its place	√ (-)	√ (+)
7	The paradox of information quantity	√ (-/+)	
8	I can see its potential	√	
9	Feedback is important		
10	If someone would just tell me		
11	Confusing terminology		
12	Getting it right…finally!	√	
13	I wouldn't do that anyway		√

From Table 4.9, we see that subjects who were highly efficient tended to comment frequently in a positive manner about the system's efficiency (Theme 5), and its organization and layout (Theme 6). Conversely, subjects who demonstrated low

efficiency provided the majority of negative comments about these two themes. Subjects who were inefficient were also more likely to blame themselves (Theme 4), to identify the system's future potential (Theme 8), and to find themselves taking longer to succeed at tasks (Theme 12). From these results, we see that several of the rationales given by users for their satisfaction are closely linked to the speed at which they were able to perform using the system. This finding lends further support to the strong negative correlation between users' overall satisfaction and efficiency (-.452, p <.01).

5.2.2 Relating Rationales to Users' Effectiveness

Some strong relationships were also found between the presence or absence of the 13 rationales and users' effectiveness (see Table 4.10). The patterns suggest that subjects who were more effective at completing tasks tended to perceive some utility offered by the system (Theme 1), they were optimistic about being able to use it given some additional practice time (Theme 2), they commented positively on the organization of its elements (Theme 6), and they tended to persist at completing tasks (Theme 12). Conversely, those subjects who were less effective tended to comment negatively about the system's organization (Theme 6), and wished that they could receive some personal assistance with the system (Theme 10). As with the analysis of themes and subjects' efficiency, the unequal number of subjects reporting each theme makes it difficult to know whether the patterns mentioned here would be maintained for a larger sample. The patterns identified, however, do provide additional support for the strong positive correlation found between users' overall satisfaction and effectiveness (r = .593, p <.01).

Table 4.10 Relating users' rationales and mean effectiveness

	Theme	Low effectiveness	High effectiveness
1	Utility and usefulness are key		√
2	Just give me some time		√
3	Clear and helpful		
4	It's my fault, not the system's	√	
5	Time equals satisfaction		
6	Everything in its place	√ (-)	√ (+)
7	The paradox of information quantity	√ (+)	√ (-)
8	I can see its potential		
9	Feedback is important		
10	If someone would just tell me	√	
11	Confusing terminology		√
12	Getting it right…finally!		√
13	I wouldn't do that anyway		√

CHAPTER V.

DISCUSSION

1. Overview

This study set out to refine the methodology of usability testing by investigating the particular challenge of measuring and understanding users' satisfaction with a web-based portal application. In an effort to address the commonly perceived phenomenon of inflated satisfaction ratings following usability tests (Nielsen & Levy, 1994; Root & Draper, 1983; Teague et al., 2001), this study compared the effects from two different priming techniques with that of a control group that received no primer. Forty-five undergraduate students enrolled at Indiana University with similar computing backgrounds participated in the study. Randomly assigned to one of three different test groups, these students performed a series of tasks using Indiana University's OneStart portal. Data on their satisfaction were collected either during the course of the session and/or after the session depending on their group membership. Through a combination of think-aloud protocols and post-test interviews, students were also asked to describe their rationales for rating satisfaction as they did.

The results of this study revealed that the priming technique had no significant impact on users' satisfaction ratings. However, several highly significant correlations were found between various pairings of users' effectiveness, efficiency, satisfaction, and demographic characteristics. Additionally, the qualitative comments from students in the study revealed thirteen common rationales for their satisfaction ratings, lending support to previous literature in some cases, while in others, contributing new empirical data for consideration. This chapter reflects on the findings from this study and discusses their

implications. Limitations of this study and suggestions for future research in this area are presented at the end.

2. The Administration of User Satisfaction Measures

2.1. Success or Failure?

A famous story about the American inventor, Thomas Edison, describes his experience inventing the incandescent light bulb. According to this story, after his 750th unsuccessful attempt at creating a working bulb, Edison was asked why he bothered to keep trying when he had "failed" so many times. The inventor responded quite simply that he never viewed his trials as failures. Rather, he preferred to think of them as having successfully discovered 750 ways that a light bulb would not work.

How one interprets the results to the first research question of this dissertation depends on ones' perspective. The results of the ANOVA between the three test groups indicated that no significant difference exists between users' satisfaction ratings, regardless of whether users were primed to consider the satisfaction instrument or not. One perspective, therefore, might be to view this study as failing to identify a new and improved approach to gathering satisfaction data. An alternative perspective, however, might be to adopt Edison's philosophy, in which case the study would be viewed as successfully eliminating one more contender from the list of possible strategies for improving the usability testing process. While the former perspective laments the lack of an addition to current methodology, the latter applauds the dismissal of one more viable candidate, demonstrating that refinements to methodologies may occur in multiple ways.

2.2. Implications of Experimental Findings

If one considers only the quantitative data, the results of this study seem rather clear. Although priming has been shown to produce observable effects in other contexts (Head, 1991; Moss & Lawrence, 1997; Salancik & Pfeffer, 1978; Yi, 1993), the priming techniques chosen for this study appeared to have no demonstrable impact on the way users rated their satisfaction in a usability test. Even though the techniques used in this study were intentionally designed to be less intrusive than methods previously mentioned in the literature (Teague et al, 2001), and they required little preparation given that they utilized previously validated instruments (ASQ and PSSUQ), the lack of a significant difference between them suggests that any time dedicated to their implementation is likely time better spent on other aspects of the usability evaluation.

There is one caveat to this conclusion. As mentioned in Chapter Four of this dissertation, subjects' qualitative reactions to the two priming methods used in this study were generally quite favorable. Subjects in Group B commonly stated that the "preview effect" of the advance primer helped them to understand what was expected of them and made them more comfortable as a result. Subjects in Group C commonly perceived that the concurrent priming activity helped them remember their experiences with the system when it came time to give their overall ratings in the end. As one subject explained,

> It's kind of like reading the questions at the end of the book before you read it just to know what you are looking for to pay special attention. I wasn't thinking about rating it, but I was thinking in my head about what I liked and didn't like.

The qualitative reaction of students to the priming techniques used in this study suggests that the real value of priming to the usability testing process may be in its ability

to make subjects feel more at ease and more confident in their ability to meet the expectations of the test, even if it doesn't have any quantifiable effect on users' ratings. Based on the results of this study, usability researchers should not expect to see the priming techniques used in this study affect subjects' post-test satisfaction ratings, but they may wish to consider incorporating them to help make the test session more comfortable and enjoyable for subjects.

2.3. Explaining the Lack of Effect

Although the analysis of the first research question suggests that priming techniques have no measurable impact on users' satisfaction ratings in a usability test, a few possible explanations for this lack of effect are worth pondering. One possibility is that any effects of priming are outweighed by users' reluctance to criticize a system. Even when usability evaluators make it clear that they have no ownership of the system being tested, many users appear unable to separate between the actual events of a usability test and the perceived *social expectancy* that may influence their retelling (Csikszentmihalyi & Csikszentmihalyi, 1988). This phenomenon was clearly evident in this study as subjects commonly apologized for giving what they perceived to be severe ratings, even when the researcher viewed them as rather positive given their observed performances.

A second possibility is that the degree or level of priming applied in this study may have been insufficient to demonstrate priming's potential effect on users' post-test ratings. Although the priming techniques chosen for this study were based on ideas from previous research (Moss & Lawrence, 1997; Root & Draper, 1983; Teague et al, 2001;

Yi, 1993), there is the possibility that other methods of priming could be identified that would realize some effect on users' satisfaction ratings.

