INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

**Remote Evaluation: A Comparison of Attended and Unattended
Usability Testing Via the Internet**

J. Steven T. Macko, B.A. (Hon)

**A thesis submitted to
the faculty of Graduate Studies and Research
in partial fulfillment of
the requirements for the degree of
Master of Arts**

**Department of Psychology
Carleton University
Ottawa, Ontario
May, 1998**

© 1998, J. Steven T. Macko

National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-32368-4

The undersigned recommend to the faculty of Graduate Studies
and Research acceptance of the thesis

Remote Evaluation: A Comparison of Attended and Unattended
Usability Testing Via the Internet

submitted by J. Steven T. Macko, B.A. (Hon)
in partial fulfillment of the requirements for
the degree of Master of Arts

A handwritten signature in cursive script, reading "Richard F. Dillon", written over a horizontal line.

Richard F. Dillon, Ph. D.

Thesis Supervisor

A handwritten signature in cursive script, reading "Kim Matheson", written over a horizontal line.

Kim Matheson, Ph. D.

Chair, Department of Psychology

Carleton University

1998

Abstract

Attended and unattended usability testing via the Internet were compared. The goal was to determine how attended Internet usability testing data would differ from unattended Internet usability testing data. Sixty subjects were randomly assigned to one of two experimental conditions and compared on 13 variables derived from 18 survey questions. Subjects in both conditions were identical on three experience variables. Performance was measured in terms of forced-choice questions and open-ended questions from the survey. Unattended testing took significantly longer to administer. Attended and unattended usability testing were equivalent on forced-choice survey questions. Unattended testing was an acceptable alternative to traditional attended usability testing regarding forced-choice questions. Conversely, attended and unattended usability testing were dissimilar on open-ended survey questions. Attended usability testing would be the preferred approach if the focus of the usability test was comment analysis because attended testing yielded more comments, unique ideas per group, unique comments per subject, and positive and negative comments, with minimal irrelevant and process comments.

Acknowledgments

I would like to express my sincere thanks and appreciation to several individuals who helped me in the process of completing this study. I would like to thank Dr. R. F. Dillon, faculty advisor, who provided guidance, comprehensive and endless feedback, and insight, and Dr. J. Wood, co-advisor, who conferred with me on usability concerns and statistical procedures. Special thanks to Dr. A. Patrick, external advisor, for sharing his expertise on survey design/analysis and providing a “third set of eyes.” I would like to thank Scott McEwen, who made it possible for me to use Cognos as a research setting, and Roger Chang, who helped me write and debug my survey’s JavaScript and HTML code. Special thanks to Corey Bernardo, Richard Hong, Tina Groves, Lesley Grignon, and Fiona Bremner for their assistance at various stages throughout the process, and to all my volunteer subjects, who invested their time to complete my survey. I would like to thank my parents, for all the Sunday dinners, concern, and support. Finally, special thanks to my fiancée Caroline, for patiently standing by throughout my many years of education, supporting my decisions.

Table of Contents

USABILITY TESTING	2
<i>Comparison of Judgment Questions</i>	3
<i>Rating Scales</i>	3
REMOTE EVALUATION	4
<i>Potential Advantages of Remote Evaluation</i>	4
Opportunity to obtain "real-world" data	4
Cost effectiveness	5
Usability engineer can not adversely influence the participants.	5
User expectations.	5
<i>Potential Disadvantages of Remote Evaluation</i>	6
Absence of direct observation	6
Annoyance	6
Absence of human assistance.	6
Honesty	6
<i>Interpersonal Roles of the Usability Engineer</i>	7
Welcome the participant.	7
Observe behaviors.	8
Probe.	8
Clarify	8
Flexibility	8
Encourage	8
Debrief the participant.	9
TYPES OF REMOTE EVALUATION	9
<i>Remote Questionnaire/Survey</i>	10
<i>Web Prototyping</i>	12
Potential advantages of Web prototyping	14
Potential disadvantages of Web prototyping.	15
CURRENT STUDY	16
METHOD	19
PARTICIPANTS	19
APPARATUS	19
PROCEDURE	28
RESULTS	32
ASSUMPTIONS	32
EXPERIENCE VARIABLES	32
TIME TO COMPLETE SURVEY	33
SURVEY ANALYSIS	33
<i>Forced-Choice Questions</i>	33
Display preference.	33
Printing preference	34
Rank the three Explanation styles	34
Correct/incorrect responses	35
Definition of outliers.	35
Rating scales.	36
<i>Open-Ended Question Analysis</i>	38
Mean number of words per question	39
Mean number of comments per question.	39
Mean number of words per comment per question	39
Mean number of unique comments per subject	39
Mean number of positive comments per question.	41
Mean number of negative comments per question.	41
Mean number of irrelevant comments per question	41

Mean number of process comments per question.....	42
FREQUENCY OF INTERPERSONAL BEHAVIORS	42
SUMMARY OF RESULTS.....	43
DISCUSSION.....	45
HYPOTHESIS 1: EQUIVALENT ANSWERS	45
<i>Positive and Negative Comments</i>	45
<i>Rating Scales</i>	46
HYPOTHESIS 2: QUANTITY OF DATA	47
<i>Mean Number Of Words Per User</i>	47
<i>Mean Number Of Words Per Comment Per User</i>	48
HYPOTHESIS 3: QUALITY OF DATA	48
<i>Irrelevant Comments</i>	48
<i>Unique Ideas Per Group</i>	48
TIME TO COMPLETE SURVEY.....	49
CONCLUSION.....	49
REFERENCES.....	54
APPENDIX A	56
ANNOUNCEMENT FOR RECRUITING	56
APPENDIX B	57
WRITTEN INTRODUCTION/INFORMED CONSENT	57
<i>Attended Condition</i>	57
<i>Unattended Condition</i>	58
APPENDIX C	59
INTERPERSONAL BEHAVIOR FREQUENCY CHECKLIST	59
APPENDIX D	60
WRITTEN DEBRIEFING	60
APPENDIX E.....	61
WEB PROTOTYPE/QUESTIONNAIRE	61
<i>Attended Condition</i>	61
<i>Unattended Condition</i>	77
APPENDIX F.....	99
ATTENDED CONDITION QUESTIONNAIRE	99
APPENDIX G	103
ATTENDED CONDITION MATERIALS	103

List of Tables

1. Remote Evaluation Techniques	10
2. Four Elements of an Effective Prototype	13
3. Attended Condition Web Browser and Monitor Resolution Dispersion.....	20
4. Control of Experimental Groups.....	29
5. Question 1A: Frequency of Category Preference	34
6. Mean Usefulness Rankings for Three Explanation Styles	35
7. Rating Scale Pass/Fail Analysis.....	36
8. Mean Frequency of Interpersonal Behaviors between the Researcher and Each Attended Condition Subject.....	43
9. Experimental Variables Tested and Major Findings	44

List of Figures

1. A 5-point rating scale that measures subjective acceptability ratings.....	3
2. Scenario's Explanation as a Tab.	21
3. Scenario's Explanation as a dockable window.	22
4. Web prototype navigation buttons: Previous and Next.	23
5. Brief Explanation style.....	24
6. Synopsis Explanation style.....	25
7. Report Explanation style.	26
8. Segment 1.....	27
9. Segment 2.....	28
10. Question 2C rating frequencies for both attended (left bar) and unattended (right bar) conditions.....	37
11. Question 3F rating frequencies for both attended (left bar) and unattended (right bar) conditions.....	38

Remote Evaluation: A Comparison of Attended and Unattended Usability Testing Via the Internet

This study was undertaken to determine if attended Internet usability testing data would differ from unattended Internet usability testing data. The Internet is a new technology paradigm that potentially has much to offer to the usability profession. It is a convenient and inexpensive communication medium that allows usability engineers to reach a remote global audience and distribute Web prototypes. The Internet also allows usability engineers to target actual users rather than representative users or internal employees.

In traditional usability testing, a usability engineer is present during the entire testing session. However, Web-based usability testing involves remote users working alone at their personal workstations across the globe. A traditional usability test can range from thirty minutes to two hours. Over the Internet, Beta users, who are often overworked business executives, will probably give no more than twenty minutes of their time for remote evaluation. As a consequence, Web-based usability testing should be divided into a small number of tests on specific topics. The current study is intended to explore the feasibility of testing one of many possible user interface components. If proven adequate, the future approach would be to break complex usability issues into smaller units. A series of small, twenty minute usability studies remotely delivered over the Internet would offer a number of advantages. There would be a time and cost savings of not requiring a usability engineer to be present. More important, the usability test would reach a large number of available users who have expertise in the application area and motivation to improve the product. Finding skilled users to test a sophisticated application can be difficult if the test must be conducted locally.

A Web prototype, a partial user interface model used to illustrate or test design alternatives, was delivered over the Internet to obtain the data in the current study. A survey was added to the Web prototype in the unattended condition. The Web prototype simulated the Explanation feature in a Cognos software product called Scenario 2.0 Beta 1, a data visualization tool. The Explanation was a new view of the user's data that provided a summary and

interpretation of a segment, a term used to describe a unit of similar data. Full details on Scenario 2.0 may be obtained from the Cognos Web site at <http://www.cognos.com>.

Usability Testing

Central to this research was a software product usability test. "Usability means that the people who use the product can do so quickly and easily to accomplish their own tasks" (Dumas & Redish, 1994, p. 4). Constant market pressure from discerning customers who demand more from the software products they purchase has resulted in an emphasis on usability among major software companies (Wiklund, 1994, chap. 1).

Usability testing typically involves the assessment of five user interface dimensions: Learnability, efficiency, memorability, errors, and satisfaction (Nielsen, 1993, chap. 2). Learnability entails the initial learning of a system so that the user can quickly become productive. The learning period should be as short and effective as possible. If a user is able to attain a high level of productivity with the user interface, it has a high level of efficiency. Memorability refers to the ability to leave the user interface for a time and remain productive upon returning, without having to relearn the features previously learned. Low overall error rates, easy recovery from errors when they do occur, and zero system-crashing errors are key elements of the error dimension. Satisfaction involves a favorable subjective opinion about working with a user interface. The user must enjoy working with the user interface. One or more of the elements outlined above are assessed during a typical usability test to uncover usability problems.

Currently, as is the case in many companies, usability testing is conducted in the Cognos corporate environment. The goal of usability testing is to identify problems and propose solutions, thus improving the usability of the product. The participants are employees who are assumed to represent real Cognos customers completing real tasks. They are observed directly without the use of video or audio tape. Comments are recorded on paper forms by a usability engineer. The data is analyzed and summarized in a report that describes each usability problem and one or more proposed solutions. This thesis presents a usability study conducted at Cognos.

Comparison of Judgment Questions

Usability data can be collected using a variety of survey techniques. Spool (1997) conducted a usability test of nine World Wide Web (Web) sites to collect data to help designers create usable Web sites. Sites were carefully selected that appealed to a wide audience, displaying a wide variety of popular, high-profile information. The goal of the usability test was to examine how easy it was for users to find specific information on the Web sites. A survey technique adopted by Spool (1997) was to ask comparison of judgment questions. For example, Spool asked the user to arrive at a single answer in a car retailer's Web site by formulating a decision (judgment) based on the comparison of two or more facts (comparison): "In your opinion, what's the best new convertible for under \$20,000" (Spool, 1997, p. 144).

The most complicated and difficult to answer question that Spool (1997) asked, comparison of judgment, had implications for this thesis. Several of the questions in the current study were designed based on this survey approach. Comparison of judgement was the process involved when analyzing two segments in the Explanation simultaneously and deciding on a correct answer to a multitude of questions.

Rating Scales

Rating scales (Alreck & Settle, 1985, pp. 146-148) are commonly used in usability evaluation to measure single dimensions such as ease of use, satisfaction or acceptability. The scale is a combination of labels and numbers. For example, an acceptability scale consists of two bipolar opposites at each end, separated by equal numeric intervals from one to five. Intermediate values are also labeled throughout the middle of the scale, as shown in Figure 1.

Extremely Unacceptable 1	Slightly Unacceptable 2	Neutral 3	Slightly Acceptable 4	Extremely Acceptable 5

Figure 1. A 5-point rating scale that measures subjective acceptability ratings.

Remote Evaluation

Hartson, Castillo, Kelso, & Neale (1996, p. 3) define remote evaluation to be a new style of usability evaluation “wherein the evaluator, performing observation and analysis, is separated in space and/or time from the user.” Many of the defining features of this new usability evaluation method can best be described as potential advantages and disadvantages.

Potential Advantages of Remote Evaluation

Opportunity to obtain “real-world” data.

Remote evaluation targets appropriate users, an expensive and often prohibitive proposition in traditional usability testing. Users are working with the software in their normal working environments, under their normal working conditions, working as diligently as they normally would, and behaving as they normally would. Remote evaluation “gives a real indication of the feelings of real users using the software to do real work...after they have used it awhile” (Elgin, 1995, p. 43). With remote evaluation, the subject can be tested as they work from within their own office (Hammontree et al., 1994). All of these factors may result in realistic data with high levels of external validity.

Factors that are impossible or not easily reproduced in a usability lab are taken into account by remote usability testing: Exact desktop computer system, office layout, interruptions, background noise, coworkers, and competing priorities (Mitropoulos-Rundus & Muszak, 1997a, p. 10).

Remote evaluation allows a usability engineer to access potential customers who have been using the product to achieve personal goals or to fulfill job requirements rather than answering specific questions arbitrarily formulated by the usability engineer. Thus, data are collected based on natural usage patterns and use cases. In addition, Beta users have been working with the software prior to the usability test, becoming familiar with the underlying usage issues. Consequently, the validity of the data is increased due to an increased knowledge of the software product.

