Investigating the Effectiveness of Applying the Critical Incident Technique to Remote Usability Evaluation

Jennifer A. Thompson

Thesis submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Master of Science

In

Industrial and Systems Engineering

Dr. Robert C. Williges, Chair

Dr. Rosson

Dr. Brian M. Kleiner

December 1, 1999
Blacksburg, Virginia

Keywords: Remote Usability Evaluation, Critical Incident Technique

Investigating the Effectiveness of Applying the Critical Incident Technique to Remote Usability Evaluation

Jennifer A. Thompson

(ABSTRACT)

Remote usability evaluation is a usability evaluation method (UEM) where the experimenter, performing observation and analysis, is separated in space and/or time from the user. There are several approaches by which to implement remote evaluation, limited only by the availability of supporting technology. One such implementation method is RECITE (the REmote Critical Incident TEchnique), an adaptation of the user-reported critical incident technique developed by Castillo (1997). This technique requires that trained users, working in their normal work environment, identify and report critical incidents. Critical incidents are interactions with a system feature that prove to be particularly easy or difficult, leading to extremely good or extremely poor performance. Critical incident reports are submitted to the experimenter using an on-line reporting tool, who is responsible for their compilation into a list of usability problems. Support for this approach to remote evaluation has been reported (Hartson, H.R., Castillo, J.C., Kelso, J., and Neale, W.C., 1996; Castillo, 1997).

The purpose of this study was to quantitatively assess the effectiveness of RECITE with respect to traditional, laboratory-based applications of the critical incident technique. A 3x2x5 mixed-factor experimental design was used to compare the frequency and severity ratings of critical incidents reported by remote versus laboratory-based users. Frequency was measured according to the number of critical incident reports submitted and severity was rated along four dimensions: task frequency, impact on task performance, impact on satisfaction, and error severity. This study also compared critical incident data reported by trained users versus by usability experts observing end-users. Finally, changes in critical incident data reported over time were evaluated.

In total, 365 critical incident reports were submitted, containing 117 unique usability problems and 50 usability success descriptions. Critical incidents were classified using the Usability Problem Inspector (UPI). A higher number of web-based critical incidents occurred during Planning than expected. The distribution of voice-based critical incidents differed among participant groups, with users reporting a greater than expected number of Planning incidents and experts reporting fewer than expected Assessment incidents. Usability expert performance was not correlated, requiring that separate analyses be conducted for each set of expert data.

Support for the effectiveness in applying critical incidents to remote usability was demonstrated, with all research hypotheses at least partially supported. Usability experts gave significantly different ratings of impact on task performance than did user reporters. Remote user performance versus laboratory-based users failed to reveal differences in all but one measure: laboratory-based users reported more positive critical incidents for the voice interface than did remote users. In general, the number of negative critical incidents decreased over time; a similar result did not apply to the number of positive critical incidents.

It was concluded that RECITE is an effective means of capturing problem-oriented data over time. Recommendations for its use as a formative evaluation method applied during the latter stages of product development (i.e. when a high fidelity prototype is available) are made. Opportunities for future research are identified.

ACKNOWLEDGEMENTS

I would like to extend my gratitude to Dr. Robert C. Williges for his continued support and guidance. Your enthusiasm for my research topic was invaluable to the completion of this thesis. I would also like to thank Dr. Brian M. Kleiner for sharing his expertise and time and Dr. Rosson for accommodating a last minute change.

So much of this work would not have been possible without the support and friendship of Greg Edwards, Eric Nash, Erik Olsen, and Lisa Cooper. I would especially like to extend my gratitude to Jason Saleem whose love and encouragement helped keep me in the eye of the hurricane.

I would also like to thank my friends and family back in Canada for their continued love, concern, and support.

This thesis is dedicated to my grandfather, Jack Everett, whose passion for engineering and the pursuit of wisdom became my own.

TABLE OF CONTENTS

LIST OF TABLES

TABLE OF FIGURES

CHAPTER 1. INTRODUCTION

The usability of a human-computer interface is a multi-dimensional attribute, comprised of such elements as affect, effectiveness, learnability, efficiency, and control. An evaluation method capable of addressing each of these elements requires careful design, particularly to obtain results that are complete, valid, and reliable. A single method for conducting all usability evaluations does not exist, but rather depends on the problem domain of interest, the resources available, and the objectives of the evaluation.

Formal evaluation is one of the most common usability evaluation approaches. According to this approach, a representative sample of users is recruited to perform specific tasks in a laboratory environment. Videotape may be used to capture audio and visual data for future analysis. A variety of quantitative and qualitative measures are used, from which usability problems and successes are identified (in the case of a formative evaluation) and/or the degree to which usability specifications have been met is verified (in the case of a summative evaluation).