3. Rationalizing Users' Satisfaction

The parsimony in naming satisfaction as a primary attribute of usability is highly attractive to those who value clear and concise definitions. The multidimensional nature of users' satisfaction quickly becomes apparent, however, when users are asked to articulate their reasoning for rating their satisfaction with a given system. In this study, some interesting qualitative data were produced that illuminate how users justify their satisfaction ratings, even when they might be inconsistent with their performances. Of the 13 themes that emerged from this study, several provide empirical support for previous findings and opinions from the literature. In a few cases, they represent new and interesting possibilities for understanding how users evaluate the systems they use. Each rationale should be considered tentative given the specific context and audience for this study, and the lack of perfect observer agreement. Nevertheless, designers and developers of other web-based portal applications, together with usability practitioners and researchers stand to benefit from considering the rationales identified in this study. In the following sections, each of the 13 rationales is discussed and their implications for designers and usability researchers are considered.

3.1. Utility and Usefulness are Key

The most frequently occurring rationale for satisfaction provided by subjects in this study was that of a perceived *utility* or *usefulness* that the OneStart portal offered them over what they were previously accustomed to. Recorded as present in 40 or more of the 45 transcripts during each of the two rating activities, this finding lends empirical

support to previous literature that suggests utility (Shackel, 1986) and the extent to which a system meets users' expectations (Rushinek & Rushinek, 1986) are important contributors to users' satisfaction and ultimately, to usability.

In the case of the OneStart portal, users described its utility in numerous ways, including the ability to personalize the layout and organization of the system, the option to choose the content that was important to them, and the opportunity to have it "all in one place" for easy access and viewing. Together, these attributes effectively comprised the *relative advantage* of the OneStart portal, a critical factor in determining whether an innovation is adopted for use (Rogers, 1995). When asked whether they expected to use the OneStart portal in the future, the vast majority of subjects in this study indicated that they did intend to use it. Unfortunately, the actual adoption rate had to be left for speculation as this question had not been part of the original plan and the necessary permissions for subsequent contact with the subjects were not in place.

For designers and developers of future web-based portal applications, this rationale emphasizes the need for fully understanding the gap between what users currently have and what they desire to be able to accomplish their goals. In the case of the OneStart portal, users ranked *personalization* and *convenience* at the top of their list.

3.2. Just Give Me Some Time

Regardless of whether users performed well or experienced difficulty during this study, they frequently held the view that they could effectively use the system given enough time and practice. Users generally seemed to hold optimism for the usability of any system, but fully expected to confront a significant learning curve given their experiences with past systems. This resignation to accepting difficult systems and

adapting to what are often poor interface designs has been acknowledged frequently in the literature (Cooper, 1999; Nielsen, 2000; Norman, 1988). In his book, *Design, Form, and Chaos,* Paul Rand (1993) refers to a conditioning effect that users in general have undergone in our society.

> The public is more familiar with bad design than good design. It is, in effect, conditioned to prefer bad design, because that is what it lives with. The new becomes threatening, the old reassuring.

This rationale of granting a new system a grace period also lends support to Raskin's (1994) suggestion that users' satisfaction is largely a product of their familiarity with an interface. He suggests that *familiar* is what we really mean when we attempt to describe *intuitive* interfaces and that users' familiarity with a new software interface contributes directly to their increased comfort, productivity, and satisfaction. For designers of web-based applications, the implications of this rationale suggest that users' familiarity with other systems should be leveraged whenever possible and that creating an interface to be used perfectly the very first time may be less critical to users' satisfaction, than making sure that it can be learned and remembered with relative ease.

3.3. Clear and Helpful

This theme saw users identify the importance of clear instructions and effective help screens to their overall satisfaction with the system. As one of the directional themes identified in this study, this theme included both negative and positive examples. Those users who found the system to be self-explanatory or who engaged the help screens and found an answer to their problems tended to be more satisfied with the system. Those users who recalled instances where they found the instructions and help screens to be incomplete or even inaccurate, typically rated the system more severely.

The expectation for clear instructions and effective help has been noted before in the literature. Nielsen (1994) advocates the importance of constructive help in his list of usability heuristics.

> Even though it is better if the system can be used without documentation, it may be necessary to provide help and documentation. Any such information should be easy to search, focused on the user's task, list concrete steps to be carried out, and not be too large. (p.30)

One common observation related to this rationale was that users frequently embarked on completing a task without referring to the instructions provided. Once difficulties were experienced, however, these users would quickly return to the instructions to compare their approach with the recommended one. The implications of this observation and the rationale in general are that clear instructions and help features are critical to a users' satisfaction with a web-based portal system, and even when they initially go unused, their presence and ease of access are expected by users.

3.4. It's My Fault, Not the System's

The tendency for users to blame themselves for any difficulties they had with the system was highly evident in this study. Users commonly referenced their own lack of ability with computers and their previous difficulties with other systems as they described the challenges they confronted using the OneStart application. In several cases, users were convinced that, while they had experienced problems using the system, most other people would likely be able to use the system without difficulty. By perceiving themselves to be the primary source of the problem, several users were observed to rationalize their satisfaction even when they had been observed to make numerous errors and expend excessive amounts of both time and energy completing tasks. This rationale

lends further empirical support to the observations of this same phenomenon by others (Cooper, 1999; Nielsen, 1993; Norman, 1988). Despite the call to arms by these authors for users to stop showing such patience with the poor design of products, a tendency clearly remains for users to feel that they are the ones at fault.

The implications of this rationale are most important for those charged with evaluating the usability of systems. To accurately interpret users' satisfaction with a software application, evaluators need to consider the degree to which users blame themselves for flaws in the design of the system. In this study, the application of a verbal protocol (Ericsson & Simon, 1993) during the post-test satisfaction rating activity was extremely important to identifying this theme. Had the post-test satisfaction instrument been completed in silence, the tendency for users to blame themselves may have gone unrecognized. For questionnaires designed to be administered without the opportunity for verbal protocol, the inclusion of items that allow subjects to self-report their level of self-blame or self-efficacy may prove valuable.

3.5. Time Equals Satisfaction

On the surface, the relationship between the time spent completing tasks and users' satisfaction may seem like an obvious one; the more responsive and well-designed a system is, the more efficient the experience and the greater one's satisfaction. Indeed, earlier satisfaction research with client-based systems confirmed that a primary attribute of users' satisfaction was the system's response time (Rushinek & Rushinek, 1986). In more recent studies, however, the correlation between users' efficiency and satisfaction has been questioned as users have demonstrated preferences for systems that they performed least efficiently with (Bailey, 1995; Frøkjær et al., 2000; Walker et al., 1998).

This raises the question of whether efficiency continues to be related to users' satisfaction for recent technologies such as web-based portal applications? The findings from this study provide empirical support that it does.

Users who perceived the system to be responsive to their actions (e.g. loading new screens, displaying available options) were generally more satisfied than those users who perceived the system to be slow to respond. Similarly, users who were able to complete their tasks in fewer attempts were more satisfied than those who required multiple attempts. These qualitative findings were supported by the correlational analysis that revealed a highly significant negative correlation between users' satisfaction and the length of time spent performing tasks (-.452, $p < .01$). For the developers of portal applications, the implication of this relationship is straightforward; whenever design decisions can be made to improve the users' level of efficiency with the system, users' satisfaction with the system is likely to increase.

3.6. Everything in its Place

Based on observations from this study, one of the greatest challenges facing the designers of web-based portals is how to present a plethora of information in a manner that is both organized and aesthetically appealing. The never-ending effort to develop content-rich systems and increasingly complex applications often leads software engineers to justify their overly busy designs on the basis of *increased functionality* and *added value*. However, the relationship between the quantity of information and the quality of its presentation need not be so linear. As Tufte (1990) reminds us,"Clutter and confusion are failures of design, not attributes of information" (p.51).

One option designers have is to simplify. In web-based interfaces, there is the potential for a repetition of elements to ultimately add to the perception of clutter without adding any real value. Known as the *1+1=3 phenomenon* (Tufte, 1990), this situation manifests when the interaction of two or more elements on a screen creates a multiplying effect that increases the viewer's perception of visual noise and clutter. One example is when a series of two colored rows of a table separated by a row of white space assumes the appearance of three colored bars on the screen (hence the name, 1+1=3).