Cost effectiveness.

Remote evaluation can reach users anywhere in the world for the same cost. The distribution expenses associated with mail and telephone surveys are eliminated through the use of existing Internet Web sites and email. Remote testing via the Internet is less expensive than paying a usability engineer to travel to distant sites, bringing real users into the lab, or bringing representative users into the lab.

Usability engineer can not adversely influence the participants.

"Participants will tend to defer to the test monitor and ask many questions if they feel that the test monitor knows all the answers. Being too knowledgeable can also intimidate a participant who may be somewhat nervous and self-conscious about his or her abilities" (Rubin, 1994, p. 72). This situation will not occur during unattended remote evaluation.

In traditional usability testing, the usability engineer might accidentally reveal information or hints. This is not possible with unattended remote evaluation.

It is often difficult for a usability engineer to remain completely neutral and impartial while evaluating a user interface. For example, if a usability engineer dislikes the user interface, he or she might subconsciously record more negative behaviors than positive ones. Conversely, if the usability engineer designed the user interface himself or herself, a vested interest in the success of the usability test may lead the engineer to subconsciously record more positive behaviors than negative ones. The usability engineer should avoid "unintentionally providing cues to the participant about correct performance" (Rubin, 1994, p. 71).

When unattended, participants may disclose data that they would not normally provide in the presence of a researcher.

User expectations.

Users can be easily contacted and asked how effectively the product met their expectations and what functionality is missing without costly long-distance telephone calls, mail surveys that are often ignored, or prohibitive travel expenses (Mitropoulos-Rundus & Muszak, 1997b, p.14).

Potential Disadvantages of Remote Evaluation

Absence of direct observation.

In unattended remote evaluation, the usability engineer is not present during the testing session. Thus, he or she will not be able to directly observe the participant and record behaviors or interpret subtle interactions. Data is limited to self-reports.

Annoyance.

It is important that the remote evaluation technique itself does not annoy the user. For example, long download times or excessive scrolling through a Web prototype may frustrate the user. This annoyance could impact the user's subjective ratings or opinions. Consequently, low satisfaction may be due to excessive scrolling rather than dissatisfaction with the product being tested. The challenge is to "design an interaction technique that will encourage an emotional user to express his or her frustration/elation (and what it is about) effectively over the network without that technique being annoying at other times" (Elgin, 1995, p. 43).

Absence of human assistance.

The usability engineer is unable to provide hints after a specified time limit, if necessary. The engineer can not answer questions from the participant if they are unable to interpret the question or require more information about the user interface. The usability engineer can not encourage participants to continue through periods of frustration.

Honesty.

The question of honesty is a problem with both remote and traditional usability evaluation. In either case, the usability engineer can never be absolutely sure that the answers a subject has given were honest. However, remote evaluation offers several unique opportunities for a respondent to be dishonest.

Perhaps the most critical issue regarding honesty is how does the usability engineer ensure usage? Did the user try all of the main features? Did the user really work with the product every day for two weeks, or did they merely use it for one hour and then set it aside?

The second most critical issue is the accuracy of self-report data. Did the participant answer the questions accurately, or was the participant only interested in answering the survey as quickly as possible to earn the free gift or to cross the survey off their “to do” list?

Third, without the presence of an evaluator, it is impossible to determine the participant's “true” environmental context while answering the survey. Was the participant talking on the phone or browsing the Internet and skimming over the survey questions at the same time? Was the participant completing the usability test with five coworkers present? These events do not occur in the presence of a usability engineer.

Finally, even if participants are asked to fill in a screening questionnaire that inquires about experience characteristics, did they answer honestly? Did they inflate their job title or job description? Were the respondents truly representative of the target users, or was the person answering the test the janitor or the son of an employee who was visiting his mother’s workplace? These events will not occur in the presence of a usability engineer.

Interpersonal Roles of the Usability Engineer

In attended remote evaluation, the usability engineer is present during the entire process to facilitate and enable the testing session. He or she can fulfill a number of interpersonal monitoring roles (Rubin, 1994, chap. 4) that are not possible in unattended usability testing. Unfortunately, the fulfillment of these roles can bias the results. The absence of a usability engineer in unattended testing will eliminate bias caused by the engineer, but the trade-off is the lack of fulfillment of eight interpersonal roles identified by Rubin (1994, chap. 4), identified below.

Welcome the participant.

A usability test begins with the usability engineer welcoming the participant to the usability lab. In doing so, he or she sets the participant at ease by making them feel secure, getting them talking, answering any questions, and responding to any reservations or concerns that the participant might have (Rubin, 1994, chap. 4).

Observe behaviors.

The usability engineer observes the behaviors of the participants as they interact with the user interface. The engineer records meaningful instances that directly relate to usability problems and disregards insignificant or trivial behaviors. Important goals of observation include finding the real “usability gems”, evaluating the implications of particular behaviors on usability, and interpreting the subtleties of speech, emphasis, and posture (Rubin, 1994, chap. 4).

Probe.

A usability engineer can probe a participant for more information when it appears that this information is not forthcoming. Probing is conducted after a time delay only if the participant does not comment on an issue spontaneously. Probing is also effective for examining issues that arise unexpectedly during a testing session or for cross-checking inconsistent comments or behaviors with previous instances of similar comments or behaviors (Rubin, 1994, chap. 4).

Clarify.

There are circumstances when a usability engineer must clarify the content of a participant's comments to determine its implications or remove ambiguity. The usability engineer must also resolve conflicting observations through clarification during a testing session, as they occur (Rubin, 1994, chap. 4).

Flexibility.

A usability testing session follows a written test plan that outlines the sequence of questions and events. A usability engineer must know when to deviate from the test plan, pursuing different avenues to expose unforeseen usability problems. A usability engineer must be prepared for the unexpected, even if the event has disastrous consequences, such as on the rare occasion when a participant becomes so disgusted with a user interface that they leave the test altogether (Rubin, 1994, chap. 4).

Encourage.

Usability engineers commonly ask participants to engage in the think aloud protocol during a usability testing session. “A think aloud protocol is the term given to a special kind of

verbal protocol in which the user says out loud what she is thinking while she is carrying out a task or doing some problem solving" (Preece, 1994, p. 622). It is used to provide information about the relationship between the cognitive activity and the overt behaviour of the participant. Thinking aloud is not a normal activity for most people. Consequently, participants often require reinforcement to think aloud so that the usability engineer can understand and document the rationale behind their behaviors. Thinking aloud may also help participants keep frustration levels to a minimum by allowing them to vent their frustrations verbally.

Participants often need to be encouraged to provide both negative and positive feedback. It is difficult for some people to make negative comments due to shyness or concern about offending the usability engineer. It is important that each participant provides as many comments as possible, withholding nothing.

The usability engineer should encourage participants to work through moments of frustration. A usability engineer can keep a participant working through periods of frustration by deflecting the problem back to the product rather than at the participant, thereby protecting their self-esteem (Rubin, 1994, chap. 4).

Debrief the participant.

Each usability test should end with a formal debriefing. The usability engineer debriefs each participant based on the events of the session, extracting "a participant's innermost thoughts and feelings about the previous two hours of work" (Rubin, 1994, p. 70). The debriefing is also a time for final questions before the participant is dismissed.

Now that the important issues regarding attended and unattended usability evaluation have been presented, various remote evaluation techniques will be described.

Types of Remote Evaluation

Remote evaluation can be conducted through a variety of techniques. Hartson et al. (1996) describe seven remote evaluation techniques (see Table 1).

Table 1

Remote Evaluation Techniques

Technique	Description	Category
Remote questionnaire/survey	A questionnaire is activated by an event in the software during usage	Unattended
Portable evaluation	The laboratory is taken to the user's normal workstation	Attended
Local evaluation at remote site	A usability firm is outsourced to conduct the test either locally or remotely, with the results transmitted to the sponsoring company via the Internet	Attended or unattended
Remote-control evaluation	The usability engineer's computer is controlled to mimic the user's actions remotely through the use of video and audio capture equipment	Attended or unattended
Video conferencing as an extension of the usability laboratory	The usability engineer conducts the testing and captures the user's actions remotely using video and audio capture equipment or shared whiteboard tools	Attended
Instrumented remote evaluation	Metering code is added to the software to record usage patterns and keystrokes	Unattended
Semi-instrumented remote evaluation	Users are trained to identify and document significant usability events themselves	Unattended

Note. Adapted from "Remote Evaluation: The Network as an Extension of the Usability Laboratory" by H. R. Hartson, J. C. Castillo, J. Kelso, & W. C. Neale, 1996, In Proceedings of the CHI '96, 1996 (Vancouver, BC, Canada, April 13-18, 1996). (pp. 228-235). New York: Association for Computing Machinery.

Remote Questionnaire/Survey

The first entry in Table 1, remote questionnaire/survey, has implications for this thesis. "A software application can be augmented to display a user questionnaire to gather subjective user

preference data about the application and its interface. The appearance of the questionnaire, requesting feedback related to usability, is triggered by an event (including task completion) during usage” (Hartson et al., 1996, p. 7). The results of the questionnaire are sent back to the usability engineer by a method unspecified in the article.

Questionnaires and surveys have one potential limitation. “Approaches based on remote questionnaires have the advantage that they capture remote user reactions while they are fresh, but they are limited to subjective data based on questions pre-written by developers or evaluators” (Hartson et al., 1996, p. 7). Contrary to Hartson et al. (1996), objective data such as feature usage frequencies, error rates, number of usability problems, and task times may also be solicited by questionnaires and surveys if they are designed creatively.

Internet-based questionnaires and surveys are prone to the same types of bias as their mail or telephone counterparts. Any bias reduces the reliability and validity of the data (Alreck & Settle, 1985, chap. 4). There are three main types of bias that may occur during remote questionnaire/survey data collection: Sampling, instrumentation, and response bias. All are equally important in both attended and unattended survey administration.

A sampling bias results from the process of distributing the survey to its respondents. For example, this bias often occurs when respondents are given the choice whether to participate or not. To help reduce a sampling bias, the degree to which the participant is able to refuse responding should be minimized. For example, users may be asked to fill out a usability survey as part of the software usage contract. The ideal situation would be for participants to choose in a random fashion whether or not to respond to a survey, indicating no bias (Alreck & Settle, 1985, chap. 4).

If a respondent’s decision to participate is based on the content of the survey, a non-response bias will result. For example, dissatisfied Beta users may be more likely to respond to the survey and voice their concerns, while satisfied users may not feel that the sponsoring company is interested in positive feedback. A non-response bias could also be caused by demographic factors (e.g. employees have more spare time to complete surveys than managers)

or psychographic factors (e.g. dissatisfied users). Virtually every survey will contain non-response bias (Alreck & Settle, 1985, chap. 4).

The second category of bias, instrumentation, indicates that several aspects of the questionnaire itself can introduce bias: Instructions, questions, scales, and response options (Alreck & Settle, 1985, chap. 4). The way in which the questions are phrased can reduce the validity and reliability of the survey. Instrumentation bias can be caused by unstated criteria, which often leads to multiple interpretations, questions that do not apply to a particular respondents experience or condition, and overgeneralizations. Using a specific example in a question can bias the results toward the context of that example, as can overspecificity, ambiguity of wording, or overemphasis of certain facts (Alreck & Settle, 1985, pp. 104-111).

The final category of bias that has implications for remote questionnaires/surveys is response bias. A response bias is caused by the intellect or predisposition of the respondent (Alreck & Settle, 1985, chap. 4). For example, some people are prone to provide mostly positive answers, while others tend to rate everything in a negative direction. A wide sampling range throughout attributes such as age, education, and experience can also assist in randomizing the effects of response bias.

Web Prototyping

Remote questionnaires/surveys are defined by Hartson et al. (1996) to be existing software products in which a survey has been added to the product's code. Web prototyping is a proposed extension to remote questionnaires/surveys (Hartson et al., 1996), introduced in the current study. Web prototyping is a new category of prototyping that utilizes static screen images distributed via the Internet to quickly and easily simulate the appearance and interactions of specific aspects of a user interface. The images are saved in either the GIF or JPEG image file format for display on World Wide Web pages.

Web prototyping is easier and faster to implement than remote questionnaires/surveys (Hartson et al., 1996) because it does not involve modifying the software application itself to insert a questionnaire. A Web prototype is a separate entity from the software that it evaluates.

A prototype is a partial user interface model that is used to illustrate or test a design or several design alternatives (Connell & Shafer, 1989, chap. 1). The prototype is created in a rapid and flexible manner so that it may easily be redesigned should it fail to demonstrate adequate usability during testing. An effective prototype is composed of four elements, as shown in Table 2.

Table 2

Four Elements of an Effective Prototype

Element
1. Functional after a minimal amount of effort
2. A means for providing users of a proposed application with a physical representation of key parts of the system before system implementation
3. Flexible- modifications require minimal effort
4. Not necessarily representative of a complete system

Note. From Structured Rapid Prototyping (p. 2) by J. L. Connell and L. B. Shafer, 1989, Englewood Cliffs, NJ: Prentice-Hall. Copyright 1994 by Prentice Hall, Inc.