Formal evaluation has often been criticized on account of the lack of generality of its results. Laboratory conditions are rarely representative of actual work conditions and the effort to match the two environmental conditions can be very difficult or impossible (Hartson, Castillo, Kelso, and Neale, 1996). Furthermore, this method of evaluation demands access to a large enough sample of representative users, all of whom must be able to come to the laboratory for the full set of usability sessions. This requirement becomes problematic when users are scattered or inaccessible (due to small numbers or lack of resources) or when test sessions must extend over a period of several days or weeks.

One means of overcoming these problems is remote evaluation. Remote evaluation is distinguished from other forms of evaluation by the fact that the evaluator or experimenter is separated in space and/or time

from the user (Hartson et al., 1996). Data is collected from the user located in his or her own work environment, thereby ensuring external validity. Relative to the user, the remote data collection process can be conducted either passively or actively. The former of these methods occurs when the user is unaffected by the collection process, as in remote inspection. Active data collection requires that the user participates in, and perhaps even directs, the data collection process. Although active data collection may disrupt task flow, it provides a direct means of soliciting usability-related information from the user. Interpretation of actions via observation or videotape analysis, and the potential biases associated with these methods, are avoided.

One method of actively collecting data from a remote environment is the critical incident technique. A critical incident is defined as an interaction with a feature or element of a system that is either particularly easy or difficult, resulting in extremely good or extremely poor performance (del Galdo, Williges, Williges, and Wixon, 1987). The technique involves the identification and reporting of critical incidents that take place relative to the general aim of the observed activity. From these reports, a list of usability problems and successes can be compiled.

The critical incident technique was proposed by Flanagan (1954) to permit systematic observations to be made in a field environment. For that reason, it makes a natural fit within the realm of remote evaluation techniques. Another advantage is that the technique was designed to be flexible, and so can be applied to a variety of problem domains. Furthermore, while Flanagan suggested the use of expert observers to carry out data collection (i.e. a passive approach), studies by del Galdo et al. (1987) and Castillo (1997) have been successful in allocating that role to the user.

While the critical incident technique is not itself a new concept, its application to the usability evaluation of human-computer interfaces has only recently been demonstrated as useful and effective (del Galdo et

al., 1987). Today, much laboratory-based formative evaluation is conducted using the critical incident technique on account of its ability to gather a rich source of usability evaluation with a minimum of time and effort (Carroll, Koenemann-Belliveau, Rosson, and Singley, 1993).

Even more recent is the application of the technique as an approach to remote evaluation. Pilot studies conducted by Castillo (1997) and Hartson et al. (1996) have provided support for this application. However, many outstanding issues remain unresolved with respect to how to best adapt the technique to remote evaluation of human-computer interfaces (see Castillo, 1997). Moreover, there is no present empirical evidence supporting its effective use in an actual remote setting. It was thus of interest to help resolve these outstanding issues, while broadening the application area to which the critical incident technique has been utilized.

Shattuck and Woods (1994) report that there exist several misunderstandings regarding the critical incident technique. This section attempts to eliminate these misunderstandings by presenting a review of the history of the technique and the research efforts that have contributed to its development. Also presented are two new domains into which this study aims to extend the critical incident technique: remote evaluation and usability evaluation of voice interfaces. The current state of each of these domains is described and the potential contributing role of the critical incident technique proposed.

1.1.1 Origins

Fitts and Jones (1947) first codified the critical incident technique in their analysis and classification of pilot error experiences in reading and interpreting aircraft instruments (although the term “error” rather than critical incident was used). The technique was formally described in a paper published 45 years ago by John Flanagan (1954). The underlying motivation for publishing this paper was to specify a method of systematically collecting observations from the field that was useful in “solving practical problems and developing broad psychological principles”. The observations of interest were critical incidents.

According to Flanagan, a critical incident is an observable and sufficiently complete human act or behavior whose consequences are such that 1) an observer can readily determine its effects and 2) it is either extremely effective or ineffective at attaining a particular objective.

Flanagan based the critical incident technique on a set of principles, rather than procedures, so to facilitate its modification to suit various problem domains. These principles stipulate initial collaboration with domain experts so as to define the general aim of the activity being studied. Trained observers are then to systematically collect data according to some predefined set of rules. Any behaviors interpreted as extremely effective or ineffective in the context of the general aim (i.e. critical incidents) are to be recorded directly or reported in subsequent interviews. Critical incidents are then to be compiled and analyzed so as to increase their usefulness. Typically, this involves an inductive process whereby the critical incidents are grouped into areas and sub-areas according to some general frame of reference. Finally, the principles stipulate that the experimenter interpret and report the results in light of biases that may have occurred during any one of the preceding stages.