In this study, users frequently rationalized their satisfaction ratings (both high and low) with references to the organization and layout of the portal's interface. Being able to locate information in consistent locations on the screen, having similar units of information chunked or compartmentalized, and being able to scan information in a logical and efficient manner were all considered positive examples of this rationale for satisfaction. Conversely, users commented negatively with regard to the portal's organization whenever new windows unexpectedly popped open, extensive scrolling was required, or the combination of screen elements produced a cluttered effect. Addressing the implications of this theme requires that designers be acutely aware of and implement visual design principles such as proximity, contrast, repetition, and alignment (Mullet & Sano, 1995) in order to optimize the organizational appearance of their interfaces. Similarly, they must guard against visual design pitfalls such as the *1+1=3 phenomenon* (Tufte, 1990) to avoid presenting users with unwanted and dissatisfying visual noise.

3.7. The Paradox of Information Quantity

As touched on earlier with the description of *Theme 5: Everything in its place,* one of the major challenges facing information portals is how to incorporate large

amounts of information without overwhelming the user. This challenge was apparent in subjects' rationales for their satisfaction as they frequently indicated a love/hate relationship with the amount of information provided in a portal. On the one hand, subjects saw the portal as valuable due to the very fact that it was information rich. On the other hand, some subjects criticized the portal for "trying to do too much" and making their experience more difficult as a result.

To address this issue, portal designers need to strive for a minimalist design approach to their system interfaces, providing users with just enough to get them started (Carroll, 1990). Once again, Neilsen's (1994) heuristics provide some helpful advice on how to achieve an *aesthetic and minimalist design:*

> Dialogues should not contain information which is irrelevant or rarely needed. Every extra unit of information in a dialogue competes with the relevant units of information and diminishes their relative visibility. (p.30)

In the case of web-based portals, one approach toward a minimalist design might include providing only brief instructions with a *more about…* hyperlink in order to allow users the option to obtain additional information without forcing them to wade through extra details on every screen.

3.8. I Can See its Potential

This theme was characterized by users' tendencies to look beyond the difficulties they had using the system and to focus on the system's ultimate potential. Subjects were generally very optimistic that any problems they experienced during the study would be eliminated before the system was released. It is worth noting that these comments were always made prior to the subjects discovering that the system was already in production and available for use by the university population.

It is also worth noting that for this theme, both external raters demonstrated a high frequency of identifying this theme where the primary researcher did not. The majority of these disagreements proved to be a case of the external raters identifying *Theme 7: I can see its potential* where the primary researcher identified either *Theme1: Utility and usefulness* or *Theme 2: Just give me some time*. These disagreements suggest that, in its current form, this theme may lack sufficient distinction from Themes 1 and 2. Further validation of this rationale in other studies is recommended.

From a designer's perspective, this rationale suggests that it is important for new systems to accurately convey to novice users the benefits they have to offer. If users are aware of these benefits, they may be more likely to look beyond any frustrations they experience during their learning phase with the application.

3.9. Feedback is Important

The recommendation for providing users with timely, informative and corrective feedback is perhaps one of the most common guidelines in software design (Nielsen, 1994; Norman, 1988; Shneiderman, 1998). Therefore, it is not surprising to find that users frequently commented on this aspect of the OneStart portal as they rationalized their satisfaction ratings for the system. Users' comments reflected both the perceived presence and absence of adequate feedback depending on their individual experiences. Users who indicated a high level of satisfaction with the system tended to perceive the feedback as being effective and adequate, while users who rated their satisfaction lower tended to criticize the system for its lack of meaningful feedback.

Discriminating between this theme and *Theme 3: Clear and helpful* proved to be a challenge for the external raters in this study (please see Table 4.6 presented in Chapter

Four). This was evidenced by the low observer agreement between raters. Following a post-rating discussion with raters, it was found that since the concept of instruction is common to both themes, it was difficult in many cases for the raters to distinguish between users' comments regarding initial instruction and subsequent feedback in response to users' actions. This was made more of a challenge for the raters by the fact that they were unfamiliar with the system being discussed, thereby making it move difficult to know what exactly was being referred to at times.

For designers of web-based portals, this rationale reinforces the importance of previous guidelines for providing users with informative feedback. This appears particularly important for novice users of a system to allow them to become familiar with new systems and reach a state of competency.

3.10. If Someone Would Just Tell Me

Similar in many ways to *Theme 4: It's my fault, not the system's,* this theme represented users' tendencies to once again remove accountability from the system by providing alternative strategies for how they might have been able to use it more easily. While this theme was not expressed as frequently as some of the others in this study (≈ 25% of subjects noted Theme 10), it produced one of the highest levels of observer agreement between raters ($\kappa = 0.89$ and $\kappa = 0.77$) suggesting that the theme as described was both present and easily recognized. Unfortunately, there seems little direct action that designers can take in response to this theme, other than to strive to have their systems successfully serve to *walk the user through* the process of learning and using the interface.

3.11. Confusing Terminology

The theme of confusing terminology was frequently used to explain users' dissatisfaction with the system. In the case of the OneStart portal, this theme was characterized by references to unfamiliar terms such as *pages*, *channels,* and *themes*. The external rater's unfamiliarity with these terms made it equally challenging for him to identify examples of this theme from the transcripts. Despite this difficulty, the observer agreement for this theme was quite high ($\kappa = 0.58$ and $\kappa = 0.57$).

The principles of using natural language and avoiding technical jargon are not new to the field of usability. As one of his ten usability heuristics, Nielsen (1994) recommends that interfaces should demonstrate a *match between the system and the real world.*

> The system should speak the users' language, with words, phrases, and concepts familiar to the user rather than system-oriented terms. Follow real-world conventions, making information appear in a natural and logical order. (p.30)

According to the findings of this study, portal designers would do well to keep this heuristic in mind and refrain whenever possible from introducing new terminology where existing terms may suffice.

3.12. Getting it Right…Finally!

Perhaps more than any other theme identified in this study, this rationale represents a finding from usability research that has received little previous attention in the literature. The tendency for users to express their satisfaction with the completion of a task, even when the process had proven to be a struggle, represents somewhat of a paradox for usability researchers. If users' satisfaction ratings with a system include a

component of achievement or triumph over a challenging interface, the most usable systems may not be the most satisfying.

Past observations have indicated that subjects in usability tests tend to persist at tasks longer than they normally would on their own (Nielsen, 1993) and that people in various contexts may become immersed in a state of *flow* as they strive to reach their goals (Csikszentmihalyi, 1990). What remains to be understood is how the positive, immersive experience that is often associated with mastering a software application interacts with the potentially frustrating and negative experience of learning and using the software to accomplish one's goals.

In the current study, users frequently responded that their most enjoyable task was one that they had visibly struggled with, but eventually completed successfully. The following excerpt illustrates one such example as the subject acknowledges that succeeding on the first try was satisfying, but successfully completing a task that required more investment on her part was even better.

> I liked succeeding on my first time. It's nice. It's like I can do this. The second time, it's like, well, I just need practice. It's just kind of frustrating. There is still a lot of satisfaction there because you're like 'It's like I beat it, I gotcha'. But the first time you get it right, it's like "I'm so smart'.

As with *Theme 9: Feedback is Important*, some caution needs to be exercised in considering this rationale due to the relatively low level of observer agreement between raters. Usability researchers would do well, however, to be aware of this phenomenon as they collect and interpret user satisfaction data. Requiring users to perform think-aloud protocols during the completion of the satisfaction instrument can assist greatly in understanding users' ratings.