The screen images in a Web prototype form a screenplay or storyboard that represents the possible interactions of the actual software product. As the users interact with the Web prototype, they are asked to fill out a questionnaire at appropriate stages in the storyboard. A storyboard is conceptually similar to usability walkthrough techniques such as cognitive walkthroughs (Wharton, Rieman, Lewis & Polson, 1994), cognitive jogthroughs (Hammontree et al., 1994), and collaborative walkthroughs (Hammontree et al., 1994). In essence, all walkthrough approaches evaluate users as they explore and learn about a proposed user interface while completing representative tasks or scenarios (Wharton et al., 1994).

There are several key elements that must be considered when creating an effective Web prototype. First, questions should be asked immediately following the applicable section of the prototype rather than saved until the end of the testing session. This will result in reliable feedback based on an appropriate context.

Second, “since the usability specialist does not typically speak directly with the sites and cannot help the sites through rough spots or ask them questions, tasks and questions need to be set up differently from typical usability test tasks....tasks need to be more detailed and directive to compensate for lack of direct contact” (Lahti & Kim, 1997, p. 478). Tasks must be very direct and explicit in their introductions, instructions and explanations.

Third, the questions must be separated from the tasks so that participants do not see them prior to finishing the task, which could potentially bias the results by either giving “away how to do the task” (Lahti & Kim, 1997, p. 478) or leading participants to respond in a certain way.

Finally, the Web prototype should be pilot tested to ensure that the tasks and questions are reasonable and that the estimated completion time is acceptable (Lahti & Kim, 1997).

Potential advantages of Web prototyping.

A remote Web prototype is passive. The prototype itself does not do anything or interfere with the user undertaking the test, which minimizes intrusion (Elgin, 1995). It is the participant who is in control of how fast and when they complete the usability test. Conversely, in a traditional usability testing session, a usability engineer controls or enables the testing process.

Web prototyping is excellent at simulating and testing one small part of the total user interface. Rapid updates can be made to the feature based on usability feedback solicited early in the development process (Dumas & Redish, 1994, chap. 5). The prototype itself can be used to demonstrate a specific design, facilitate communication about the design, and serve as a “live” functional specification. When two or more competing design solutions exist, evaluators can quickly and easily decide which Web prototype demonstrates higher usability.

Because a Web prototype is scripted and follows a specific storyboard of events, the usability engineer knows the exact context and moment in which the usability problem occurred, reducing ambiguity in interpreting the cause of the usability problem. “It is, of course, desirable to know as much as possible about the user’s exact interactions, configuration and goals, but in many cases it is not practical to ask the user to formulate these details” (Elgin, 1995, p. 43).

Web prototyping is expeditious. At its simplest level, storyboard screens can be copied from an existing software product or drawn in a Paint program. Web prototyping is also inexpensive. The technique requires minimal development effort and time to create the storyboard screens.

Electronic questionnaires improve the efficiency of data collection. Collection is easily automated through the use of HTML forms, server-side program languages such as Common Gateway Interface (CGI), and spreadsheets or database software. The process can essentially be paperless.

Potential disadvantages of Web prototyping.

Simulating an interactive user interface feature with static images requires a high level of creativity to effectively represent the original interactive functionality in a static representation. In its implementation, Hypertext Markup Language (HTML), the language of World Wide Web pages, may require an unacceptable number of images or amount of scrolling to implement an appropriate storyboard. This could increase the user's frustration, resulting in lower subjective ratings. Users may also lose their context or train of thought.

Static images may require an unacceptable amount of text explanation to provide sufficient context to ensure that all users will interpret the interface issue in the same way.

Web prototyping is a new and untested method of usability testing. Until a process that incorporates templates, standards, guidelines, and recent technology is developed, the time required to implement a Web-based usability test will be high.

As the size of the feature set under investigation increases, the number of images required increases, resulting in an unwieldy prototype. If a usability test requires longer than fifteen minutes to complete, the entire test should be subdivided into two or more separate and smaller tests. The subdivision may also increase the time required to create the whole test.

Using static Web-based screens might be slower than using a fully interactive prototype or Beta product because of the need to scroll from screen to screen, download images, and retrace previous steps to maintain context.

If the storyboard only tests specific parts of a user interface, the usability test might overestimate or underestimate the true usability of the user interface based on the subset of available interactions.

A storyboard offers limited opportunities for a user to explore, become misguided, or even lost, perhaps limiting the total number of usability problems that would be uncovered if the user was working with the full version of the software (Dumas & Redish, 1994, pp. 73-74). A storyboard is unable to verify the occurrence of global usability problems (Nielsen, 1990, Dumas & Redish, 1994, p. 74) because it can only test a subset of the total user interface.

All comments solicited from a remote HTML survey must be typed by the respondent. If some respondents are not effective typists, they will be reluctant to provide comments, and the amount of feedback provided by unattended remote evaluation will be reduced. However, in the future, voice recording could eliminate the need for respondents to type their own comments, suggestions, and rationales.

The security of a Web-based survey can be dubious. If the sponsoring company's competitors discover the survey, they can send fake responses to either confuse the usability engineer or waste their time by adding noise. If the usability test involves placing a Web prototype on the Internet, a company's competitors may be able to find the site and copy the designs. Passwords given to Web prototype users can prevent unauthorized access, but the potential for sabotage is not completely eliminated.

Current Study

The Internet is a convenient and inexpensive communication medium that allows usability engineers to reach a remote global audience virtually instantaneously and without incurring travel expenses. In traditional usability testing, a usability engineer is present during the testing session. In remote survey evaluation, the survey is conducted in the absence of a usability engineer, freeing their time for other tasks. Remote evaluation targets "real" users who have been using the software product under study rather than representative users or internal employees who have never been introduced to the product. Internet-based surveys assess real users completing real

tasks, resulting in realistic data high in external validity. In addition, Beta test customers have worked with the product prior to answering the survey, increasing the validity of the data.

At present, there are no published studies that systematically compare attended and unattended usability evaluation. The current study was conducted to determine whether the data collected from attended usability evaluation differs from unattended usability evaluation data.

It was hypothesized that attended and unattended subjects would be equivalent in their preferences, usefulness rankings, scale ratings, pass/fail decisions, number of correct responses, and mean number of positive and negative comments per user per question. It was postulated by the researcher that these types of usability issues would be independent of the condition from which they are collected.

It was hypothesized that the quantity of data obtained from open-ended questions in the attended condition would be higher than the quantity obtained in the unattended condition. In the attended condition, the researcher would record all comments made by the subject. Conversely, unattended subjects would have the additional impediment of typing their own comments, resulting in reduced quantity. Hesitant attended subjects would be probed by the researcher for answers if necessary, but hesitant unattended subjects could leave questions unanswered.

It was hypothesized that the quality of data obtained from open-ended questions in the attended condition would be higher than the quality obtained in the unattended condition. Specifically, it was hypothesized that attended subjects would provide more unique comments per subject than unattended subjects. A unique comment is not overlapped in content by another comment from the same subject in the same question. If one subject makes two comments about two different issues in the same sentence, this will count as two unique comments. For example, a subject might respond to a survey question with "A is a good layout, but I like the way B handles printing." These are two unique comments about the user interface: A is the best screen layout and B printing is preferred over A printing. Thus, higher quality is represented by higher numbers of unique comments per subject.

Second, it was hypothesized that attended subjects would provide fewer irrelevant comments than unattended subjects. An irrelevant comment is extraneous with regard to the specific usability issue under investigation. For example, an irrelevant comment might take the form of a comment about the user interface that has nothing to do with the current question being asked. Common examples could be comments in the form of questions posed to the researcher or off-topic comments. Attended subjects would be able to ask the researcher questions prior to responding, allowing them to make informed and directed comments. Conversely, unattended subjects would make more irrelevant comments because they would be unable to seek clarification or ask questions of the researcher, limiting their ability to grasp the full context of the prototype and resulting in more irrelevant comments.

Third, it was hypothesized that attended subjects would make fewer process comments than unattended subjects. A process comment requires a response from the researcher in order to transform it into a useful piece of information. In the unattended condition, no guidance is possible. In this case, process comments are comments that would result in guidance if a researcher was present. For example, a process comment might be a subject's statement that is unclear. Consequently, the researcher would probe the subject until he or she fully understood the content of the comment. Conversely, unattended subjects might make more "no opinion" comments throughout the survey because it would be an effortless way to end the question and complete the survey faster. A "no opinion" comment is typically followed up with probing by a researcher in attended testing to obtain a firm opinion. Higher quality is represented by lower numbers of process comments. Process comments reduce quality because they require further probing or clarification by the researcher to turn them into useful usability-related information.

Method

Participants

Sixty subjects were randomly selected from an established contact list of employees willing to participate in usability testing. All subjects were Ottawa-area Cognos employees. In total, sixty-three employees were contacted, but three were unavailable for testing due to encroaching deadlines.

All subjects were treated in accordance with the ethical standards of the American Psychological Association. Deception was not used, and subjects were given the opportunity to withdraw at any time.

Apparatus

The study was conducted at the Cognos Riverside location. All subjects completed the experiment using their personal workstation. A typical workstation at Cognos at the time of the study was a Pentium Pro 150 desktop computer with a 21 inch colour monitor. The corporate operating system is either Windows 95 or Windows NT, and the corporate Web browser is Netscape Navigator 3.0, although employees are free to obtain an alternate browser of their choice. The Web prototype was accessed from the corporate Intranet server. Video taping equipment is not used while usability testing at Cognos.

Information about the subject's Web browser and monitor resolution was obtained by the researcher in the attended condition and summarized in Table 3. To reduce interaction between the researcher and the subject as much as possible in the unattended condition, this information was not solicited from unattended subjects. There was no reason to expect that the two experimental conditions differed on Web browser or monitor resolution because subjects were drawn from the same population and randomly assigned to the experimental conditions.

Table 3

Attended Condition Web Browser and Monitor Resolution Dispersion

Element	Number of subjects
Web browser	
Internet Explorer 3.0	4
Internet Explorer 4.0	2
Netscape Navigator 3.0	20
Netscape Communicator 4.0	4
Monitor resolution	
640 by 480	1
800 by 600	4
1024 by 768	21
1280 by 1024	3
1600 by 1200	1

There were two versions of the Web prototype, as shown in Appendix E. Both were identical in content except the Web prototype in the attended condition did not contain any survey questions because the researcher asked all questions verbally. Attended subjects were given two pages of supplemental materials such as diagrams and rating scales to assist them while answering the questions (see Appendix G). Conversely, the survey questions and supporting materials were included within the Web prototype for the unattended condition. The survey in Appendix E contained three comparison of judgment questions: 3A, 3B, and 3E. Five-point rating scales were used in questions 1C, 2C, and 3F.

The Web prototype in both experimental conditions was a series of static GIF images that represented the Explanation feature in Scenario 2.0 Beta 1. Three usability issues were explored: Explain display, Explanation style, and Explanation interpretation.

The first issue involved the two different display locations for the Explanation: Tab and dockable. As shown in Figure 2, the Explanation could be displayed prominently as a Tab. Tabs are useful for maximizing the display of mutually exclusive information on a computer screen. Scenario had four other mutually exclusive Tabs: Graph, Tree, Bookmarks, and Data.

Figure 2. Scenario's Explanation as a Tab.

As shown in Figure 3, the Explanation could also be displayed as a dockable window. A dockable window was useful if the user needed to see the Explanation and one of the other Tabs at the same time.

Figure 3. Scenario's Explanation as a dockable window.

The Explanation's two display options were presented to the subjects in a linear storyboard format with one idea per page. The steps required to change the display were outlined in detail, and the results of specific actions were illustrated. The subject moved through each stage in the storyboard by clicking on Next or Previous buttons, shown at the bottom of Figure 4, which eliminated the need for scrolling through the browser window via the scrollbars. After following the storyboard, subjects were asked which display they preferred and two related questions.

Figure 4. Web prototype navigation buttons: Previous and Next.

The second usability issue presented the three levels of detail for the Explanation style: Brief, Synopsis, and Report. The Brief explanation contained the lowest level of detail (see Figure 5). Synopsis and Report explanations both provided an equivalent amount of detail, but Synopsis used an easily scannable bullet and footnote format (see Figure 6) and Report used a text-based format (see Figure 7) better suited for printed copies. Subjects cycled through GIF images of each Explanation style and ranked the usefulness of each. They were asked how acceptable it was if there was only one Explanation style.

Figure 5. Brief Explanation style.

Figure 6. Synopsis Explanation style.

Figure 7. Report Explanation style.

In the previous two usability issues, the subject was presented with alternative display formats for the same data. The final usability issue, Explanation interpretation, presented two different data sets. GIF images of the Explanations for two data segments were shown to the subject, shown in Figure 8 and Figure 9, and then subjects were asked to compare the content of the two Explanations and make a correct judgment about the data.

Figure 8. Segment 1.

Figure 9. Segment 2.

Procedure

One categorical independent variable with two experimental conditions was tested. In one condition, attended, the researcher was present during the usability testing session. In another condition, unattended, the researcher was not present during the usability testing session. Three experience variables were measured: Statistics knowledge, Scenario expertise, and Internet browsing expertise.

Experimental control was enforced by keeping as many factors constant as possible while at the same time maintaining the integrity of the two experimental conditions (see Table 4).