1.1.2 Reliability and Validation

Following its introduction, the critical incident technique was frequently used in job analyses. Its reliability and validity, however, were not confirmed until the work of Andersson and Nilsson (1964) nearly a decade later. This study was conducted by collecting critical incidents pertaining to the behavior of store managers. Data was obtained from several different groups of people using a variety of interview techniques. A questionnaire was also distributed to solicit additional critical incidents from three of the four groups. All critical incidents were classified into areas, categories, and subcategories, the accuracy of which was verified by asking university students to recategorize the incidents. Further analyses confirmed the comprehensiveness of the behaviors addressed and indicated that only a relatively small number of incidents were required to define the majority of relevant categories. The type and number of critical incidents reported were only slightly affected by having different interviewers report

the critical incidents. The number, but not the type, of critical incidents was affected by the data collection method selected. Based on these results, Andersson and Nilsson (1964) concluded that the method was both reliable and valid.

1.1.3 Transformation Into a Usability Evaluation Tool

Despite the now pervasive awareness of the critical incident method, Shattuck and Woods (1994) report that the technique has been rarely used in the form in which Flanagan originally described. Rather, the technique has been modified to accommodate the needs and constraints of a particular study. Facilitating these modifications was the flexibility inherent to the technique (as was anticipated and encouraged by Flanagan).

In recent years there has been an increasing interest in the study of systems involving people and machines engaged in goal-directed activities, and in the usability of the interfaces there within (Shattuck and Woods, 1994). This trend has resulted in a shift away from behaviorally oriented research (that in which the critical incident technique was rooted) towards one that focused on cognitive processes. A rethinking, as opposed to simple modification, of the critical incident technique was in order.

Formalizing this effort was the work of del Galdo, Williges, Williges and Wixon (1987) in their development of a method to evaluate software documentation and aid in its subsequent redesign. The critical incident method was selected as a basis for collecting data on account of its ability to be introduced iteratively, as well as its focus on users and their immediate reaction to software problems during interaction. However, before the critical incident technique could be applied to the problem domain of interest, it had to first undergo several fundamental modifications.

The first of these modifications was a refinement of the concept of a critical incident. Recall that Flanagan (1954) defined a critical incident as extreme *behavior* that is either outstandingly effective or

ineffective with respect to the aim of an activity. In contrast, del Galdo et al. (1987) treated the critical incident as an *interaction* with a feature or element of the system that proved to be particularly easy or difficult. In other words, the notion of a critical incident was expanded to encompass the combined and interdependent behaviors of the interface and its user. Additional modifications were made to the way in which the critical incidents were collected. For example, the people experiencing the critical incidents (i.e. the users) were allocated the task of identifying and reporting critical incidents rather than trained observers. Users were required to describe the critical incident *and* rate its criticality (with respect to task performance) through the use of an on-line reporting tool. Reporting was done immediately after the critical incident took place or while the critical incident was in progress, rather than in the subsequent interviews proposed by Flanagan. At the completion of the reporting process, a usability expert was responsible for compiling the critical incident reports and translating them into usability problems and successes. A ranking scheme based on frequency and criticality ratings was used to prioritize the problems and successes, which helped direct redesign efforts.

Del Galdo et al. (1987) implemented this revised critical incident technique twice to allow for two design iterations. A reduction in the number of critical incidents and their criticality ratings from one iteration to the next was found and used as a measure of success of the revised technique. Thus, del Galdo et al. (1987) demonstrated the effectiveness of modifying the critical incident method to obtain qualitative data pertaining to human-computer interaction.

Expanding upon these modifications were Koenemann et al. (1994) and Carroll et al. (1993). This group proposed that a critical incident infrequently occurs as an isolated event. Rather it is more likely to be a manifestation of a series of earlier events, each of which is unremarkable, and hence undetectable. They named this series of events a “critical thread”. In order to capture the events comprising a critical thread,

data prior to the critical incident must be reviewed and claims analysis used to extract commonality from identified events. The critical incident can then be more fully and accurately explained.

To support this proposal, an evaluation case study was conducted based on observational data of Smalltalk™ learners using two different training materials. The occurrence of a critical incidence unique to one user was identified and selected as a basis of comparison. Data recorded prior to this event was reviewed for both participants. Claims analysis was subsequently implemented to delineate the critical thread. The outcome of the case study was that, while demanding and laborious, the critical thread approach is more systematic. The richer source of data that it provides also ensures that all aspects of a usability problem are identified and addressed, leading to a more successful redesign.

The successes of changes implemented by del Galdo et al. (1986) and Koenemann-Belliveau et al. (1994) comprise the stepping-stones along which the critical incident was brought to human-computer interface usability evaluation. The work of Hartson et al. (1996) and Castillo (1997) has now indicated a potential successful extension of the technique to the realm of remote evaluation. Before this research is reviewed the concept of remote evaluation will be addressed.