3.13. I Wouldn't Do That Anyway

The final theme identified in this study and one that deserves some consideration from usability researchers was the tendency for users to rationalize their satisfaction based on the relevance of the tasks they had performed. Although this theme varied in its levels of observer agreement between rating activities, this may be partially explained by the post-activity discussion with the first external rater revealing his misinterpretation of this theme. Rather than identifying examples where subjects expressed ambivalence about tasks that they had been unable to complete, the external rater interpreted this rationale to include any indication that users had disliked a given task used in the test. Since users frequently commented about certain tasks being frustrating to complete, it is not surprising to find that the external rater had identified 14 examples of this rationale that went uncounted by the primary researcher. The second external rater appeared to be much clearer about this theme, although her observed frequency of appearance was lower overall.

The implications of this rationale for usability practitioners are that the tasks included in a usability test must be carefully chosen to reflect not only what the stakeholders of the application feel are important and authentic, but also what users are going to find meaningful and relevant to their experience with the application. Tests which include tasks that are perceived to be irrelevant by users, may be more likely to produce inflated satisfaction ratings that are inconsistent with users' observed experiences.

4. The Relationship Between Satisfaction and Objective Performance

4.1. Correlating Efficiency, Effectiveness, and Satisfaction

The third and final research question in this study found significant correlations between each pairing of users' subjective satisfaction, efficiency (time spent per task) and effectiveness (success rate). Satisfaction was found to be negatively correlated with time spent (-.452, p <.01) and positively correlated with effectiveness (.593, p <.01). Similarly, time spent (efficiency) and effectiveness were found to be negatively correlated (-.394, p <.01).

At a glance, these relationships seem intuitive. Users are more satisfied when they perform quickly with a system, they are more satisfied when they perform accurately with a system, and there is a strong tendency for more efficient performances to be more accurate as well. While each of these relationships may seem obvious, and enjoy some support from previous literature (Nielsen & Levy, 1994), they stand in stark contrast to more recent results suggesting that efficiency, effectiveness, and users' subjective satisfaction are not correlated (Frøkjær et al., 2000; Walker et al., 1998). They also represent what have otherwise been elusive findings for those hoping to make a business case for users' satisfaction with information systems (Melone, 1990).

There are a number of possibilities for the discrepancy between the correlations found in this study and missing from others. One possible factor may have to do with the complexity of the system. While web-based enterprise portals represent high-density information environments and typically provide users with a wide variety of personalization features, they may not constitute the type of *complex* systems that Frøkjær et al. (2000) describe in their research.

A second possibility for the difference in correlations found may be the result of different methodologies used. While this study required subjects to use a single application and rate their satisfaction levels with it following a standard usability test, the other studies implemented a comparative study in which users had the luxury of more objectively evaluating each system as it related to the other. This may have influenced users to rate their satisfaction differently than had they used only a single system.

A third possible factor for the different correlational findings may have to do with the statistical requirement for variance within the variables being examined. For example, in a case with little or no variance where every subject rated their satisfaction with the system about the same (either high or low), there is no opportunity for a correlation to exist – one could not predict users' satisfaction based on another variable since all users rated their satisfaction the same regardless of the other variable. In the current study, a sufficient amount of variance existed for the variables of satisfaction, efficiency, and effectiveness. Other studies that have shown no significant correlations between these variables may not have benefited from such conditions.

Finally, the significant correlations found in this study that indicate a strong relationship between users' satisfaction and their objective performances may be evidence that people are gradually becoming more sensitive to software usability issues. While users in this study remained generally optimistic and positive about the OneStart portal despite the usability problems they may have encountered, many of them were quick to identify and discuss what they perceived to be usability problems, often relating them to their previous experiences with other systems. In the case of those students who had recently graduated from the W200: Computers in Education course, there may have

been some exposure to usability concepts during their coursework. In the case of those subjects who had just begun the W200 course at the time of this study, any awareness of usability issues would have had to come from their own prior experiences with using software. From the research literature, we know that usability evaluators with experience in human-computer interaction typically do a better job of identifying usability problems through heuristic evaluations than evaluators without similar experience (Nielsen, 1992). It may be that the more aware test participants are of software usability issues, the more likely they are to identify those factors commonly perceived to be responsible for a usable system.

5. Limitations of the Study

As with any research, this study has its limitations. They include the sample of the user population chosen to participate, the methodology applied to the research questions, the context chosen for the investigation, and the instruments used to measure users' satisfaction. Each of these limitations is discussed in more detail in the following paragraphs.

5.1. The Sample

Generalizing the results of this study must be done with caution due to the sample of the population that was considered. Participants in this study represented but a small segment of the larger population who use web-based enterprise portals and an even smaller segment of those who use web-based applications in general. The sample's homogenous nature with regards to their age and their level of education and computer experience begs the question whether this study's findings would hold true for others. It is possible that both novice and expert computer users would rationalize their satisfaction

differently than the *intermediate* level participants in this study. Additionally, the relatively small number of subjects per test condition (N = 15) resulted in less statistical power than might normally be preferred for an experimental comparison study.

5.2. The Context

The context chosen for this study represents a second limiting factor for readers to consider. As with any type of research in which the subjects are aware that they are being studied, there is potential for the Hawthorne effect (Gall et al, 1996). In this study, there was the related phenomenon of users' tendencies to persist at tasks during a usability test. Together, these variables may have influenced any or all of users' effectiveness, efficiency, and satisfaction with the system.

The selection of a web-based portal application to examine the questions in this study also limits the extent to which the results of this study can be generalized. Given that the majority of the participants had no prior experience with an enterprise portal application, there may have been some novelty effect on their ultimate satisfaction with the system that might not have been there had the software been something they were more familiar with (e.g. word processing, Email, regular website, etc.).

5.3. The PSSUQ and ASQ Satisfaction Instruments

A number of satisfaction instruments are available to usability practitioners. Of these, this study implemented the Post-Study System Usability Questionnaire (PSSUQ) and After-Scenario Questionnaire (ASQ) developed by IBM Corporation. Although less popular than some other instruments, the PSSUQ and ASQ had been shown in previous research to be reliable. The brevity of each instrument, particularly the 3-item ASQ,

added to their attraction for this study. Despite these positive characteristics, however, a few limitations of the instruments were observed.

Subjects in this study frequently requested clarification for a few of the items on the PSSUQ as well as the third item on the ASQ. By having subjects think-aloud during their rating activities, these uncertainties were recognized and users were helped to understand the items so as to rate them appropriately. Nevertheless, it is unclear to the researcher that his efforts to clarify certain items were always successful. It remains possible that some uncertainties about items on the PSSUQ persisted and may have influenced users' ratings on occasion.

A second issue with the two instruments was the tendency for users to reverse the scale without realizing it. The original Likert Scales for the PSSUQ and ASQ instruments contained seven intervals with anchors of 1 = strongly agree (satisfied) and 7 = strongly disagree (not satisfied). A number of subjects began the rating activity by selecting a larger number on the scale, thereby indicating disagreement with that item and dissatisfaction with the system, when in fact they were truly quite satisfied. The think-aloud protocol allowed the researcher to identify such inconsistencies and confirm with users that their ratings matched their intentions. However, there is the possibility that the counterintuitive scale may have led to the occasional user slip that went unnoticed by both the subject and the researcher. For this reason, it is recommended that future users of the PSSUQ and ASQ instruments reverse the scales in advance to provide anchors of 1 = strongly disagree (not satisfied) and 7 = strongly agree (satisfied).

A final limitation associated with the satisfaction instruments is that neither the PSSUQ nor ASQ provided users with labels other than the two anchor points. The lack of

additional labels made it difficult at times for users to determine how to interpret the various intervals and may have led to less than accurate ratings in some cases.

6. Suggestions for Further Research

While the results of this study suggest that priming has no discernible impact on users' post-test satisfaction ratings in a usability test, certain limitations identified earlier in this chapter may warrant some further exploration of this issue. Future studies might consider investigating the use of priming with systems other than web-based portals, particularly systems that users have some familiarity with in order to remove any potential novelty effect associated with using a new system. Similarly, one might explore whether priming effects are more noticeable when users have more or less computer experience than those who participated in this study.