Table 4

Control of Experimental Groups

Factor	Attended	Unattended
Held constant		
Announcement for recruiting	See Appendix A	See Appendix A
Written introduction/informed consent	See Appendix B	See Appendix B
Testing environment	User's workstation	User's workstation
Subjects	Cognos employees	Cognos employees
Browser	User's browser	User's browser
Questionnaire	Version 1.0	Version 1.0
Welcome the participant	Yes	Yes
Written debriefing	See Appendix D	See Appendix D
Not held constant		
Web Prototype	No questionnaire	Questionnaire
Researcher present	Yes	No
Subjects read questionnaire	No	Yes
Subjects record own responses	No	Yes
Researcher observes behaviors	Yes	No
Researcher probes subjects	Yes	No
Researcher clarifies subjects	Yes	No
Researcher answers questions	Yes	No
Researcher provide hints to subjects	Yes	No
Flexibility in the testing plan	Yes	No
Researcher encourages subjects	Yes	No

A usability test can potentially assess an unlimited number of issues. "Usability testing can be thought of as a collaboration between users and product development teams" (Hammontree, Weiler & Nayak, 1994, p. 22). In the current study, the testing focus was reduced to three core issues, derived from interviews with one of Scenario's developers and Scenario's product manager. The product manager requested usability data on the new Explanation feature because it had never been assessed before due to its newness. The developer outlined three issues of contention among the development team that could be resolved through usability testing: Explanation display, Explanation style, and Explanation interpretation.

Each potential subject received an identical "announcement for recruiting" email message, as shown in Appendix A. Additional communication by email, telephone or workstation visits was necessary on 17 occasions due to the complexity of the particular subject's work schedule. Reminder messages, identical to the first email message, were resent to five subjects one week after the initial message.

At a pre-arranged time, the researcher met the subject at their workstation. The researcher assigned the subject to one of the two testing conditions: Attended or unattended. Each subject was given a written introduction dependent upon their experimental condition that was also an informed consent form (see Appendix B).

In the attended condition, 10 subjects were asked at random to consent to have the session audio taped on a Sony micro cassette recorder to evaluate the reliability of event transcription. All 10 subjects agreed. The total number of differences between the notes taken by the researcher and the notes transcribed from the audio tapes among the 10 audiotaped subjects was recorded. There was an average of 3.1 differences ($SD = 1.2$) for each of the 10 subjects. All differences took the form of comments that were missed by the researcher. The unattended condition did not require audio tape transcription because the subjects typed their own comments directly, eliminating the second-person translation inherent in the attended condition.

The researcher recorded the usability testing session start and end times in the attended condition. In the unattended condition, the subject was asked to enter their starting time, and the

browser added the end time automatically. In both experimental conditions, the subject followed the Web prototype shown in Appendix E and answered questions when appropriate. The researcher remained present during testing in the attended condition, and asked all of the survey questions verbally from a paper form, as shown in Appendix F. If the question included several response options, these were read to the subject word-for-word. One supplemental graphic, an acceptability scale, a satisfaction scale, and three demographic scales were made available as reference materials for the subject on a sheet of paper during testing, as shown in Appendix G. The researcher recorded all responses. In the attended condition, all survey questions were removed from the Web prototype because the researcher asked the questions verbally, but the content and usability issues remained identical in both testing conditions.

As shown in Appendix C, the researcher recorded eight of his interpersonal behaviors with each subject in the attended condition in a frequency checklist. The results formed an index of the number of behaviors that occurred only during attended testing because in the unattended condition, the researcher left the subjects alone to complete the survey once they began reading the introduction screen.

When the subjects completed the experiment they were thanked for their time and excused, either verbally in the attended condition or through the prototype itself in the unattended condition. Subjects in both conditions were emailed a debriefing on the day of their participation, as shown in Appendix D.

Results

Each question or group of related questions was analyzed separately through the use of an appropriate statistical technique: ANOVA, MANOVA, t test, or chi-square. The probability of a Type I error was set at .05.

Assumptions

The general approach to addressing the assumptions of the statistical tests used in this research was to investigate normality and homogeneity of variance. Both were violated occasionally. However, the statistics were robust to violations of these assumptions due to equal sample sizes per group and a large overall sample size (Keppel, 1991).

Experience Variables

As shown in Appendix E, subjects in both experimental conditions were compared on three self-reported experience variables: Statistics knowledge, Scenario expertise, and Internet browsing expertise. There was no evidence that the two groups differed on these important variables.

Statistics knowledge was measured on a scale from 1 (No Knowledge) to 10 (I am a Statistician). Attended subjects were not significantly different in statistics knowledge ($M = 4.4$, $SD = 2.1$) from unattended subjects ($M = 3.8$, $SD = 1.7$), $F(1, 58) = 1.48$, ns , $\eta^2 = .03$.

Scenario expertise was measured on a scale from 1 (No Expertise) to 10 (I am a Power User). Attended subjects were not significantly different in Scenario expertise ($M = 3.1$, $SD = 2.5$) from unattended subjects ($M = 2.6$, $SD = 2.2$), $F < 1$.

Internet browsing expertise was measured on a scale from 1 (No Expertise) to 10 (I am a Power Browser). Attended subjects were not significantly different in Internet browsing expertise ($M = 7.9$, $SD = 1.9$) from unattended subjects ($M = 7.6$, $SD = 2.0$), $F < 1$.

To determine if the experience of subjects was related to performance answering survey questions, univariate ANOVAs were performed with testing condition (attended versus unattended) as the independent variable and a subset of dependent variables derived from the survey questions. In addition, the same analyses were performed with statistics knowledge,

Scenario knowledge, and Web browsing expertise as covariates to see if statistically controlling for the effects of experience had any effect on the results for testing condition. For all dependent variables, the conclusions were the same for ANOVAs and ANCOVAs. Therefore, only results for ANOVAs are presented.

Time to Complete Survey

To determine if both experimental groups were answering the survey in an equivalent time period, the time required by each subject to complete all of the questions was recorded to the nearest minute. Attended subjects required significantly less time in minutes to complete the survey ($M = 21.6$, $SD = 6.8$) than unattended subjects ($M = 25.9$, $SD = 9.5$), $F(1, 58) = 4.07$, $p < .05$, $\eta^2 = .07^1$.

Survey Analysis

For ease of presentation, the data collected from the survey was divided into two categories: Forced-choice questions and open-ended questions.

Forced-Choice Questions

Forced-choice questions provided the subject with a specific set of response options. The results of each of the 11 forced-choice question in the survey are described below.

Display preference.

Question 1A asked subjects which type of display they preferred: Tab or dockable window (see Figure 2 and Figure 3, respectively). The results are presented in Table 5. A Pearson chi-square test for independence was conducted to determine if display preference was independent of experimental group. There was no evidence that display preference was related to group membership, $\chi^2(1, N = 60) = 0.62$, ns , $\eta^2 = .01$.

¹ For the purposes of this thesis, an outlier was defined statistically as a data point two or more standard deviations above or below the mean. There were three outliers in the time to complete the survey. When time outliers in both experimental conditions greater than or equal to 40 minutes were discarded, attended subjects again required significantly less time in minutes ($M = 20.3$, $SD = 4.6$) to complete the survey ($M = 23.9$, $SD = 7.4$), $F(1, 58) = 4.82$, $p < .05$, $\eta^2 = .08$.

Table 5

Question 1A: Frequency of Category Preference

Category	Frequency	
	Attended	Unattended
Tab	47% (14/30)	37% (11/30)
Dockable	53% (16/30)	63% (19/30)

Printing preference.

Question 1E asked subjects if they wanted to print from both the Tab and dockable window displays. In both attended and unattended conditions, 29 out of 30 subjects said yes.

Rank the three Explanation styles.

Question 2A asked subjects to rank the Brief, Synopsis, and Report Explanation styles in order of usefulness to their individual needs on a scale from 1 (Most Useful) to 3 (Least Useful). (see Figure 5, Figure 6, and Figure 7, respectively). The mean rankings are presented in Table 6.

To assess the effect of testing condition on the usefulness ranking of each Explanation style, a two (group: attended vs. unattended) by three (Explanation style: Brief, Synopsis, Report) mixed factorial ANOVA was conducted on subject's usefulness ratings. The sphericity assumption was not violated. The main effect for Explanation style was significant ($F(1, 57) = 6.63, p < .05, \eta^2 = .10$). The interaction between experimental group and Explanation style, shown in Table 6, was not significant, $F < 1$.

This test did not allow an overall comparison of the groups because each subject's rankings summed to 6 ($1 + 2 + 3$) over the three Explanation styles. The nonsignificant interaction between group and Explanation styles indicated that the rankings given were independent of group. This was confirmed by directly comparing groups on each of the three explanation styles using three between-subjects ANOVAs. There were no significant differences, all $F_s < 1$.

Paired comparisons were used to follow up the significant main effect for Explanation style. Subjects rated the Synopsis style significantly higher than the Brief style, $F(1, 57) = 20.70$,

$p < .01$). Subjects also rated the Report style significantly higher than the Brief style, $F(1, 57) = 6.66$, $p < .05$). The Synopsis and Report styles were not rated significantly different ($F(1, 57) = 2.37$, ns). If developers were interested in implementing only one Explanation style, it could be either Synopsis or Report. Both were rated significantly more useful than Brief, but equal in usefulness to each other.

Table 6

Mean Usefulness Rankings for Three Explanation Styles

Condition	Synopsis	Report	Brief
	<u>M</u> (<u>SD</u>)	<u>M</u> (<u>SD</u>)	<u>M</u> (<u>SD</u>)
Attended ^a	1.6 (0.7)	2.0 (0.9)	2.3 (0.7)
Unattended ^b	1.7 (0.8)	1.9 (0.8)	2.4 (0.7)

^a $n = 29$. ^b $n = 30$

Correct/incorrect responses.

There were four questions on the survey that required a correct or incorrect response: 3A, 3B, 3C, and 3E in Appendix E. Each subject was assigned a score between 0 and 4 that corresponded to the total number of correct responses. Attended subjects ($M = 3.4$, $SD = 0.9$) were not significantly different in their total number of correct responses than unattended subjects ($M = 3.6$, $SD = 0.6$), $F < 1$.

Definition of outliers.

Question 3D asked subjects what outliers represented. The question was intended to determine whether subjects were using the Explanation as it was meant to be used: to interpret their data. A definition of what outliers represented, "Outliers may represent either unusual data or invalid records" was offered in a footnote at the bottom of the Explanation for both segments. In both attended and unattended conditions, 20 out of 30 subjects were correct in identifying the definition from the Explanation. The other 10 subjects in each condition also gave a correct definition of an outlier, but provided their own definition rather than the definition from the

Explanation itself. Therefore, all of the users were correct because there were two ways to answer the question correctly.

Rating scales.

There were three 5-point rating scales on the survey. Each scale was summarized according to two criteria based on the ratings provided, as shown in Table 7. The first criterion analyzed whether the attended and unattended subjects provided significantly different mean ratings when analyzed using an ANOVA. The second criterion involved applying a standard usability “pass” or “fail” decision to each question. Rather than using a statistical test, a rating scale was labeled “pass” from a usability perspective if 80% of all respondents answered the question with a rating of “Neutral” (3) or above. If the scale failed to meet the criteria, the question was labeled “fail.” This is a typical approach to a usability goal, and 80% with ratings of neutral or above is fairly liberal in terms of attaining a pass decision. In the case of a failure, the usability engineer would recommend that the software developers address the issue raised by the question.

Table 7

Rating Scale Pass/Fail Analysis

Question	<u>F</u>	Attended		Unattended	
		<u>M</u>	% Neutral or Above	<u>M</u>	% Neutral or Above
1C	0.49	3.3	63% (19/30) Fail	3.6	77% (23/30) Fail
2C	4.24*	3.4	63% (19/30) Fail	2.7	57% (17/30) Fail
3F	0.79	3.4	70% (21/30) Fail	3.7	90% (27/30) Pass

* $p < .05$

Question 1C asked subjects how acceptable it was that the Explain Tab used a different display format than the other four Tabs. Higher ratings indicated higher acceptance. With respect to the mean difference criterion, attended subject’s ratings (M = 3.3, SD = 1.3) were not significantly different from the unattended subject’s ratings (M = 3.6, SD = 1.3, F < 1. Regarding the pass/fail criterion, both experimental conditions failed to achieve the “pass” criteria (see Table 7).

Question 2C asked subjects how acceptable it was if there was only one Explanation style rather than three. It was left to the subject to decide how one style would be created from the current three: Brief, Synopsis, and Report. Higher ratings indicated higher acceptance. Frequencies of ratings are shown in Figure 10. With respect to the mean difference criterion, attended subject's ratings ($M = 3.4$, $SD = 1.4$) were significantly higher than the unattended subject's ratings ($M = 2.7$, $SD = 1.4$, $F(1, 58) = 4.26$, $p < .05$, $\eta^2 = .07$). However, concerning the pass/fail criterion, both experimental conditions failed to meet the "pass" criteria (see Table 7).

Figure 10. Question 2C rating frequencies for both attended (left bar) and unattended (right bar) conditions.

Question 3F asked subjects how satisfied they were with the "interpretive" or "analyst in a box" aspect of the Explanation. Higher ratings indicated higher satisfaction. With respect to the mean difference criterion, attended subject's ratings ($M = 3.4$, $SD = 1.3$) were not significantly different from the unattended subject's ratings ($M = 3.7$, $SD = 1.1$), $F < 1$. However, regarding the pass/fail criterion, the unattended subjects successfully met the "pass" criteria but the attended subjects were unsuccessful (see Table 7). Figure 11 shows that subjects in the unattended condition shifted from a 3 rating, which would have been a pass, to a 2 rating, which was a fail. The difference was caused primarily by only one unattended subject rating their satisfaction 2, as

compared to 7 attended subjects (see Figure 11). Additionally, 57% of the comments from unattended subjects focused on their dissatisfaction with the Explanation, compared to only 46% of the comments from attended subjects.

Figure 11. Question 3F rating frequencies for both attended (left bar) and unattended (right bar) conditions.