1.2 REMOTE EVALUATION

Usability evaluations employing the critical incident technique have typically been conducted in a laboratory. In this environment, the experimenter has control over such elements as noise, lighting, and system configuration. Furthermore, a trained expert (i.e. the experimenter or delegated usability expert) is able to directly observe the participant, in adherence to Flanagan's principles. The limitation of this environment is that it fails to capture usage patterns pertinent to actual work environments. Moreover, it does not address situations in which the users of interest are distributed and remote.

Distributed and international customer bases are becoming more common. For example, Hammontree, Weiler, and Nayak (1994) report that several U.S.-based Hewlett Packard development teams are producing software for user groups located in Europe, Australia, and Asia-Pacific. In this past, this trend would have represented a major stumbling block for human factors practitioners wishing to conduct usability evaluations with representative users. Recent advances in the areas of information sharing and collaborative work tools have provided a feasible solution: the use of computer networks, and particularly the Internet, as a framework upon which to construct a usability lab and the use of work stations connected to these networks as "windows" into a remote user site (Hammontree et al., 1994). This approach is called remote evaluation.

1.2.1 Definition

Remote evaluation is defined to be "usability evaluation wherein the evaluator, performing observation and analysis, is separated in space and/or time from the user" (Hartson et al., 1996). Figure 1-1 illustrates this concept.

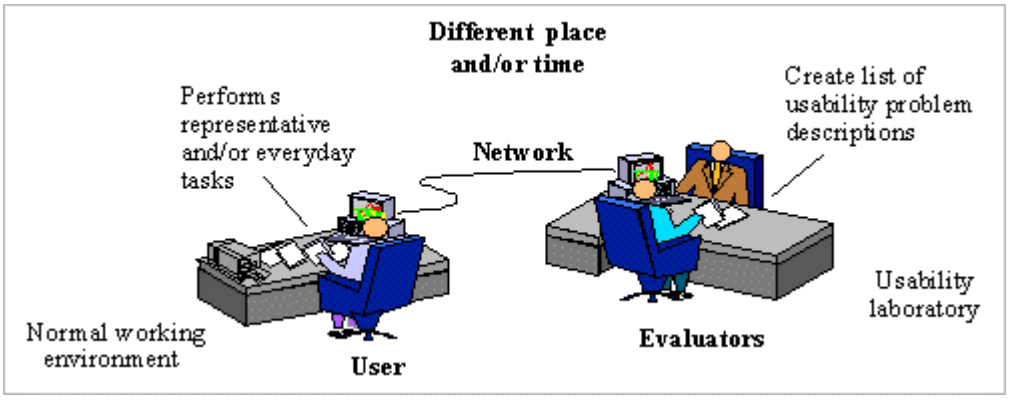

Figure 1-1: Remote Evaluation (Castillo, 1998)

In remote evaluation it is important to define a common frame of reference. In this study, location is specified with respect to the evaluator, who is local. The user, working in his or her working environment that is separate from that of the evaluator, is remote.

Another important clarification to make is that while remote evaluation strives to capture data with the user working in his or her normal environment, it is not the same as usability testing in the field. Field testing implies that the evaluator and user share the same environment (usually that of the user). The risk inherent to field testing is that the presence of the evaluator can modify user behavior through the act of observation. Remote testing minimizes this risk by eliminating the physical presence of the evaluator, although the user behavior may still be affected if he or she is required to participate in the data collection process.

1.2.2 Supporting Technology

Remote evaluation would not be feasible without the increasingly availability of supporting technology. Required of this technology is that it permits some form of collaboration between the experimenter(s) and user(s), regardless of their separation in time and/or space. Meeting these requirements are window/application sharing tools, whiteboard applications, and computer-based video conferencing.

Tools that support sharing of windows and applications in real-time operate via a network or modem links such that two or more computers are linked. Provided sharing tool software is active on all participating computers, users can view a shared window or application as if it were actually running on each of their respective machines (Hammontree, Weiler, and Nayak, 1994).

Similar in concept is the shared whiteboard, which allows multiple users to simultaneously view, point at, and annotate, the contents of a common drawing or writing surface (Hammontree et al., 1994). Typical whiteboard contents include snapshots of documents, diagrams, or in the case of a usability evaluation, prototype screens.

Window/application and whiteboard sharing tools allow users to view a common computer screen, but not each other. Live video of one or several people can be displayed as windows on the computers of another person or group of people by means of computer-based videoconferencing tools. The advantage of this technology is that it permits the conveyance of facial gestures and body movements, from which valuable information can be ascertained, and a more natural and engaging form of communication can be supported.