Perhaps the more fruitful line of research to come out of this study has to do with the findings from the second research question. Future research might investigate whether usability testing other systems and other user categories produces the same rationales for users' satisfaction as those that were identified in this study. Particularly for those rationales that were identified as directional in this study, additional research is needed to determine how best to increase the percentage of positive reactions and decrease the percentage of negative reactions for themes such as *Clear and helpful instructions, The paradox of information quantity*, and *Time equals satisfaction*. Through such efforts, the rationales identified in this study might be further validated and refined to assist usability researchers and software designers in their efforts at creating both usable and satisfying system.

Additional testing of other enterprise portals using the PSSUQ instrument might be conducted to determine if the correlations found in this study between efficiency, effectiveness, and satisfaction exist for other web-based enterprise portal systems. Along those same lines, further testing of the OneStart portal system using other satisfaction questionnaires (e.g. SUMI, QUIS, etc.) may help assess the reliability of the strong correlations discovered in this study.

Finally, additional research geared toward web-portal development might examine how users' satisfaction ratings with a new portal system relate to their future adoption and use of that system. Efforts in this area would do well to consider the lessons learned from this study, making sure that they include a substantial qualitative component in order to help designers and evaluators understand users' decision-making criteria.

REFERENCES

Andre, A. D., & Wickens, C. D. (1995). When users want what's not best for them. *Ergonomics in Design, October*, 37-41.

Baecker, R., Grudin, J., Buxton, W., & Greenberg, S. (1995). A historical and intellectual perspective. In R. Baecker & J. Grudin & W. Buxton & S. Greenberg (Eds.), *Readings in human-computer interaction: Toward the year 2000* (pp. 35-47). New York: Morgan Kaufmann.

Bailey, G. (1995). *Performance vs. preference.* Paper presented at the Conference of the Human Factors and Ergonomics Society, New York, NY.

Bailey, G. D. (1993). *Iterative methodology and designer training in human-computer interface design.* Paper presented at the ACM INTERCHI'93 Conference on Human Factors in Computing Systems, New York.

Bainbridge, L., & Sanderson, P. (1995). Verbal protocol analysis. In J. Wilson & N. Corlett (Eds.), *Evaluation of Human Work* (2 ed.): Taylor and Francis.

Baroudi, J. J., Olson, M. H., & Ives, B. (1986). An empirical study of the impact of user involvement on system usage and information satisfaction. *Communications of the ACM, 29*(3), 232-238.

Berners-Lee, T. (1998). *The World Wide Web: A very short personal history* [World Wide Web]. Tim Berners-Lee. Retrieved November 16, 2001, from the World Wide Web: http://www.w3.org/People/Berners-Lee/ShortHistory.html

Bevan, N. (1995). Measuring usability as quality of use. *Software Quality Journal, 4*, 115-130.

Bevan, N., & Macleod, M. (1994). Usability measurement in context. *Behaviour & Information Technology, 13*(1-2), 132-145.

Beyer, J., & Holtzblatt, K. (1998). *Contextual design: Defining customer-centered systems*. San Francisco, CA: Morgan Kaufmann.

Bias, R. G. (1994). The pluralistic usability walkthrough: Coordinated empathies. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 63-76). New York: John Wiley & Sons, Inc.

Bias, R. G., & Mayhew, D. J. (Eds.). (1994). *Cost-justifying usability*. Toronto: Academic Press.

Billingsley, P. A. (1993). Reflections on ISO 9241: Software usability may be more than the sum of its parts. *StandardView, 1*(1), 22-25.

Borrie, W. T., & Roggenbuck, J. W. (1998). The problem of verbal reports in recreation research: Review, recommendations, and new directions. *Tourism Analysis, 2*, 175-183.

Bouch, A., Kuchinsky, A., & Bhatti, N. (2000, April 1-6). *Quality is in the eye of the beholder: Meeting users' requirements for internet quality of service.* Paper presented at the CHI 2000: Conference on Human Factors in Computing Systems, The Hague, Netherlands.

Carey, J. M., Mizzi, P. J., & Lindstrom, L. C. (1996). Pull-down versus traditional menu types: An empirical comparison. *Behaviour & Information Technology, 15*(2), 84-95.

Carroll, J. M. (1990). *The Nurnberg funnel: Designing minimalist instruction for practical computer skill*. London, England: The MIT Press.

Chen, H., & Nilan, M. (1998). *An exploration of web users' internal experience: Application of the experience sampling method to the web environment.* Paper presented at the WebNet 98 World Conference, Orlando, FL.

Chin, J., Diehl, V., & Norman, K. (1988). *Development of an instrument measuring user satisfaction of the human-computer interface.* Paper presented at the ACM CHI 1988 Conference on Human Factors in Computing Systems.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Connolly, C. G. (2000). From static web site to portal. *Educause Quarterly, 23*(2), 38-43.

Cooper, A. (1999). *The inmates are running the asylum*. Indianapolis, IN: SAMS.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper Perennial.

Csikszentmihalyi, M., & Csikszentmihalyi, I. (1988). Introduction to Part IV. In M. Csikszentmihalyi & I. Csikszentmihalyi (Eds.), *Optimal experience: Psychological studies of flow in consciousness* (pp. 251-265). Cambridge, UK: Cambridge University Press.

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the experience-sampling method. *The Journal of Nervous and Mental Disease, 175*(9), 526-536.

Datz-Kauffold, B., & Henry, S. L. (2000, August 14-18). *Waving magic wands: Interaction techniques to improve usability testing low-fidelity prototypes.* Paper presented at the Usability Professionals' Association, Asheville, NC.

Doll, W. J., & Torkzadeh, G. (1988). The measurement of end-user computing satisfaction. *MIS Quarterly, 12*(2), 259-274.

Dumas, J. S., & Redish, J. C. (1993). *A practical guide to usability testing*. Norwood, NJ: Ablex.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Revised ed.). Cambridge, MA: MIT Press.

Ethridge, R. R., Hadden, C. M., & Smith, M. P. (2000). Building a personalized education portal. *Educause Quarterly, 23*(3), 12-19.

Ettema, J. S. (1985). Explaining information system use with system-monitored vs. self-reported use measure. *Public Opinion Quarterly, 49*, 381-387.

Fontana, a., & Frey, J. (2000). The interview: From structured questions to negotiated text. In N. K. Denzin & Y. S. Lincoln (Eds.), *Handbook of qualitative research* (pp. 645-672). Newbury Park, CA: Sage.

Foster, J. C., Mcinnes, F. R., Jack, M. A., Love, S., Dutton, R. T., Nairn, I. A., & White, L. S. (1998). An experimental evaluation of preferences for data entry method in automated telephone services. *Behaviour & Information Technology, 17*(2), 82-92.

Frazee, J. P. (2001). Charting a smooth course for portal development. *Educause Quarterly, 24*(3), 42-48.

Frick, T. W., & Semmel, M. I. (1978). Observer agreement and reliabilities of classroom observational measures. *Review of Eduacational Research, 48*(1), 157-184.

Frøkjær, E., Hertzum, M., & Hornbaek, K. (2000, April 1-6). *Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated?* Paper presented at the ACM CHI 2000 Conference on Human Factors in Computing Systems, The Hague, Netherlands.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). White Plains, NY: Longman.

Gould, J. D., Boies, S. J., Levy, S., Richards, J. T., & Schoonard, J. (1987). The 1984 Olympic Message System: A test of behavioral principles of system design. *Communications of the ACM, 30*(9), 758-769.

Gould, J. D., Boies, S. J., & Lewis, C. (1991). Making usable, useful, productivity: Enhancing computer applications. *Communications of the ACM, 34*(1), 74-85.

Gould, J. D., & Lewis, C. (1985). Designing for usability: Key principles and what designers think. *Communications of the ACM, 28*(3), 300-311.

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction, 13*, 203-261.

Gutwin, C., & Greenberg, S. (1998). *Effects of awareness support on groupware usability.* Paper presented at the ACM CHI 1998 Conference on Human Factors in Computing Systems, Los Angeles, CA.