Open-Ended Question Analysis

The survey asked a total of seven open-ended questions. An open-ended question does not have a specific response option; the content is provided at the discretion of the respondent. A multivariate ANOVA was conducted on eight dependent variables that represented all of the comment categories summarized below to determine if group differences existed. The first four comment categories are summaries: mean number of words per user, mean number of comments per user, mean number of words per comment per user, and mean number of unique comments per subject. The last four comment categories were subjectively assigned: positive comments, negative comments, irrelevant comments, and process comments.

The researcher enlisted a second usability engineer to assist in categorizing comments. The inter-rater reliability was 89%. In the case of a disagreement, the comment category was assigned by the author.

Mean number of words per question.

A total of 8476 words were used to answer the seven open-ended questions by all sixty subjects: 4333 from the attended condition and 4143 from the unattended condition. Averaged across all seven open-ended questions, each subject in the attended condition used a mean of 20.6 words ($SD = 4.9$) to answer each question. The unattended subjects used a mean of 19.7 words per question ($SD = 8.0$). The two experimental groups were not significantly different in mean number of words per question, $F < 1$.

Mean number of comments per question.

A comment was defined as one idea spoken or written by the subject. In some cases, one sentence contained two ideas and was counted as two comments: "While I prefer the dockable window because it allows me to see my data for a reference base, I really prefer the 'undocked' or 'floating' versions." A total of 677 comments were collected from all 60 subjects throughout the seven open-ended survey questions: 393 from the attended condition and 283 from the unattended condition. Averaged across all seven open-ended questions, each subject in the attended condition made a total of 1.9 comments ($SD = 0.4$) per question. The unattended subjects submitted a significantly lower mean of 1.3 comments ($SD = 0.5$) per question, $F(1, 58) = 23.72$, $p < .01$, $\eta^2 = .29$.

Mean number of words per comment per question.

Averaged across all seven open-ended questions, each subject in the attended condition used a mean of 11.2 words per comment ($SD = 2.4$) when answering each question. The unattended subjects used a mean of 13.5 words per comment ($SD = 8.0$), which was significantly higher $F(1, 58) = 4.79$, $p < .05$, $\eta^2 = .08$. This finding follows from the results presented above; unattended subjects provided fewer comments but the same number of words for a significantly higher ratio of words per comment.

Mean number of unique comments per subject.

A unique comment is not overlapped in content by another comment from the same subject. For example, in question 1B, one subject stated "If I'm looking at an explanation of

something, I would like to be able to refer back to the original 'something'. But, why can't I print it?" These are two unique comments that do not overlap in content from the same subject in the same question. One sentence from a subject can contain one, two, or more unique comments as long as the content of each does not overlap.

Conversely, a typical example of two non-unique comments was "Why do you need a split screen anyway? That much flexibility is almost too much." These two comments are not unique because each comment overlaps the content of the other and discusses the same central issue. The first comment creates the context for the second comment.

Averaged across all seven open-ended questions, each subject in the attended condition made a mean of 1.7 unique comments ($SD = 0.3$) per question. The unattended subjects submitted a significantly lower mean of 1.2 unique comments ($SD = 0.4$) per question, $F(1, 58) = 36.69$, $p < .01$, $\eta^2 = .39$.

Recall that a unique comment is not overlapped in content by another comment from the same subject in the same question. Often, several different subjects made the same comment. The researcher examined all of the total ideas from all of the subjects in each experimental group and counted how many unique, non-overlapping ideas were provided. For example, seven out of 30 attended subjects suggested that the Explanation should include definitions for important statistical terms such as skewness. Regardless of how many subjects made a particular comment, that particular comment was counted only once as a unique idea for the group.

When the content of similar comments was grouped across all subjects, an index of the number of unique ideas per group was calculated. Attended subjects made a total of 394 comments across all seven open-ended question. Of those, 216 (55%) were unique, non-overlapping ideas and the remainder were duplications of the same central themes by different subjects. Unattended subjects made a total of 283 comments across all seven open-ended question, and 160 (57%) were unique ideas per group.

Mean number of positive comments per question.

A positive comment praised the usability of any aspect of the user interface presented in the usability test. For example, comments such as “I like dockable because I have the explanation and what is being explained at the same time” were categorized as positive. Positive comments were used to inform the developers about aspects of their software that excelled.

Averaged across all seven open-ended questions, each subject in the attended condition provided 0.8 positive comments ($SD = 0.3$) when answering each question. The unattended subjects provided an average of 0.6 positive comments per question ($SD = 0.2$), which was significantly lower $F(1, 58) = 18.50, p < .01, \eta^2 = .24$.

Mean number of negative comments per question.

A negative comment criticized the usability of any aspect of the user interface presented in the usability test. For example, comments such as “the View Options dialog is too wordy” were categorized as negative. Examples of negative comments included opinions that directly opposed the current product or suggestions for missing items or new features. Negative comments were used to identify areas that required improvement.

Averaged across all seven open-ended questions, each subject in the attended condition provided 0.9 negative comments ($SD = 0.4$) when answering each question. The unattended subjects provided an average of 0.5 negative comments per question ($SD = 0.2$), which was significantly lower $F(1, 58) = 16.37, p < .01, \eta^2 = .22$.

Mean number of irrelevant comments per question.

An irrelevant comment was extraneous with regard to the usability issue under investigation. For example, comments such as “I would like to see how Word handles printing” or “It depends if printing is important to the user or not” were categorized as irrelevant because they did not directly reference a concrete aspect of the user interface under investigation.

Averaged across all seven open-ended questions, each subject in the attended condition provided 0.1 irrelevant comments ($SD = 0.2$) when answering each question. The unattended

subjects provided an average of 0.2 irrelevant comments per question ($SD = 0.2$), which was not significantly different $F(1, 58) = 2.30$, ns , $\eta^2 = .04$.

Mean number of process comments per question.

A comment categorized as process required further action from the researcher because it was not relevant to the usability of Scenario as stated. For example, a comment such as “I have no opinion” would need to be followed up by the researcher through the use of probing to encourage the subject to keep thinking the question through to a conclusion. All questions were designed so that even a person new to Scenario could formulate an opinion. An unattended comment full of spelling and grammar mistakes such as “It tells me what canned algoithms can” would need to be followed up by the researcher to clarify what was stated and to verify the spelling. However, probing and clarification were not possible in the unattended condition because the researcher wasn’t present, resulting in no opportunity to enhance the content of these comments through the use of probing or clarification.

Averaged across all seven open-ended questions, each subject in the attended condition provided 0.02 process comments ($SD = 0.1$) when answering each question. The unattended subjects provided an average of 0.1 process comments per question ($SD = 0.2$), which was significantly higher $F(1, 58) = 6.94$, $p < .05$, $\eta^2 = .11$.

Frequency of Interpersonal Behaviors

The purpose of recording the frequencies of interpersonal behaviors was to obtain a descriptive index of the number of interactions between the researcher and each subject in the attended condition. These interactions were not possible in the unattended condition because the researcher was not present.

It was important to record interpersonal behaviors, a core component of attended usability testing, because they were an index of interactions lost when conducting an identical test unattended. If the survey data from both conditions was similar, then it might be acceptable to lose those interactions and conduct unattended usability tests.

The mean frequencies of interpersonal behaviors between the researcher and each subject in the attended condition, along with descriptive statistics, are presented in Table 8. For example, each subject was probed an average of 5.4 times during the course of usability testing. Two of the eight items in the checklist shown in Appendix C, observe behaviors and provide hints, were omitted because there were no instances of those behaviors. All subjects received one debriefing.

Table 8

Mean Frequency of Interpersonal Behaviors between the Researcher and Each Attended Condition Subject

Statistic	Probe the subject for more information	Clarify the subject's comments	Answer questions from the subject	Be flexible and modify the written testing plan	Encourage the subject
<u>M</u>	5.4	2.7	4.4	0.1	0.1
<u>SD</u>	2.1	2.1	2.4	0.3	0.2
Minimum	0.0	0.0	0.0	0.0	0.0
Maximum	10.0	9.0	11.0	1.0	1.0

Summary of Results

All of the variables tested in the current study and the major findings of each are summarized in Table 9.

Table 9

Experimental Variables Tested and Major Findings

Variable	Finding
Forced-choice questions	
Display preference	Attended = Unattended
Printing preference	Attended = Unattended
Explanation style usefulness	Attended = Unattended
Correct/incorrect responses	Attended = Unattended
Rating scales	Attended $\neq$ Unattended
Open-ended questions	
Mean number of words	Attended = Unattended
Mean number of comments	Attended > Unattended
Mean number of words per comment	Attended < Unattended
Mean number of unique comments per subject	Attended > Unattended
Number of positive comments	Attended > Unattended
Number of negative comments	Attended > Unattended
Number of irrelevant comments	Attended = Unattended
Number of process comments	Attended < Unattended

Discussion

The goal of this study was to determine if attended Internet usability testing data differed from unattended Internet usability testing data. Specifically, three research hypotheses were evaluated. The results supported the hypotheses in some survey questions and not in others. Each hypothesis is presented separately.

Hypothesis 1: Equivalent Answers

It was hypothesized that attended and unattended subjects would be equivalent in their preferences, usefulness rankings, scale ratings, pass/fail decisions, number of correct responses, and mean number of positive and negative comments per user per question. The results demonstrated that these types of usability data were similar whether the testing was attended or unattended, except for positive and negative comments and rating scales. Given the same Web prototype, there was no reason to believe that subjects viewing the prototype alone would answer survey questions differently than subjects viewing the prototype in the presence of the researcher. The survey questions were written carefully so that all of the necessary context was available for the subject to make an informed decision, regardless of the presence or absence of a usability engineer. Consequently, if a usability engineer is willing to limit a usability analysis to predominately forced-choice questions, unattended testing would be an acceptable alternative to traditional attended usability testing.

Positive and Negative Comments

Unattended subjects made less positive and negative comments than attended subjects. This was an important finding: Unattended usability testing data is less positive and negative in tone than attended usability testing data. Perhaps unattended subjects were more hesitant to disclose comments to an unknown and unseen researcher, despite the fact that anonymity was guaranteed. Possibly the promise of anonymity was more credible when stated by a human researcher in the attended condition than when stated via a sentence in the Web prototype in the unattended condition.

Maybe attended subjects assumed that the researcher was interested in positive or negative feedback based on his posture, cues, or manner, and were attempting to please the experimenter by providing more positive or negative comments. For example, the primary goal of usability testing is to identify problems in the form of negative feedback, and improvements to the user interface are based on this feedback. It would be easy for the researcher to cue for negative comments during attended testing and not possible at all in unattended testing. A description on the Web prototype would have to entice unattended subjects to provide more positive and negative comments.

Rating Scales

There were three rating scales on the survey. As shown in Table 7, the rating scales were diverse in their conclusions. First, only one of the three scales supported the hypothesis of equal mean rankings between experimental groups, as shown in Table 7 by the significant F value for question 2C. Second, only two of the three scales supported the hypothesis of equal pass/fail decisions, as shown in Table 7. Question 1C and question 2C were labeled “fail” in both experimental groups, but question 3F was labeled “fail” in the attended condition and “pass” in the unattended condition. In sum, all three rating scales failed to achieve the “pass” criteria in the attended condition. Conversely, two rating scales failed to achieve the “pass” criteria in the unattended condition, but one scale was successful.

The comment data did not provide reasons for the different conclusions across the three rating scales, and it was not possible to determine a general pattern of differences between the two treatment groups based on the results presented in Table 7. Both experimental groups provided similar comments on the same central themes, yet their mean ratings or percentage of rankings neutral and above differed. The implication is to expect different conclusions from attended and unattended subjects based on data from rating scales. Rating scales must be interpreted with caution. Further research should be conducted using a Web prototype incorporating a larger number of rating scales to determine if a general pattern of responding can be identified between attended and unattended subjects.

Hypothesis 2: Quantity of Data

It was hypothesized that the quantity of data obtained from open-ended questions in the attended condition would be higher than the quantity obtained in the unattended condition. This hypothesis was evaluated through the mean number of words per user, the mean number of comments per user, and the mean number of words per comment per user. The hypothesis was supported by the mean number of comments per user. The hypothesis was not supported by the mean number of words per user and the mean number of words per comment per user.

Mean Number Of Words Per User

The two experimental groups were not significantly different in mean number of words per question. In the attended condition, subjects spoke freely with the researcher while the researcher recorded their responses, which minimized the effort on the part of the subject. Conversely, it was postulated that unattended subjects would use less words because they had to type their comments themselves, verify their own spelling, and correct their own grammar. The result would be an increased effort toward providing open-ended comments that was not matched in the attended condition. However, despite this increased effort, it was discovered that unattended subjects provided an equivalent mean number of words per question. Possibly, unattended subjects used an equivalent number of words because they were conscientiously trying to provide enough context so that the researcher would understand their opinions as they were intended. The effort of typing would be for naught if the researcher did not understand their comments. An examination of the comments provided by the unattended subjects confirmed this assumption. In sum, typing their own comments was not a hindrance to the unattended subjects.

Perhaps many of the unattended subjects were trained typists, a common skill in the software industry, and typing answers to open-ended questions did not require any additional effort. It is also common for some software professionals to enroll in communication courses for their professional development, which could improve their spelling, grammar, and ability to complete open-ended questions, again reducing the effort required to complete open-ended questions. To explore these issues further, future research could ask each subject to rate their

typing expertise and identify any communication course qualifications. Both could be considered experience variables.