1.2.3 Approaches

Many approaches to remote evaluation are well suited to the three forms of technologies currently available. A list of such approaches, as identified by Hartson et al. (1996) and Hammontree et al. (1994), is given in Table 1. Each approach imposes a different level of involvement on the part of the user. For instance, the user's role may be passive (i.e. unaffected by the data collection process) or active (i.e. the user participates in data collection). This classification scheme is applied to the approaches listed in Table 1-1.

Table 1-1: Approaches to Remote Evaluation

Remote Usability Approach	Type Data Collection	User Role in Data Collection Process
Portable Usability Evaluation	A portable usability evaluation kit is taken to the users in their normal working environment.	Active
Remote Inspection	Clients use the network to deliver design documents, software samples, and/or prototypes to the remote evaluators (typically commercial services).	Not applicable – usability evaluation is typically conducted in the form of heuristic evaluation or intuitive inspection. Design guidelines, user profiles, and software standards may or may not be used.
Remote Questionnaire/ Survey	A questionnaire intended to elicit user preference data is built into the software application being evaluated. A particular event or sequence of events triggers its appearance. Responses are sent to the developers via the internet. The usefulness of data is limited due to the lack of qualitative data.	Active
Remote-control Evaluation	Evaluator obtains control of the user's computer by means of an Internet connection and software such as Timbuktu™. Video capture is achieved by means of a video camera and scan converter; audio capture is made via computer or telephone. Data-capture may be ongoing or may be triggered based on usage of a particular application.	Passive (without audio capture) Active (with audio capture)
Video Conferencing	Similar to local, laboratory-based usability evaluation, except that user and evaluator are not in adjacent rooms but connected using the network and video conferencing software.	Passive (with the exception of requiring the user to set-up the videoconference).
Instrumented Remote Evaluation	Application to be tested is coded to automatically record data pertaining to user actions (ex. keystrokes and mouse movements). Data is compiled into journals or logs, which are later analyzed to determine if and when usability problems occurred. Swallow, Hameluck, and Carey (1998) have worked on the development of problem indicator criteria to facilitate in automatic (and accurate) problem detection.	Passive
Semi-instrumented Remote Evaluation	Users are trained to identify and report specific usage events while interacting with application being evaluated. Reports are transmitted to developers, along with context information (ex. system state, task, interface history and state), and usability problems are identified.	Active
Remote Thinking Aloud	Video and audio links are established between users and evaluators. Users perform representative tasks while concurrently describing their actions and impressions aloud while the evaluators observe.	Active
Collaborative Walkthroughs	The evaluator remotely assists the user methodically step through a design storyboard based on a predefined series of linear paths that characterize the task domain of interest.	Active

1.2.4 Assessment of Remote Evaluation Approaches

As the technology for supporting remote evaluation becomes increasingly more sophisticated and cost-effective, the impetus for validating the effectiveness of the various approaches has taken on increasing importance. Efforts to achieve this goal have recently been initiated.

For instance, the remote thinking-aloud approach has been implemented at Sun and Hewlett-Packard sites with successful results (Hammontree et al., 1994). Several factors contributed to the success of this approach. For example, the establishment of a live video link between users and evaluators not only helped build a rapport with the users, but also encouraged a more active and engaging discussion. Another key factor was the dynamic presentation of instructions, which helped avoid potential response biases resulting from reading the scenarios beforehand. Finally, the capture of screenshots illustrating problematic features or interactions was shown to be a useful means by which to elicit design suggestions from the user (Hammontree et al., 1994).

Hammontree et al. (1994) also report successful implementation of the collaborative walkthrough. Attributed to this success was the use of shared whiteboard software, which allows users to directly input their problems or design suggestions during the evaluation session. Scenarios helped define a framework from which user feedback could be more easily elicited.

The aforementioned successes were observed in the field and lack empirical support. Hartson et al. (1996) conducted two case studies, the purpose of which was to use empirical methods to assess the validity of videoconferencing and the semi-instrumented critical incident gathering. Specifically, the objective was to examine the “degree to which the quality and quantity of data of data collection were affected by implementing these new approaches” (Hartson et al., 1996).

Videoconferencing was evaluated by having participants answer a questionnaire and complete benchmark tests using a WWW browser while located in the lab or remote from the lab. An evaluator recorded usability problems in both conditions, with remote data made available from digital video sent via the Internet and voice transmitted via a telephone connection. Time and navigation errors were logged automatically. Results of this case study indicated a lack of significant differences in questionnaire data and in the number of usability problems recorded between conditions. These results provided support for the feasibility of videoconferencing as a usability evaluation method.