Hackos, J. T., & Redish, J. C. (1998). *User and task analysis for interface design*. Toronto: John Wiley & Sons, Inc.

Hartson, R., Castillo, J. C., & Kelso, J. (1996, April 13-18). *Remote evaluation: The network as an extension of the usability laboratory.* Paper presented at the ACM CHI'96 Conference on Human Factors in Computing Systems, Vancouver, BC, Canada.

Head, T. C. (1991). An examination of the priming effect as it pertains to the leader behavior description questionnaire form XII. *Psychological Reports, 68*(2), 515-520.

Hertzum, M., & Frøkjær, E. (1996). Browsing and querying in online documentation: A study of user interfaces and the interaction process. *ACM Transactions on Computer-Human Interaction, 3*(2), 136-161.

Ives, B., Olson, M. H., & Baroudi, J. J. (1983). The measurement of user information satisfaction. *Communcations of the ACM, 26*(10), 785-793.

Jacobson, C. (2000). Institutional Information Portals. *Educause Review, 35*(4), 58-59.

Jacques, R., Preece, J., & Carey, T. (1995). Engagement as a design concept for multimedia. *Canadian Journal of Educational Communication, 24*(1), 49-59.

Jafari, A. (2000). Optimizing campus web sites. *Educause Quarterly, 23*(2), 56-58.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. M. (1991). *User interface evaluation in the real world: A comparison of four techniques.* Paper presented at the ACM CHI'91 Conference on Human Factors in Computing Systems, New Orleans, LA.

Kirakowski, J., & Corbett, M. (1993). SUMI: The software usability measurement inventory. *British Journal of Educational Technology, 24*(3), 210-212.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3nd ed). Monterey, CA:Brooks/Cole.

Kissel, G. V. (1995). *The effect of computer experience on subjectiev and objective software usablity measures.* Paper presented at the ACM CHI 1995 Conference on Human Factors in Computing Systems, Denver, CO.

Kraemer, H. C. (1982). Kappa coefficient. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of Statistical Sciences*. New York: John Wiley & Sons.

Kvavik, R. B., & Handberg, M. N. (2000). Transforming student services. *Educause Quarterly, 23*(2), 30-37.

Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for coputer usability studies: The ASQ. *SIGCHI Bulletin, 23*(1), 78-81.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Pyschometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7*(1), 57-78.

Lin, H. X., Choong, Y., & Salvendy, G. (1997). A proposed index of usability: A method for comparing the relative usability of different software systems. *Behaviour & Information Technology, 16*(4/5), 267-278.

Looney, M., & Lyman, P. (2000). Portals in higher education. *Educause Review, 35*(4), 28-36.

Mandel, T. (1997). *The elements of user interface design*. New York,NY: John Wiley & Sons.

Maslow, A. H. (1973). Theory of human motivation. In R. J. Lowry (Ed.), *Dominance, self-esteem, self-actualization: Germinal papers of A.H. Maslow*. Monterey, CA: Brooks-Cole.

Mayhew, D. J., & Mantei, M. (1994). A basic framework for cost-justifying usability engineering. In R. G. Bias & D. J. Mayhew (Eds.), *Cost-justifying usability* (pp. 9-43). Boston: Academic Press.

McKnight, C., Dillon, A., & Richardson, J. (1996). User-centered design of hypertext/hypermedia for education. In D. H. Jonassen (Ed.), *Handbook of Research for Educational Communications and Technology*. New York: Macmillan.

Melone, N. P. (1990). A theoretical assessment of the user-satisfaction construct in information systems research. *Management Science, 36*(1), 76-91.

Moss, S. E., & Lawrence, K. G. (1997). The effects of priming on the self-reporting of perceived stressors and strains. *Journal of Organizational Behavior, 18*(4), 393-403.

Mullet, K., & Sano, D. (1995). *Designing visual interfaces: Communication oriented techniques*. Englewood Cliffs, NJ: SunSoft Press.

Nielsen, J. (1989). What do users really want? *International Journal of Human-Computer Interaction, 1*(1), 137-147.

Nielsen, J. (1993). *Usability engineering*. Cambridge, MA: Academic Press.

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R. L. Mack (Eds.), *Usability Inspection Methods*. New York: John Wiley & Sons, Inc.

Nielsen, J. (2000). *Designing web usability: The practice of simplicity*. Indianapolis, IN: New Riders Publishing.

Nielsen, J. (2001). *Usability metrics* [Web article]. Jakob Nielsen. Retrieved March 4, 2001, from the World Wide Web: http://www.useit.com/alertbox/20010121.html

Nielsen, J., & Levy, J. (1994). Measuring usability: Preference vs. performance. *Communications of the ACM, 37*(4), 66-75.

Nielsen, J., & Phillips, V. L. (1993). *Estimating the relative usability of two interfaces: Heuristic, formal, and empirical methods compared.* Paper presented at the ACM INTERCHI'93, New York.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

Norman, D. A. (1988). *The design of everyday things*. New York, NY: Doubleday Currency.

Norman, D. A. (1998). *The invisible computer*. Cambridge, MA: MIT Press.

Norman, D. A., & Draper, S. W. (1986). Introduction. In D. A. Norman & S. W. Draper (Eds.), *User centered system design: New perspectives on human-computer interaction* (pp. 1-6). Hillsdale, NJ: Lawrence Erlbaum.

Norman, K., & Shneiderman, B. (1989). Questionnaire for user interaction satisfaction (QUIS 5.0). HCI Lab, College Park, University of Maryland.

Nunnally, J. C. (1970). *Introduction to Psychological Measurement.* New York: McGraw-Hill.

Omoigui, N., He, L., Gupta, A., Grudin, J., & Sanocki, E. (1999). *Time-compression: Systems concerns, usage and benefits.* Paper presented at the ACM CHI 2000 Conference on Human Factors in Computing Systems, Pittsburgh, PA.

Oppenheim, A. N. (1992). *Questionnaire design, interviewing and attitude measurement* (2 ed.). London: Pinter.

Phifer, G. (2000). *Best practices in deploying enterprise portals*. Gartner Group, Inc. Retrieved November 4, 2001, from the World Wide Web: http://gartner3.gartnerweb.com/public/static/hotc/hc00091248.html

Preece, J. (Ed.). (1993). *A guide to usability: Human factors in computing*. Wokingham: Addison-Wesley.

Raghavan, S., & Perlman, G. (2000). *The difference of preference vs. performance can be different for concurrent vs. retrospective ratings.* Paper presented at the IEA 2000/HFES 2000 Conference.

Rand, P. (1993). *Design, Form, and Chaos*. New Haven, CT: Yale University Press.

Raskin, J. (1994). Intuitive equals familiar. *Communications of the ACM, 37*(9), 17-18.

Reeves, B., & Nass, C. (1996). *The media equation*. Cambridge, UK: Cambridge University Press.

Rivard, S., & Huff, S. L. (1988). Factors of success for end-user computing. *Communications of the ACM, 31*(5), 552-561.

Rogers, E. (1995). *Diffusion of innovations* (4th ed ed.). New York: Free Press.

Root, R. W., & Draper, S. (1983). *Questionnaires as a software evaluation tool.* Paper presented at the ACM CHI 1983 Conference on Human Factors in Computing Systems.

Rosenbaum, S., Rohn, J. A., & Humburg, J. (2000). *A toolkit for strategic usability: Results from workshops, panels, and surveys.* Paper presented at the ACM CHI 2000 Conference on Human Factors in Computing Systems, The Hague, Netherlands.

Rowley, D. E., & Rhoades, D. G. (1992). *The cognitive jogthrough: A fast-paced user interface evaluation procedure.* Paper presented at the ACM CHI'92 Conference on Human Factors in Computing Systems, Monterey, CA.

Rubin, J. (1994). *Handbook of usability testing: How to plan, design, and conduct effective tests*. New York: John Wiley & Sons, Inc.