Mean Number Of Words Per Comment Per User

Unattended subjects provided more words per comment than attended subjects, a finding that was contrary to the original hypothesis. As mentioned above, unattended subjects used an equivalent number of words as the attended subjects, but they provided fewer comments overall, resulting in a higher ratio of words per comment.

Hypothesis 3: Quality of Data

Although the quantity of data in open-ended comments is interesting, it is secondary in usefulness in a usability evaluation to the quality of the comments. Quality is the fundamental issue. It was hypothesized that the quality of data obtained from open-ended questions in the attended condition would be higher than the quality obtained in the unattended condition. This hypothesis was evaluated through the mean number of unique comments per subject, mean number of irrelevant comments per user per question, and mean number of process comments per user per question. The hypothesis was not supported by the mean number of irrelevant comments per user per question.

Irrelevant Comments

Unattended subjects provided the same mean number of irrelevant comments per question as attended subjects, a response direction that was not anticipated. Recall that irrelevant comments are extraneous with regard to the specific usability issue under investigation. In retrospect, irrelevant comments should have been included under hypothesis one rather than hypothesis three. Given the same Web prototype, both experimental groups had the same opportunities to make irrelevant comments.

Unique Ideas Per Group

Attended subjects provided more unique ideas per group than unattended subjects. All usability engineers are interested in maximizing the number of unique ideas obtained from

usability testing. The implication is that future research should be designed to determine how to increase the number of unique ideas provided by unattended usability testing subjects.

Time to Complete Survey

Attended subjects required significantly less time in minutes to complete the survey ($\bar{M}$ = 21.6) than unattended subjects ($\bar{M}$ = 25.9 minutes). The significant difference might have been due to a number of factors. First, the unattended subjects may have spent more time re-reading questions and summaries by navigating forward and back within the prototype. Attended subjects were able to ask the researcher their questions directly, without having to verify information via re-reading or scrolling. Second, unattended subjects may have spent more time correcting the spelling and grammar of their responses. Attended subjects did not have to record their own responses. Third, unattended subjects may have stopped participating in the survey momentarily to answer telephone calls, check their e-mail, or to obtain a cup of coffee. Conversely, all thirty attended subjects gave the researcher their full attention and time during the testing session.

The survey in the current study was optimized through pilot testing of four coworkers so that it could be completed unattended, on average, within 15 minutes. In actual testing, attended and unattended subjects completed the survey in 21.6 and 25.9 minutes, respectively. The attended sessions took approximately 44% longer than planned. Unattended sessions took approximately 73% longer than planned. This is much too long for a Web-based usability test.

In sum, unattended usability testing will take longer than attended usability testing with an equivalent test. The implication is to reduce the time required to take the test as much as possible. A reasonable target would be a pilot test of 10 minutes, which would result in an unattended usability test with a duration of approximately 17 minutes.

Conclusion

Unattended usability testing is a viable alternative to attended usability testing, the traditional approach. As shown in Table 9, there were no major group differences in the experience variables and the forced-choice question variables. However, group differences were the most common finding in the open-ended question variables. These differences are attributed

to the high frequency of probing and clarification in the attended condition, as shown in Table 8. Therefore, the group differences in open-ended questions were caused by the behavior of the researcher, a behavior inherent to the attended testing condition. Conversely, probing and clarification were not necessary in the forced-choice questions because these types of questions were clear enough as presented, resulting in no group differences on forced-choice questions.

Both testing styles yielded similar answers to forced-choice survey questions, but 5-point satisfaction and acceptability rating scales must be interpreted with caution because they produced divergent results. If a usability engineer is able to limit a usability analysis to forced-choice questions, unattended testing would be an acceptable alternative to traditional attended usability testing.

The quantity of comment data collected from unattended usability testing was lower than the quantity obtained from an equivalent attended test. In particular, unattended usability testing yielded a lower mean number of comments per question, a lower mean number of unique comments per subject, and a lower mean number of both positive and negative comments per question. Unattended testing resulted in an equivalent quantity in the form of mean number of words per question and mean number of irrelevant comments per question. Unattended subjects provided a higher quantity in two categories: mean number of process comments per question and mean number of words per comment.

A total of 677 comments were collected from all 60 subjects throughout the seven open-ended survey questions: 393 from the attended condition and 283 from the unattended condition. In spite of the lower comment quantity, 283 comments were enough to be useful and much was learned from the unattended subjects that could be used to increase the usability of Scenario. Both experimental groups discussed similar and overlapping themes and reported similar observations. Unattended usability testing did yield enough useful comment data that unattended testing could be considered a viable, albeit less prolific, alternative to traditional attended usability testing. Nevertheless, if a usability engineer is primarily interested in obtaining a large quantity of comment data, attended usability testing would be the preferred approach.

The quality of comment data collected from unattended usability testing was lower than the quality obtained from attended usability testing. The higher the quality of the data, the greater its potential to increase the usability of a software product. In particular, unattended usability testing yielded less unique ideas per group and a lower mean number of unique comments per subject. Unattended and attended subjects made an equivalent mean number of irrelevant comments. In spite of the predominantly lower comment quality, unattended usability testing did yield enough useful opinions, observations, and unique ideas to justify the process. Nonetheless, if a usability engineer is primarily interested in obtaining higher quality data, attended usability testing would be the favored approach.

To increase the quality of answers to open-ended questions, it might be possible in some circumstances to convert open-ended questions such as "Please explain..." into forced-choice questions with multiple choice answer categories. The challenge is in not limiting the number of possible responses from subjects or to influence their answers.

Another possible solution would be to provide specific instructions within the question itself to provoke discussion. For example, instead of simply asking "Please explain," the question could ask "Please explain why you prefer one display option over another." Alternatively, the question could include an example of how the subjects might possibly respond. This approach would give the subjects an idea of what is expected of them. However, a bias could be introduced because subjects may feel that the only appropriate response is to comment on the example, rather than spontaneously formulate their own opinions. Future research could test the validity of the approaches described above.

Despite the rating scale, negative comment, quantity, and quality limitations of unattended usability testing discovered in the current study, there are several advantages that make a trade-off worthwhile. There would be a time and cost savings of not requiring a usability engineer to be present. More important, the usability test would potentially reach a large number of available users who have expertise in the application area and motivation to improve the product. Finding

skilled users to test a sophisticated application can be difficult if the test must be conducted locally.

The unique advantages of unattended usability testing can be leveraged even further when conducted with Beta users. Beta testing users are “real” users rather than artificial subject sets such as company employees. The intended audience of remote evaluation at Cognos is Beta users. Beta testing is the process of offering potential customers an incomplete but functional product to test how well it suits their needs, prior to purchasing the released product. During the Beta period, users report problems and software “bugs.” Beta testing users are intrinsically motivated to improve the usability of software products because they will be using them to meet their daily job requirements in the future. Beta testing users complete self-determined, job-related tasks with their software, rather than the traditional targeted tasks typically created by usability engineers. Beta testing users have worked with the software product for several weeks prior to answering our surveys.

It is possible that the results from the current study would be different if the subject pool was composed entirely of Beta testing users rather than internal employees of software companies. This could be a topic of future research. For example, a case study was planned to augment the current study. Volunteers from 25 registered Cognos Beta sites currently testing Scenario in their workplace were contacted and an equivalent Web prototype was made available. Unfortunately, none of the Beta users responded. These potential subjects were highly motivated Beta testers but they shunned the Web prototype despite an offer of incentive. It is possible that the potential population of tens of millions of Web users may be difficult to attract.

To increase response rates, perhaps a Web prototype could be made available to the entire Internet audience, a subject pool of millions of people. However, given the response rate from the registered Beta sites, further research would be necessary to determine the usefulness of such a proposition.

This study was undertaken to determine if attended Internet usability testing data would differ from unattended Internet usability testing data. The results show that unattended Internet

usability testing is a valid and valuable alternative to traditional attended testing. However, attended and unattended testing both provide different results in certain cases. Choosing one method over another will depend on the types of questions being asked, the acceptability of inherent trade-offs, and the feedback that is required by the software's developers.

References

- Alreck, P. L., & Settle, R. B. (1985). The survey research handbook. Homewood, IL: Irwin.
- Bly, S. (1997). Field work: Is it product work? Interactions, 5 (1), 25-30.
- Connell, J. L., & Shafer, L. (1989). Structured rapid prototyping: An evolutionary approach to software development. Englewood Cliffs, NJ: Prentice-Hall.
- Dumas, J. S., & Redish, J. C. (1994). A practical guide to usability testing. Norwood, NJ: Ablex.
- Elgin, B. (1995). How can networked users provide their own usability feedback? Subjective usability feedback from the field over a network: A CHI '95 SIG. SIGCHI Bulletin, 27 (4), 43-44.
- Hammontree, M., Weiler, P., & Nayak, N. (1994). Remote usability testing. Interactions, 2 (7), 21-25.
- Hartson, H.R., Castillo, J.C., Kelso, J., & Neale, W.C. (1996). Remote evaluation: The network as an extension of the usability laboratory. In Proceedings of the CHI '96, 1996 (Vancouver, BC, Canada, April 13-18, 1996). (pp. 228-235). New York: Association for Computing Machinery.
- Lahti, E., & Kim, K. (1997). Obtaining usability feedback from Beta sites: making your products "Beta and Beta". In The Art of Seeing Conference Proceedings, 1997 (Monterey, California, 12 August - 15 August, 1997). (pp. 477-479). Dallas, TX: Usability Professionals' Association.
- Keppel, G. (1991). Design and analysis: A researcher's handbook. New Jersey: Prentice-Hall.
- Mitropoulos-Rundus, D., & Muszak, J. (1997a). Criteria for determining if consumer "in-home" usability testing is feasible for your product. Common Ground, 7 (1), 10-12.
- Mitropoulos-Rundus, D., & Muszak, J. (1997b). How to design and conduct a consumer in-home usability test. Common Ground, 7 (2), 1-19.
- Monk, A., Wright, P., Haber, J., & Davenport, L. (1993). Improving your human-computer interface: A practical technique. New York: Prentice-Hall.
- Morkes, J. & Nielsen, J. (1997, November 7). Concise, scannable, and objective: How to write for the Web [On-line]. Available: <http://www.useit.com/papers/webwriting/writing.html>

- Nielsen, J. (1990). Paper versus computer implementations as mockup scenarios for heuristic evaluation. In Proceedings of Interact '90. (pp. 315-320). New York: Association for Computing Machinery.
- Nielsen, J. (1993). Usability engineering. San Diego: Academic Press.
- Preece, J. (1994). Human-computer interaction. Wokingham, England: Addison-Wesley.
- Rubin, J. (1994). Handbook of usability testing: How to plan, design, and conduct effective tests. New York: Wiley.
- Spool, J. M. (1997). Web site usability: A designer's guide. North Andover, MA: User Interface Engineering.
- Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In J. Nielsen & R. L. Mack (Eds.), Usability inspection methods (chap. 5). New York: Wiley.
- Wiklund, M. E. (1994). Usability in practice: How companies develop user-friendly products. Cambridge, MA: AP Professional.

Appendix A

Announcement For Recruiting

Each potential subject was e-mailed the following text message:

Good Morning! This is Steve Macko calling from the UI group on the 4th floor...

We are currently conducting a brief usability test on Scenario Beta 2.0 using a new style of Internet usability testing as part of a joint research project with Carleton University. We were hoping that you could spare about 15 minutes on one of the following days at your convenience:

- Monday
- Thursday
- Friday

Please let me know what time and day is best for your schedule (e.g. Thursday at 10AM), and provide directions to your desk. Hope to see you soon!
Steve

Appendix B

Written Introduction/Informed Consent

Attended Condition

Research Personnel

Steve Macko (Principal Investigator, 738-1338 Ext. 3244)

Dr. R. F. Dillon (Faculty Advisor, 520-6629)

Ethical questions or concerns:

Dr. Mary Gick (Ethics Chair, 520-2600 Ext. 2664)

Dr. Kim Matheson (Department Chair, 520-2600 Ext. 2648)

Introduction

Today we will be conducting a brief usability test on a new feature from Scenario Beta 2.0- the Explain Tab. However, a *unique* aspect of this usability test is that we are going to use the Internet to administer the test itself, which is what Cognos would like to do in the future with its Beta users around the world!

We will test half of the participants with me present. The other half will be tested without me here at all, which would simulate how Beta users would be taking the test. In addition, Beta users around the world will also be administered an identical usability test and questionnaire via the Internet (experiment 2), and these results will be compared to those from experiment 1.

All feedback will be kept confidential and anonymous. Your name will not be associated with your answers. If you wish to cease participation in this usability test at any time, please do so.

I will remain present during the session. To help me transcribe accurate information, I would like your permission to audio tape your comments. You can ask me to clarify anything at any time, but I can not answer any of the specific questions on the questionnaire for you. Do you have any final questions before we begin?

Participant's Name: _____ Date: _____
Signature: _____

Unattended Condition

Research Personnel

Steve Macko (Principal Investigator, 738-1338 Ext. 3244)

Dr. R. F. Dillon (Faculty Advisor, 520-6629)

Ethical questions or concerns:

Dr. Mary Gick (Ethics Chair, 520-2600 Ext. 2664)

Dr. Kim Matheson (Department Chair, 520-2600 Ext. 2648)

Introduction

Today we will be conducting a brief usability test on a new feature from Scenario Beta 2.0- the Explain Tab. However, a *unique* aspect of this usability test is that we are going to use the Internet to administer the test itself, which is what Cognos would like to do in the future with its Beta users around the world!