A second case study was conducted to determine if similar results were applicable to the semi-instrumented remote evaluation. In this study, users were engaged in the identification of critical incidents and the quality of their data compared with that collected in a laboratory-based usability evaluation. Critical incidents were defined as occurrences during system usage reflecting usability problems, missing functionality, or other ways in which the system fails to meet user needs. The rationale for involving users is that they are the only ones to have knowledge of critical incidents as they occur. They can therefore provide a very important source of qualitative data from which usability problems can be isolated. Of interest was to assess the ability of users to identify critical incidents in comparison to expert observers and the ability of evaluators to transform reported critical incident data into real usability problems. del Galdo et al. (1987) pursued similar research objectives in their evaluation of critical incidents reporting by users in a laboratory environment.

The method by which Hartson et al. (1996) implemented the semi-instrumented remote evaluation technique built upon findings of previous critical incident feasibility studies (i.e. del Galdo et al., 1987 and Koenemann et al., 1994). For example, to facilitate the reporting of critical incidents by remote users performing normal working tasks, Hartson et al. (1996) added a "Report Incident" button to the application being evaluated. This button was visible and accessible during all phases of interaction.

When activated, an on-line reporting tool similar to that developed by del Galdo et al. (1987) was accessed. The purpose of this tool was to extract information from the user regarding the critical incident.

Also required was that the critical incident data be amenable to quick and easy conversion into usability problems. To meet this requirement, the concept of a critical thread (Koenemann et al., 1994) was built into the functionality of the "Report Incident" button. Specifically, the button was designed to trigger an instrumentation routine that captured a screen-sequence video clip showing 60 seconds of screen activity surrounding the critical incident. This video clip could be combined with the critical incident description to create a contextualized critical incident (CCI). The CCI could then be sent to the evaluator via the Internet for review. This process is illustrated in Figure 1-2.

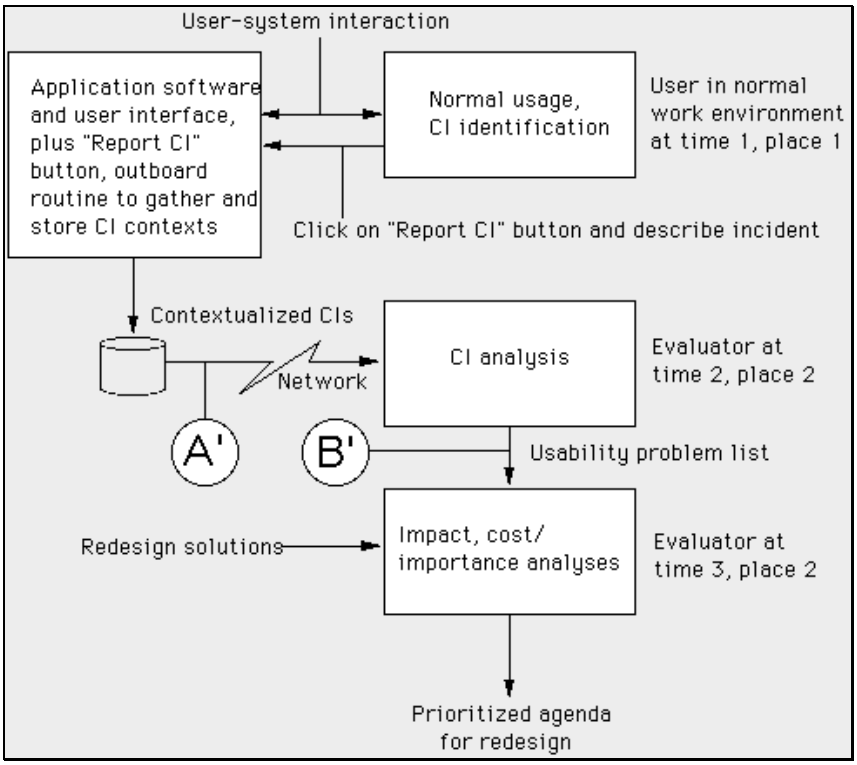

Figure 1-2: Semi-instrumented Remote Evaluation Implemented via the Critical Incident Technique (Hartson et al., 1996)

In the case study, Hartson et al (1996) used two kinds of subjects: user subjects and expert subjects. User subjects were non-usability experts who received training to identify critical incidents while performing a series of WWW browser-related tasks. The actual means by which the critical incidents were reported was simulated. User subjects were required to click on the space bar rather than a "Report Incident" button and to verbalize a description of the critical incident rather than interacting with an on-line reporting tool. Both audio and video data were recorded on videotape. Usability experts were recruited to review the videotapes and identify any critical incidents not identified by the user subjects. Combining user-reported critical incident descriptions and video clips created contextualized critical incidents. These were sent to a second set of usability expert for transformation into one or more usability problems. The success of this transformation process was evaluated.

Informal observations indicated that user subjects were equally as capable of identifying critical incidents than expert observers, with the exception of lower severity incidents (which they tended to ignore). The ability of expert users to identify usability problems associated with CCIs was impeded by a lack of knowledge regarding task information (ex. intended task and task context). The amount of data needed to capture sufficient information without incurring high network transmission costs was identified as an issue requiring further investigation.