Rushinek, A., & Rushinek, S. F. (1986). What makes users happy? *Communications of the ACM, 29*(7), 594-598.

Salancik, G. R., & Conway, M. (1975). Attitude inferences from salient and relevant cognitive content about behavior. *Journal of Personality and Social Psychology, 32*(5), 829-840.

Salancik, G. R., & Pfeffer, J. (1978). A social information procesing approcch to job attitudes and task design. *Administrative Science Quarterly, 23*, 224-253.

Schuman, H. (1970). The random probe: A technique for evaluating the validity of closed questions. In D. P. Forcese & S. Rocher (Eds.), *Stage of social research*. Englewood Cliffs, NJ: Prentice-Hall.

Sears, A. (1997). Heuristic walkthroughs: Finding problems without the noise. *International Journal of Human-Computer Interaction, 9*(3), 213-234.

Shackel, B. (1986, September 23-26). *People and computers: Designing for usability.* Paper presented at the Conference of the British Computer Society: Human Computer Interaction Specialist Group, University of York, UK.

Shackel, B. (1991). Usability - context, framework, definition, design and evaluation. In B. Shackel & S. Richardson (Eds.), *Human Factors for Informatics Usability* (pp. 21-38). Cambridge: Cambridge University Press.

Shackel, B. (1997). Human-computer interaction – whence and whither? *Journal of the American Society for Information Science, 48*(11), 970-986.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer interaction* (3rd ed.). Reading, MA: Addison-Wesley Longman.

Steinbrenner, K. (2001). Unlocking ERPs with portals. *Educause Quarterly, 24*(3), 55-57.

Sugar, W. A., & Boling, E. (1995). *User-centered innovation: A model for early usability testing* (ERIC Document Reproduction Service No. ED 383 340).

Tanin, E., Lotem, A., Haddadin, I., Shneiderman, B., Plaisant, C., & Slaughter, L. (2000). Facilitating data exploration with query previews: A study of user performance and preference. *Behaviour and Information Technology, 19*(6), 393-403.

Teague, R., De Jesus, K., & Nunes-Ueno, M. (2001). *Concurrent vs. post-task usability test ratings.* Paper presented at the ACM CHI 2001 Conference on Human Factors in Computing Systems, Seattle, WA.

Torkzadeh, G., & Doll, W. J. (1991). Test-restest reliability of the end-user computing satisfaction instrument. *Decision Sciences, 22*(1), 26-37.

Tufte, E. R. (1990). *Envisioning Information*. Cheshire: Graphics Press.

Tulving, E., & Schacter, D. L. (1990). Priming and human memory system. *Science, 247*, 301-306.

Uebersax, J. S. (1987). Decision-making models and the measurement of interrater agreement. *Psychological Bulletin, 101*, 140-146.

van Welie, M., van der Veer, G. C., & Eliëns, A. (1999). *Breaking down usability.* Paper presented at the INTERACT '99, Edinburgh.

Walker, M. A., Fromer, J., Di Fabbrizio, G., Mestel, C., & Hindle, D. (1998). *What can I say? Evaluating a spoken interface to email.* Paper presented at the ACM CHI 1998 Conference on Human Factors in Computing Systems, Los Angeles.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In J. Neilsen & R. L. Mack (Eds.), *Usability Inspection Methods* (pp. 105-140). New York: John Wiley & Sons, Inc.

Wiklund, M. E. (Ed.). (1994). *Usability in practice: How companies develop user-friendly products*. Toronto: Academic Press.

Wildemuth, B. M., Friedman, C. P., & Downs, S. M. (1998). Hypertext versus boolean access to biomedical information: A comparison of effectiveness, efficiency, and user preferences. *ACM Transactions on Computer-Human Interaction, 5*(2), 156-183.

Yi, Y. (1993). Contextual priming effects in print advertisements: The moderating role of prior knowledge. *Journal of Advertising, 22*(1), 1-10.

APPENDICES

Appendix A. List of Tasks for Groups A, B and C.

Task A

Log into OneStart with your IU UserID and password.

Task B

Use the OneStart Tutorial to answer each of the following questions:

- Explain the difference between a page and a channel.
- Explain the difference between portal icons and channel icons.
- Describe the two different methods for selecting content for the portal?

Close the popup Tutorial window before proceeding with the next task.

Task C

Add a new page to your portal and name it “My stuff”. Add the 4 channels below and change the screen’s colors to red and gold as seen below.

Appendix A (cont'd). List of Tasks for Groups A, B and C.

Task D
Add a channel to the wide column (right side) of the OneStart page named "IUB" that lets you view the "Indiana Digital Student" online newspaper.

Task E
You would like to print out a copy of your course schedule for this semester.
Use the OneStart portal to add the necessary channel to your "IUB" page and use it to print out your schedule.

Task F
Add a page and name it "Computer Info". Add the 3 channels as below.
Use the "Accessible Option" on the Content & Layout screen for adding your content this time.

Task G
You would like to be able to view movie show-time details in your portal. Currently, no such option is available in the list of OneStart channels.

Use "My Custom Channels" on your portal's "Front Page" to create a channel that will contain the Kerasotes website (http://www.kerasotes.com/Showtimes.asp).
Once created, make it so you can view this new channel on your "MyStuff" page.

Use the new channel to identify two movies playing at the College Mall theatres in Bloomington.

Appendix A (cont'd). List of Tasks for Groups A, B and C.

Task H

Use the Bookmarks channel on the "Front Page" to create a bookmark for "Amazon.com" (www.amazon.com). Store this bookmark in a folder called "Online Stores". Your channel should like the diagram below when you are done.

Once you have completed this, check to see that the Amazon.com bookmark works properly.

Task I

Using the Bookmarks channel on your "Front Page", rename the "Amazon.com" bookmark to simply "Amazon", and transfer it to a new folder called "Books". Your channel should like the diagram below when you are done.

Appendix B. Post-Study System Usability Questionnaire (PSSUQ)

(Lewis, 1995)

1) Overall, I am satisfied with how easy it is to use this system.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

2) It was simple to use this system.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

3) I could effectively complete the tasks and scenarios using this system.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

4) I was able to complete the tasks and scenarios quickly using this system.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

5) I was able to efficiently complete the tasks and scenarios using this system.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

6) I felt comfortable using this system.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

7) It was easy to learn to use this system.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

8) I believe I could become productive quickly using this system.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

9) The system gave error messages that clearly told me how to fix problems.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

10) Whenever I made a mistake using the system, I could recover easily and quickly.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

11) The information (such as on-line help, on-screen messages, etc.) provided with this system was clear.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

12) It was easy to find the information I needed.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

13) The information provided for the system was easy to understand.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

14) The information was effective in helping me complete the tasks and scenarios.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

15) The organization of information on the systems screens was clear.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

NOTE: the interface includes those items that you use to interact with the system. For example, some components of the interface are the keyboard, the mouse, and the screens (including their use of graphics and language).

16) The interface of this system was pleasant.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

17) I like using the interface of this system.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

18) This system has all the functions and capabilities I expect it to have.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

19) Overall, I am satisfied with this system.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

Appendix C. After-Scenario Questionnaire (ASQ)

(Lewis, 1991)

1) Overall, I am satisfied with the ease of completing the tasks in this scenario.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

2) Overall, I am satisfied with the time it took to complete the tasks in this scenario.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

3) Overall, I am satisfied with the support information (instructions, online help) when completing the tasks.

strongly agree <----------> strongly disagree not applicable

1 2 3 4 5 6 7 N/A

Comments:

Appendix D. Demographic Questionnaire

Name: _____

Email: _____ @ indiana.edu

Gender: ❑ M ❑ F Age: _____

Year in University (e.g. 1,2,3,4…): _____

Major area of study (e.g. Math Education): _____

What activities do you use the Web for: (Check all that apply)

- ❑ News
- ❑ Sports & entertainment
- ❑ Email/chat
- ❑ School/library research
- ❑ General searching (e.g. Google)
- ❑ Online purchases
- ❑ Banking
- ❑ Personal portal (e.g. MyYahoo)
- ❑ Develop websites or applications
- ❑ Other _____
- ❑ Other _____

List ALL of the computer applications that you are skilled in using:
(e.g. Microsoft Word, Telnet (Pine), Photoshop, Front Page, Dreamweaver, etc…)

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

_____ _____

Appendix E. Introductory Protocol for Subjects in Group A

For the session today, we will be looking at the OneStart portal application for Indiana University. This system has recently been released for students, staff and faculty to use, but we are always looking for ways to improve it. That is why we are testing it with students like yourself.