We will test half of the participants with me present, and the other half without me here at all, which would simulate how Beta users would be taking the test. In addition, Beta users around the world will also be administered an identical usability test and questionnaire via the Internet (experiment 2), and these results will be compared to those from experiment 1.

All feedback will be kept confidential and anonymous. Your name will not be associated with your answers. If you wish to cease participation in this usability test at any time, please do so.

I am going to leave you once you begin the test. When you are finished, please click the "Submit" button. Do you have any final questions before we begin?

Participant's Name: _____ Date: _____
Signature: _____

URL: http://www.cognos.com/stage/scenario_explain_survey/intro.htm

Appendix C

Interpersonal Behavior Frequency Checklist

Core Behavior	Frequency
Observe Behaviors	
Probe (ask the user for more information)	
Clarify (help the user when they appear confused)	
Answer Questions (when asked by the subject)	
Provide Hints	
Flexibility	
Encourage	
Debrief	

Appendix D

Written Debriefing

The following written debriefing was e-mailed to all subjects on the day of their participation:

Research Personnel

Steve Macko (Principal Investigator, 738-1338 Ext. 3244)

Dr. R. F. Dillon (Faculty Advisor, 520-6629)

Ethical questions or concerns:

Dr. Mary Gick (Ethics Chair, 520-2600 Ext. 2664)

Dr. Kim Matheson (Department Chair, 520-2600 Ext. 2648)

Debriefing

Thank you for participating in the Scenario 2.0 Beta survey. The usability test employed a new style of Internet usability testing as part of a joint research project with Carleton University. In experiment 1, half of the participants will be tested in the presence of a usability engineer, and the other half will be tested in the absence of a usability engineer. The results from both testing conditions will be compared using statistical procedures to determine if any significant differences exist.

In addition, Beta users around the world will also be administered an identical usability test and questionnaire via the Internet (experiment 2). These results will be compared to those from experiment 1.

If you would like to see a copy of the results of this study, please email me directly. I will send them out as soon as they become available (approximately April 1998).

Steve Macko (Steve.Macko@Cognos.Com)
Usability Engineer/Human Factors Research
738-1338 Ext. 3244

Appendix E

Web Prototype/QuestionnaireAttended Condition**Introduction- Attended Condition**

The **purpose** of this survey is to analyze the usability of the new Explain feature in Scenario:

The screenshot shows a web application window titled "Scenario - [GO-II.ana: Target = Profit]". The interface includes a toolbar with icons for various functions. On the left, there is a "Path" window showing a tree structure with "All Data" (100 rows, 860 sales) and "Sale Price: Less than 0.00" (2 rows, 304 sales). Below this is a "Factors" window listing various factors with their values: Sale Price (54), Cost of Goods Sold (48), Date (36), Advertising Cost (29), Channel (17), Month of Year (11), Month of Quarter (8), and Average Sales (5). The main area displays the "Explain" feature for the selected segment. It shows the title "Target = Profit" and "Sale Price: Less than 0.00". Below this, it states "The current segment contains rows with: Sale Price: Less than 0.00". A "Summary of Profit" table compares the current segment with all data. Below the table, it says "Above and beside this segment:" followed by a list of bullet points explaining the segment's characteristics.

	Current	All Data
average	-304	860
median	-301	827
size	8	480

Above and beside this segment:

- The parent segment was split by *Sale Price* into 7 segments.
- The impact of *Sale Price* at the parent segment is 54, which is high.¹
- The 8 rows in this segment represent a relatively small proportion of the 480 rows of the parent segment.²
- This segment contains distinct values of *Profit*.³
- There is an increasing trend. That is, *Profit* increases as *Sale Price* increases.

At the bottom of the window, there is a status bar showing "Overall Fit: 54%".

The Explanation is a new view of your data that provides a summary and interpretation of your data **segments** (groups of data that are similar in both factor and target values) in the Graph and Tree Tabs. As you navigate through the segments in the Graph or Tree, the Explanation changes to reflect the selected segment.

In all examples:

- The ToolBar, Status Bar, Path and Factors window have been turned off to conserve space
- the sample data file *GO-II.ana* has been loaded
- the target is Profit (a continuous variable)

We need your input on 3 usability issues. All feedback will be kept confidential and anonymous. Your name will not be associated with your answers.

Begin Survey Now

Usability Issue 1: Explain Display

There are two ways to display Explain information:

- in a Tab display just like the Tree and Graph Tabs
- in a dockable window

You set the type of display from the **View** menu...

Select Options...

The current segment shows with:
Sale Price: Less than 0.00

Summary of segment data:

	Current	All Data
average	-304	860
median	-301	827
size	8	480

Above and beside this segment:

- The parent segment was split by *Sale Price* into 7 segments.
- The impact of *Sale Price* at the parent segment is 54, which is high.¹
- The 8 rows in this segment represent a relatively small proportion of the 480 rows of the parent segment.²
- This segment contains distinct values of *Profit*.³
- There is an increasing trend. That is, *Profit* increases as *Sale Price* increases.

Below this segment:

- The segment has not been drilled down.

This dialog determines the Explanation's display location.

If you select the "Tab display" option...

...you would see the Explanation as a Tab along with Graph, Tree, Bookmarks, and Data. As a Tab,

- you *can not* see the explanation and any other Tab at the same time unless you select Window | New Window and Window | Tile
- you *can* print the Explanation or copy it to the clipboard

Scenario - [GO-II.ana: Target = Profit]

Graph Tree Bookmarks Data **Explanation**

Target = Profit
Sale Price: Less than 0.00

The current segment contains rows with:
 Sale Price: Less than 0.00

Summary of *Profit*

	<u>Current</u>	<u>All Data</u>
average	-304	860
median	-301	827
size	8	480

Above and beside this segment:

- The parent segment was split by *Sale Price* into 7 segments.
- The impact of *Sale Price* at the parent segment is 54, which is high.¹
- The 8 rows in this segment represent a relatively small proportion of the 480 rows of the parent segment.²
- This segment contains distinct values of *Profit*.³
- There is an increasing trend. That is, *Profit* increases as *Sale Price* increases.

Below this segment:

- The segment has not been drilled down.

If you select the "dockable window display" option...

You would see the Explanation as a standard window **docked** at the right edge of the main Scenario window. When docked,

- you *can* see both the Explanation and the other Tabs
- you *can not* print the Explanation or copy it to the clipboard

You can also drag the window yourself and dock it in 3 other locations: Left, Bottom, and Top:

Or, if you double-click on the window's title bar, you can make an **undocked** or "floating" window:

The screenshot shows a software interface with a main window titled "Scenario - [GO-II.ana: Target = Profit]". The window has a menu bar with "File", "View", "Segment", "Analysis", "Tools", "Window", and "Help". Below the menu bar are tabs for "Explain", "Tree", "Histogram", and "Data". The main area displays "Average Profit" and "All Data". A table with 8 rows is visible, with a selection box around the first row. An "Explain" dialog box is open, showing details about the current segment.

Explain

The current segment contains rows with:
Sale Price: Less than 0.00

Summary of Profit

	Current	All Data
average	-304	860
median	-301	827
size	8	480

Above and beside this segment:

- The parent segment was split by *Sale Price* into 7 segments.
- The impact of *Sale Price* at the parent segment is 54, which is high.
- The 8 rows in this segment represent a relatively small

At the bottom of the dialog box, there are two buttons: "< Previous" and "Next >".

Summary:

	View Explanation and data at same time	Print and copy Explanation to clipboard
Explanation as a Tab	No	Yes
Explanation as an independent, dockable window	Yes	No

Usability Issue 2: Explanation Style

You can also set the level of detail included in the Explanation. There are three levels. Here is the **View Options** dialog again.

The **Brief** style presents Explanations in a compact list format, with no supplemental information or advice...

Scenario - [G0-II.ana: Target = Profit]

File View Search Analyze Tools Window Help

Graph Tree Bookmark Data Explain

Target = Profit
Sale Price: Less than 0.00

The current segment contains rows with:
 Sale Price: Less than 0.00

Summary of *Profit*

	<u>Current</u>	<u>All Data</u>
average	-304	860
median	-301	827
size	8	480

Above and beside this segment:

- The parent segment was split by *Sale Price* into 7 segments.
- The impact of *Sale Price* at the parent segment is 54, which is high.
- The 8 rows in this segment represent a relatively small proportion of the 480 rows of the parent segment.
- This segment contains distinct values of *Profit*.
- There is an increasing trend. That is, *Profit* increases as *Sale Price* increases.

Below this segment:

- The segment has no factors.

< Previous Next >

The **Synopsis** style presents Explanations in a compact list format, with supplemental information and advice offered as footnotes...

Scenario - [GO-II ana: Target = Profit]

File Edit View Segment Analysis Help

Graph Tree Bookmarks Data Editor

Summary of Profit

	<u>Current</u>	<u>All Data</u>
average	-304	860
median	-301	827
size	8	480

Above and beside this segment:

- The parent segment was split by *Sale Price* into 7 segments.
- The impact of *Sale Price* at the parent segment is 54, which is high.¹
- The 8 rows in this segment represent a relatively small proportion of the 480 rows of the parent segment.²
- This segment contains distinct values of *Profit*.³
- There is an increasing trend. That is, *Profit* increases as *Sale Price* increases.

Below this segment:

- The segment has no factors.

Footnotes ←

¹ A high impact means that a factor explains a lot of the variation in the target.

² Predictions are more reliable when based on large data sets.

³ The shaded quartiles box of this segment does not overlap with other segments' in the Graph Tab. Predictions are more reliable when the values are distinct.

The **Report** style presents Explanations in an easily-readable format, with supplemental information and advice included in the body of the text...

Scenario - [GO-11.ana: Target = Profit]

File Edit View Search Analysis Tools Window Help

Graph Tree Bookmarks Data Editor

	<u>Current</u>	<u>All Data</u>
average	-304	860
median	-301	827
size	8	480

Above and beside this segment
 The parent segment was split by *Sale Price* into 7 segments.

The impact of *Sale Price* at the parent segment is 54, which is high. A high impact means that a factor explains a lot of the variation in the target.

The 8 rows in this segment represent a relatively small proportion of the 480 rows of the parent segment. Predictions are more reliable when based on large data sets.

This segment contains distinct values of *Profit*. The shaded quartiles box of this segment does not overlap with other segments' in the Graph Tab. Predictions are more reliable when the values are distinct.

There is an increasing trend. That is, *Profit* increases as *Sale Price* increases.

Below this segment
 The segment has no factors.

Summary: There are 3 Explanation styles:

- **Brief:** a compact list format, with no supplemental information or advice
- **Synopsis:** a compact list format, with supplemental information and advice offered as footnotes
- **Report:** an easily-readable format, with supplemental information and advice included in the body of the text

Usability Issue 3: Explanation Interpretation

Now we want you to actually **use** the Explanation information. Here are the Synopsis-style Explanations for two different segments:

- Sale Price: 2500.00 to 10500.00
- Sale Price: 10500.00 or greater

All other screen elements have been removed for clarity...

5/20/2012

5/20/2012

The current segment contains rows with:
Sale Price: 2500.00 to 10500.00

Summary of Profit

	Current	All Data
average	1120	860
median	1219	827
size	230	480

Observations about this segment:

- There is one low outlier.¹
- The values of Profit slightly skewed.

Above and beside this segment:

- The parent segment was split by Sale Price into 7 segments.
- The impact of Sale Price at the parent segment is 54, which is high.²
- The 230 rows in this segment represent a relatively large proportion of the 480 rows of the parent segment.³
- This segment does not contain distinct values of Profit.
- There is an increasing trend. That is, Profit increases as Sale Price increases.

Below this segment:

- There are 8 significant factors at this segment, compared with 13 columns included in the analysis.
- This segment is further split by Date into 15 child segments.

Footnotes

- Outliers may represent either unusual data or invalid records. Verify using the Data Tab.
- A high impact means that a factor explains a lot of the variation in the target.
- Predictions are more reliable when based on large data sets.
- The shaded quartiles box of this segment overlaps other segments' in the Graph Tab. Predictions are more reliable when the values are distinct.

The current segment contains rows with:
Sale Price: 10500.00 or greater

Summary of Profit

	Current	All Data
average	2033	860
median	1237	827
size	19	480

Observations about this segment:

- There are 4 low outliers, and 4 high outliers.¹
- The average lies above the shaded quartiles box.²

Above and beside this segment:

- The parent segment was split by Sale Price into 7 segments.
- The impact of Sale Price at the parent segment is 54, which is high.²
- The 19 rows in this segment represent a relatively small proportion of the 480 rows of the parent segment.³
- This segment does not contain distinct values of Profit.
- There is an increasing trend. That is, Profit increases as Sale Price increases.

Below this segment:

- The segment has no factors.

Footnotes

- Outliers may represent either unusual data or invalid records. Verify using the Data Tab.
- This may be due to the low outliers. Consider using the median as a measure of typical instead of the average.
- A high impact means that a factor explains a lot of the variation in the target.
- Predictions are more reliable when based on large data sets.
- The shaded quartiles box of this segment overlaps other segments' in the Graph Tab. Predictions are more reliable when the values are distinct.

< Previous

Unattended Condition**Introduction- Unattended Condition**

The **purpose** of this survey is to analyze the usability of the new Explain feature in Scenario:

The Explanation is a new view of your data that provides a summary and interpretation of your data **segments** (groups of data that are similar in both factor and target values) in the Graph and Tree Tabs. As you navigate through the segments in the Graph or Tree, the Explanation changes to reflect the selected segment.