The significance of the Hartson et al. (1996) case study was not only that it provided further support for remote evaluation, but also that it represented a first attempt to integrate the critical incident technique into the area of remote evaluation. The result was the development of a new type of remote evaluation method called the user-reported critical incident method.

1.3 THE USER-REPORTED CRITICAL INCIDENT TECHNIQUE FOR REMOTE EVALUATION

In general, the user-reported critical incident technique is a usability evaluation method for capturing critical incidents that satisfies the following criteria:

- Data are centered on critical incidents that occur during task performance.
- Real users perform tasks.
- Users are located in normal working environments.
- Data are captured on a daily basis and in a cost-effective way.
- Direct interaction between user and evaluator is not needed.
- Data are high quality and hence easily converted into usability problems.

Castillo (1997) developed a method of implementing the user-reported critical incident technique by refining the semi-instrumented remote evaluation technique implemented and evaluated by Hartson et al. (1996). Several refinements were made, as described below.

1.3.1 On-Line Critical Incident Reporting

The first refinement made was to develop an on-line critical incident reporting tool that could be implemented versus just simulated. This reporting tool was comprised of a “Report Incident” button that opened a textual form in a separate window from the application being evaluated. This form was comprised of a series of carefully structured questions designed to capture various contextual factors typically associated with a critical incident. A list of these questions, their associated contextual factors, and the format in which they were presented is provided in Table 1-2.

Table 1-2: Contents of the Critical Incident Reporting Form

Question	Contextual Factor(s)	Format
Explain what you were trying to do when the critical incident occurred	Specific beginning of the task	Text box
Describe what you expected the system to do just before the critical incident occurred.	User expectations	Text box
In as much detail as possible, describe the critical incident that occurred and why you think it happened.	Detailed description of the critical incident	Text box
Describe what you did to get out of the critical incident.	Ability to recover and specific ending of a critical incident	Text box
Where you able to recover from the critical incident?	Ability to recover	Yes/No
Are you able to reproduce the critical incident and make it happen again?	Critical incident reproducibility	Yes/No
Indicate in your opinion the severity of this critical incident.	Critical Incident Severity/Task Frequency	5-point Rating Scale

Activation of the "Report Incident" button prompted the user to respond to each of the questions listed in Table 1-2 and then submit the report by pressing on a Submit button. A cancel feature was also added in the event that the user did not want the report form submitted.

Castillo (1997) also designed additional windows to enhance the critical incident identification and reporting process. A Welcome Window, for example, was designed to provide both general information and instructions relevant to the evaluation study as a whole. This screen was followed by a Critical Incident Instructions Window. As its name implies, the purpose of this window was to provide information about how to identify and report a critical incident, and as such served as an implicit reminder to participants of their role in the usability evaluation.

1.3.2 User Training

A second major refinement implemented by Castillo was a more structured way of training user subjects. Training was considered to be an integral component on account of its hypothesized ability to increase the effectiveness of the user as a critical incident reporter.

Castillo adopted a minimalist approach in the design of his training program. Developed by Caroll and his associates (Carroll, Mack, Lewis, Grischkowski, and Robertson (1985); Carroll, Smith-Kerker, Ford, and Mazur-Rimetz (1987)) this approach advocates that people prefer active learning or learning by doing rather than by reading a manual (Widenbeck, Zila, and McConnell, 1995). The following features characterize this approach: a focus on real tasks, reduction in the verbiage of training materials, and support for error recovery and recognition. Effectiveness of the minimalist has experimental support (Carroll et al. (1985); Carroll et al. (1987)).

Conducted in advance to groups of participants, Castillo's training program began with a 20-minute video presentation. The purpose of this video was to demonstrate a user identifying critical incidents while performing several tasks, such as:

- deleting a file from a personal document retrieval system;
- formatting of a diskette in DOS using a Macintosh; and
- counting game penalties using a Web-based counter.

In each case, a narrator provided an explanation as to why a particular critical incident was considered as such, what its severity was, and what were possible means by which it could be resolved. The videotape concluded with a final usage scenario that was stopped by the experimenter before an explanation of the critical incident was given. The participants were then asked to derive their own explanation.

Participants were then given a five-minute review of how to identify *and* report critical incidents, and then were given an opportunity to practice while performing a representative task using the software being evaluated. Hands-on training, while not an explicit component of the minimalist training, fits well within the scope of this approach (Wiedenbeck et al., 1995).