The main purpose of this session is to determine how satisfied you are with the system. At the end of the session, I will be asking you to complete a questionnaire to rate your overall satisfaction level with the system.

Keep in mind that we are testing the application and NOT you. If there are parts of the application that are difficult for you to use, it is ok to say so. Any problems you may have help us come up with ways to make the program better.

Once I have gone over the instructions with you and you are comfortable with them, we will do the first task together so you know how things are going to work. After that task, I will go into the next room and leave you to work alone in this room. I will be able to see your computer screen though and hear you through the speakers and you will be able to hear me through the computers speakers if I need to give you instructions.

You will perform a series of about 12 tasks during the session. Each task is written on a sheet of paper that you may refer to as you perform the task. I'd like you to read each task to yourself and then tell me in your own words what it asks you to do. Some tasks are short and some are rather long. Work on a task for as long as you normally would. When you feel you have completed it, say "I'm done" and you can proceed with the next task. If you ever feel like you have spent long enough on a task and want to move on, just say "I'd like to pass" and you can begin the next task.

To help me understand how you use the application, I would like you to 'think aloud' as you complete the tasks. This helps me understand what you are looking at on the screen, what you are looking for, and why you make the choices you do. If you forget to think out loud, I may remind you by asking you to tell me what you are thinking.

After you have finished the task portion of the session, I will have you complete a short questionnaire rating your satisfaction with the system. Then I'd like to ask just a few follow-up questions. After our discussion, I can answer any questions you may have about the study or the OneStart portal and I will get you to sign a couple more forms for your phone card and IUWare CD-ROM. Do you have any questions for me now?

Let's begin with the practice task.

Appendix F. Introductory Protocol for Subjects in Group B

For the session today, we will be looking at the OneStart portal application for Indiana University. This system has recently been released for students, staff and faculty to use, but we are always looking for ways to improve it. That is why we are testing it with students like yourself.

The main purpose of this session is to determine how satisfied you are with the system. At the end of the session, I will be asking you to complete a questionnaire to rate your overall satisfaction level with the system.

Keep in mind that we are testing the application and NOT you. If there are parts of the application that are difficult for you to use, it is ok to say so. Any problems you may have help us come up with ways to make the program better.

Once I have gone over the instructions with you and you are comfortable with them, we will do the first task together so you know how things are going to work. After that task, I will go into the next room and leave you to work alone in this room. I will be able to see your computer screen though and hear you through the speakers and you will be able to hear me through the computers speakers if I need to give you instructions.

You will perform a series of about 12 tasks during the session. Each task is written on a sheet of paper that you may refer to as you perform the task. I'd like you to read each task to yourself and then tell me in your own words what it asks you to do. Some tasks are short and some are rather long. Work on a task for as long as you normally would. When you feel you have completed it, say "I'm done" and you can proceed with the next task. If you ever feel like you have spent long enough on a task and want to move on, just say "I'd like to pass" and you can begin the next task.

To help me understand how you use the application, I would like you to 'think aloud' as you complete the tasks. This helps me understand what you are looking at on the screen, what you are looking for, and why you make the choices you do. If you forget to think out loud, I may remind you by asking you to tell me what you are thinking.

After you have finished the task portion of the session, I will have you complete a short questionnaire rating your satisfaction with the system. Then I'd like to ask just a few follow-up questions. After our discussion, I can answer any questions you may have about the study or the OneStart portal and I will get you to sign a couple more forms for your phone card and IUWare CD-ROM. Do you have any questions for me now?

Before we begin with the session, I would like you to become familiar with the satisfaction questionnaire that you will be filling out at the end of the session. I'd like you to fill out a sample questionnaire right now for one of the systems you listed on the form earlier. Take the next few minutes to think of that system and fill out this questionnaire with that system in mind.

Let's begin with the practice task.

Appendix G. Introductory Protocol for Subjects in Group C

For the session today, we will be looking at the OneStart portal application for Indiana University. This system has recently been released for students, staff and faculty to use, but we are always looking for ways to improve it. That is why we are testing it with students like yourself.

The main purpose of this session is to determine how satisfied you are with the system. At the end of the session, I will be asking you to complete a questionnaire to rate your overall satisfaction level with the system.

Keep in mind that we are testing the application and NOT you. If there are parts of the application that are difficult for you to use, it is ok to say so. Any problems you may have help us come up with ways to make the program better.

Once I have gone over the instructions with you and you are comfortable with them, we will do the first task together so you know how things are going to work. After that task, I will go into the next room and leave you to work alone in this room. I will be able to see your computer screen though and hear you through the speakers and you will be able to hear me through the computers speakers if I need to give you instructions.

You will perform a series of about 12 tasks during the session. Each task is written on a sheet of paper that you may refer to as you perform the task. I'd like you to read each task to yourself and then tell me in your own words what it asks you to do. Some tasks are short and some are rather long. Work on a task for as long as you normally would. When you feel you have completed it, say "I'm done" and you can proceed with the next task. If you ever feel like you have spent long enough on a task and want to move on, just say "I'd like to pass" and you can begin the next task.

After each task, I'd like you to complete 3 short questions rating your satisfaction with that task. For each task, I'd like you to indicate your rating on a log sheet provided for you. After giving your rating, you may proceed with the next task sheet.

To help me understand how you use the application, I would like you to 'think aloud' as you complete the tasks. This helps me understand what you are looking at on the screen, what you are looking for, and why you make the choices you do. If you forget to think out loud, I may remind you by asking you to tell me what you are thinking.

After you have finished the task portion of the session, I will have you complete a short questionnaire rating your satisfaction with the system. Then I'd like to ask just a few follow-up questions. After our discussion, I can answer any questions you may have about the study or the OneStart portal and I will get you to sign a couple more forms for your phone card and IUWare CD-ROM. Do you have any questions for me now?

Let's begin with the practice task.

Todd W. Zazelenchuk

Education

Ph.D. March, 2002. Instructional Systems Technology, Indiana University, Bloomington

Dissertation: Measuring satisfaction in usability tests: A comparison of questionnaire administration methods and an investigation into users' rationales for satisfaction.

M.Ed. August, 1994. Educational Technology, University of Saskatchewan, Saskatoon

B.Ed. April, 1989. University of Saskatchewan, Saskatoon

B.Sc. April, 1986. University of Saskatchewan, Saskatoon

Work Experience

Research, Design and Evaluation

1999-2002. Usability Consulting Services, University Information Technology Services, Indiana University.

2001. Contract usability consulting: Palm Inc., Santa Clara, CA., Centers for Disease Control, Atlanta, GA.

1999. Contract human performance consulting: Cornerstone Information Systems, Bloomington, IN.

1999. Contract instructional systems consulting: Conexant Systems, Newport Beach, CA.

1996-1998. Extension Division, University of Saskatchewan, Saskatoon, SK, Canada.

1994-1996. Andersen Consulting, Calgary, AB, Canada.

Teaching

1999-1999. Lecturer, Teacher Education Preparation, Indiana University, Bloomington, IN.

1998. Lecturer, Instructional Design, University of Saskatchewan, Saskatoon, SK, Canada.

1994. Lecturer, Instructional Technology, University of Saskatchewan, Saskatoon, SK, Canada.

1989-1993. Teacher, Grades 6-7, Westberry School, Kindersley, SK, Canada.