In all examples:

- The ToolBar, Status Bar, Path and Factors window have been turned off to conserve space
- the sample data file *GO-II.ana* has been loaded
- the target is Profit (a continuous variable)

We need your input on 3 usability issues. All feedback will be kept confidential and anonymous. Your name will not be associated with your answers.

Begin Survey Now

Please enter the current time now: e.g. 1:35 PM

Usability Issue 1: Explain Display

There are two ways to display Explain information:

- in a Tab display just like the Tree and Graph Tabs
- in a dockable window

You set the type of display from the **View** menu...

Select Options...

Scenario - [GO-II.ana: Target = Profit]

View Segment Analysis Tools Window Help

Graph Data Edit

Target = Profit
Sale Price: Less than 0.00

The current segment shows with:
Sale Price: Less than 0.00

Options...

Summary of segment

	Current	All Data
average	-304	860
median	-301	827
size	8	480

Above and beside this segment:

- The parent segment was split by *Sale Price* into 7 segments.
- The impact of *Sale Price* at the parent segment is 54, which is high.¹
- The 8 rows in this segment represent a relatively small proportion of the 480 rows of the parent segment.²
- This segment contains distinct values of *Profit*.³
- There is an increasing trend. That is, *Profit* increases as *Sale Price* increases.

Below this segment:

- The segment has not been drilled down.

<Previous >Next>

This dialog determines the Explanation's display location.

If you select the "Tab display" option...

...you would see the Explanation as a Tab along with Graph, Tree, Bookmarks, and Data. As a Tab,

- you *can not* see the explanation and any other Tab at the same time unless you select Window | New Window and Window | Tile
- you *can* print the Explanation or copy it to the clipboard

Scenario - [GO-IL ana: Target = Profit]

Graph Tree Bookmarks Data **Explanation**

Target = Profit
Sale Price: Less than 0.00

The current segment contains rows with:
 Sale Price: Less than 0.00

Summary of *Profit*

	<u>Current</u>	<u>All Data</u>
average	-304	860
median	-301	827
size	8	480

Above and beside this segment:

- The parent segment was split by *Sale Price* into 7 segments.
- The impact of *Sale Price* at the parent segment is 54, which is high.¹
- The 8 rows in this segment represent a relatively small proportion of the 480 rows of the parent segment.²
- This segment contains distinct values of *Profit*.³
- There is an increasing trend. That is, *Profit* increases as *Sale Price* increases.

Below this segment:

- The segment has not been drilled down.

<Previous Next>

If you select the "dockable window display" option...

You would see the Explanation as a standard window **docked** at the right edge of the main Scenario window. When docked,

- you *can* see both the Explanation and the other Tabs
- you *can not* print the Explanation or copy it to the clipboard

You can also drag the window yourself and dock it in 3 other locations: Left, Bottom, and Top:

Or, if you double-click on the window's title bar, you can make an **undocked** or "floating" window:

Summary:

	View Explanation and data at same time	Print and copy Explanation to clipboard
Explanation as a Tab	No	Yes
Explanation as an independent, dockable window	Yes	No

1A. Which type of display do you prefer?

- ☐ Tab display
☐ Dockable window

1B. Please explain.

<div></div>	<div></div>
<div></div>	<div></div>

[< Previous](#) [Next >](#)

1C. The Explain Tab is the only Tab that can become a dockable window. The Graph, Tree, Bookmarks, and Data Tabs can only be displayed as "true" Tabs. How **acceptable** is it to you that the Explain Tab is *different*?

(Please click the radio button that corresponds to your opinion)

Very Unacceptable 1	Slightly Unacceptable 2	Neutral 3	Slightly Acceptable 4	Very Acceptable 5
☐	☐	☐	☐	☐

1D. Please explain.

☐ Previous
 ☐ Next

1E. Would you want to be able to print from **both** the Tab and dockable window displays?

- ☐ Yes
☐ No

1F. Please explain.

☐ Previous
 ☐ Next

Usability Issue 2: Explanation Style

You can also set the level of detail included in the Explanation. There are three levels. Here is the **View Options** dialog again.

The Brief style presents Explanations in a compact list format, with no supplemental information or advice...

Scenario - [GO-II.ana: Target = Profit]

File Edit View Search Analysis Tools Window Help

Graph Trees Bookmarks Data Editor

Target = Profit
Sale Price: Less than 0.00

The current segment contains rows with:
 Sale Price: Less than 0.00

Summary of *Profit*

	<u>Current</u>	<u>All Data</u>
average	-304	860
median	-301	827
size	8	480

Above and beside this segment:

- The parent segment was split by *Sale Price* into 7 segments.
- The impact of *Sale Price* at the parent segment is 54, which is high.
- The 8 rows in this segment represent a relatively small proportion of the 480 rows of the parent segment.
- This segment contains distinct values of *Profit*.
- There is an increasing trend. That is, *Profit* increases as *Sale Price* increases.

Below this segment:

- The segment has no factors.

Previous Next

The **Synopsis** style presents Explanations in a compact list format, with supplemental information and advice offered as footnotes...

Scenario - [60-II ana: Target = Profit]

File Edit View Segment Project Info Window Help

Graph Tree Bookmark Data Export

Summary of Profit

	<u>Current</u>	<u>All Data</u>
average	-304	860
median	-301	827
size	8	480

Above and beside this segment:

- The parent segment was split by *Sale Price* into 7 segments.
- The impact of *Sale Price* at the parent segment is 54, which is high.¹
- The 8 rows in this segment represent a relatively small proportion of the 480 rows of the parent segment.²
- This segment contains distinct values of *Profit*.³
- There is an increasing trend. That is, *Profit* increases as *Sale Price* increases.

Below this segment:

- The segment has no factors.

Footnotes ←

¹ A high impact means that a factor explains a lot of the variation in the target.

² Predictions are more reliable when based on large data sets.

³ The shaded quartiles box of this segment does not overlap with other segments' in the Graph Tab. Predictions are more reliable when the values are distinct.

The **Report** style presents Explanations in an easily-readable format, with supplemental information and advice included in the body of the text...

Scenario - [GO-II.ana: Target = Profit]

File Edit View Settings Analysis Data Window Help

Graphs Data Bookmarks Data Export

	<u>Current</u>	<u>All Data</u>
average	-304	860
median	-301	827
size	8	480

Above and beside this segment
 The parent segment was split by *Sale Price* into 7 segments.

The impact of *Sale Price* at the parent segment is 54, which is high. A high impact means that a factor explains a lot of the variation in the target.

The 8 rows in this segment represent a relatively small proportion of the 480 rows of the parent segment. Predictions are more reliable when based on large data sets.

This segment contains distinct values of *Profit*. The shaded quartiles box of this segment does not overlap with other segments' in the Graph Tab. Predictions are more reliable when the values are distinct.

There is an increasing trend. That is, *Profit* increases as *Sale Price* increases.

Below this segment
 The segment has no factors.

Summary: There are 3 Explanation styles:

- **Brief:** a compact list format, with no supplemental information or advice
- **Synopsis:** a compact list format, with supplemental information and advice offered as footnotes
- **Report:** an easily-readable format, with supplemental information and advice included in the body of the text

2A. How would you **rank** each of the 3 Explanation styles in order of usefulness to your individual needs?

(Please type a "1" for most useful, "2" for second most useful, and "3" for third most useful.)

- ☐ Brief
- ☐ Synopsis
- ☐ Report

2B. Comments:

[< Previous](#) [Next >](#)

2C. How **acceptable** would it be to you if there was only *one* Explanation style rather than three?

Very Unacceptable 1	Slightly Unacceptable 2	Neutral 3	Slightly Acceptable 4	Very Acceptable 5
☐	☐	☐	☐	☐

2D. Please explain.

[< Previous](#) [Next >](#)

Usability Issue 3: Explanation Interpretation

Now we want you to actually **use** the Explanation information. Here are the Synopsis-style Explanations for two different segments:

- Sale Price: 2500.00 to 10500.00
- Sale Price: 10500.00 or greater

All other screen elements have been removed for clarity...

[< Previous](#) [Next >](#)

3A. Using the information contained within the two Explanations, which of the two segments is **skewed**?

- ☐ Sale Price: 2500.00 to 10500.00
- ☐ Sale Price: 10500.00 or greater
- ☐ Neither segment is skewed

3B. Which of the two segments has the **most** outliers?

- ☐ Sale Price: 2500.00 to 10500.00
- ☐ Sale Price: 10500.00 or greater
- ☐ Neither segment has outliers

3C. How many outliers does your choice of segments from 3B have?

3D. What do outliers represent?

[< Previous](#)

[Next >](#)

[Show Segments Again](#)

3E. Is there a **trend** in the data?

- ☐ No
☐ Yes: Increasing
☐ Yes: Decreasing

3F. How **satisfied** are you with the "interpretive" or "analyst in a box" aspect of the Explanation (e.g. Did the Explanation identify trend in a satisfactory manner)?

Very Unsatisfied	Slightly Unsatisfied	Neutral	Slightly Satisfied	Very Satisfied
1	2	3	4	5
☐	☐	☐	☐	☐

3G. Please explain.

[< Previous](#) [Next >](#) [Show Segments Again...](#)

3H. Is there any information **missing** from the Explanation that you require?

[< Previous](#) [Next >](#) [Show Segments Again...](#)

Demographics

4A. How would you rate your **knowledge** of statistics (e.g. means, medians, modes, ranges, interpreting charts, etc.)?

(Please choose the number that best applies)

No Knowledge 1	2	3	4	5	6	7	8	9	I am a Statistician 10
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

4B. How would you rate your **expertise** with Scenario (e.g. what percentage of features are you competent with)?

(Please choose the number that best applies)

No Expertise 1	2	3	4	5	6	7	8	9	I am a Power User 10
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

<Previous	Next>
------------------------------	--------------------------

4C. How would you rate your **expertise** with browsing the Internet (e.g. I can navigate through most Web pages)?

(Please choose the number that best applies)

No Expertise 1	2	3	4	5	6	7	8	9	I am a Power Browser 10
☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Please submit your answers to Cognos by clicking this button:

Thank you for your time!

Appendix F

Attended Condition Questionnaire

Browser and version: _____

Monitor Resolution: _____

Start Time: _____ (page2.htm loaded)

Usability Issue 1: Explain Display1A. Which type of display do you **prefer**?☐ Tab display☐ Dockable window

1B. Please explain.

1C. The Explain Tab is the only Tab that can become a dockable window. The Graph, Tree, Bookmarks, and Data Tabs can only be displayed as “true” Tabs.

(Show diagram)

How **acceptable** is it to you that the Explain Tab is *different*?

Very Unacceptable	Slightly Unacceptable	Neutral	Slightly Acceptable	Very Acceptable
1	2	3	4	5

1D. Please explain.

1E. Would you want to be able to print from **both** the Tab and dockable window displays?☐ Yes☐ No

1F. Please explain.

Usability Issue 2: Explanation Style

2A. How would you **rank** each of the 3 Explanation styles in order of usefulness to your individual needs, where a "1" is most useful, "2" is second most useful, and "3" is third most useful?

_____ Brief
_____ Synopsis
_____ Report

2B. Comments:

2C. How **acceptable** would it be to you if there was only *one* Explanation style rather than three?

Very Unacceptable	Slightly Unacceptable	Neutral	Slightly Acceptable	Very Acceptable
1	2	3	4	5

2D. Please explain.

Usability Issue 3: Explanation Interpretation

3A. Using the information contained within the two Explanations, which of the two segments is **skewed**?

- ☐ Sale Price: 2500.00 to 10500.00
- ☐ Sale Price: 10500.00 or greater
- ☐ Neither segment is skewed

3B. Which of the two segments has the most **outliers**?

- ☐ Sale Price: 2500.00 to 10500.00
- ☐ Sale Price: 10500.00 or greater
- ☐ Neither segment has outliers

3C. How **many** outliers does your choice of segments from 3B have? _____

3D. What do outliers **represent**?

3E. Is there a **trend** in the data?

- ☐ No
- ☐ Yes: Increasing
- ☐ Yes: Decreasing

3F. How **satisfied** are you with the “interpretive” or “analyst in a box” aspect of the Explanation (e.g. Did the Explanation identify trend in a satisfactory manner)?

Very Unsatisfied	Slightly Unsatisfied	Neutral	Slightly Satisfied	Very Satisfied
1	2	3	4	5

3G. Please explain.

3H. Is there any information **missing** from the Explanation that you require?

Demographics

4A. How would you rate your **knowledge** of statistics (e.g. means, medians, modes, ranges, interpreting charts, etc.)?

*(Please choose the number that **best** applies)*

4B. How would you rate your **expertise** with Scenario (e.g. what percentage of features are you competent with)?

*(Please choose the number that **best** applies)*

4C. How would you rate your **expertise** with browsing the Internet (e.g. I can navigate through most Web pages)?

*(Please choose the number that **best** applies)*

End Time: _____

Thank you for your time!

Appendix G

Attended Condition Materials

The usability test is accessed from the following location:

<http://gibraltar/users/mackos/scenario/intro2.htm>

Figure 1:

Scale 1:

Very Unacceptable	Slightly Unacceptable	Neutral	Slightly Acceptable	Very Acceptable
1	2	3	4	5

Scale 2:

Very Unsatisfied	Slightly Unsatisfied	Neutral	Slightly Satisfied	Very Satisfied
1	2	3	4	5

Scale 3:

Scale 4:

Scale 5:

IMAGE EVALUATION TEST TARGET (QA-3)

APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989

© 1993, Applied Image, Inc., All Rights Reserved