A case study was conducted to ascertain the effectiveness of the training program. Only half of the participants were exposed to the videotape presentation, while all participants participated in the follow-up review and practice sessions. The performance of these two groups was compared. It was found that participants reported a similar amount of critical incidents, the quality of which was approximately equivalent. These findings seem to indicate that the videotape presentation provided no additional training benefit to the participants. However, many of participants (75%) who were exposed to the video presentation expressed that they felt better prepared to identify critical incidents.

What may also have been critical to the success of the training program was the hands-on practice session. Charney and Reder (1986), for example, have shown that methods involving problem-solving practice are superior to merely reading or typing worked-out examples. A formal evaluation of the effectiveness of the practice exercises was not conducted.

1.3.3 Qualitative Feasibility Study

A case study was implemented by Castillo (1997) to assess the feasibility of the new method with respect to the user's ability to identify and report critical incidents and to the expert's ability to transform this data into usability problems. Ways in which the method could be further refined were also investigated.

A sample of user subjects was selected and trained for the purpose of this case study. The nature of this training varied: half of the subjects were exposed to both the videotape and practice session (plus review) while the other half were exposed to the practice session only. All user subjects were then assigned a series of tasks to perform using a token experimental application (i.e. the Internet Movie Database). Mandated was that they complete all tasks by a certain date, but were at liberty to complete all tasks in one or multiple sittings. Video cameras and a scan converter were used to collect visual and audio data of the participant and screen clips, respectively. The experimenter reviewed this data and selected six

reports from different user subjects that demonstrated completeness and accuracy. A CCI was created for each of these selected reports by adding a 3-minute screen capture and video clip. CCI packages were then distributed for evaluation to four different evaluator subjects for evaluation, two of which received only the critical incident reports (and not the video clips). All subjects completed questionnaires.

Results showed that remote users are capable of reporting high severity incidents encountered during task performance, as well as low and medium severity incidents. This capability was not significantly affected by differences in training. Also revealed was that evaluator subjects were capable of analyzing critical incident data to produce usability problem descriptions. The number of usability problems identified was not significant affected by the ability of videotape clips, although clip and report evaluators required less analysis time than report only evaluators.

Despite the refinements that Castillo (1997) integrated into his version of the user-reported critical incident technique, the results obtained made it clear that further development and refinement is required. For example, a lack of granularity in the severity rating scale suggests the development of a more comprehensive and reliable severity rating scheme, perhaps in the form of smaller and more easily judged 6-point scale ratings (to avoid overuse of a middle rating). Considerable variation in the timing of report submission with respect to critical incident occurrence was shown to exacerbate the process of automatically capturing meaningful screen data. This finding suggests that supplementing critical incident descriptions with additional data may not be worth the additional cost and effort. Other suggested refinements included a more seamless integration of the critical incident reporting tool with the application being evaluated and a more flexible means of reporting critical incidents. Added flexibility would include the ability to send quick questions or comments and a means of separating the act of identifying the occurrence of an incident from the act of reporting it. In conclusion, the work of Castillo

(1997) provided a foundation from which further development and refinement of the remote user-reported critical incident technique can be pursued.

1.4 VOICE INTERFACES

In the development of a usability evaluation method, it is important to assess whether and how usability problems vary with the type of interface (Gray and Salzman, 1998). To date, the remote critical incident technique has been applied only within the context of computer interfaces, such as graphical user interfaces and web interfaces. Its effectiveness in the evaluation of other types of interfaces (ex. speech, auditory, and haptic) has not been explored.

Voice interfaces are one example of an alternate interface modality to which the application of the critical incident could be of benefit. A voice interface provides output in the form of natural or synthesized speech. Telephone applications are one example for which voice is the only feasible output modality. Voice interfaces can also be integrated into computer interfaces as a means of providing an alternate mode of interaction (ex. to support hands-free interaction).

1.4.1 Voice Interface Technology

The characteristics of a voice interface depend on its underlying technology. The voice-processing field encompasses five broad areas of technology: 1) voice coding, 2) voice synthesis; 3) speaker recognition, 4) speech recognition, and 5) spoken language translation (Doe, 1998). Of particular interest is speech synthesis, which allows for the dynamic generation of speech, and voice recognition. When implemented into a single application, these technologies together support speaking and listening activities. As such they can be used to mimic human-human dialogue.

1.4.1.1 Speech Recognition

The ability to listen and understand what is being said is the underlying objective of speech recognition technology. More specifically, speech recognition is the process of extracting information in

a voice signal so as to identify a speaker and/or control the actions of a computer (Doe, 1998). Human speech recognition is the model computer-based speech recognition is designed.

Automatic speech recognition (ASR) is the mapping of a continuous-time signal to a series of discrete entities (ex. phonemes). The general process by which this mapping is achieved consists of three components:

1. A structural model that encapsulates knowledge of language structure, speech production, and speech perception (Makhoul and Schwartz, 1994).
2. A statistical variability model that accounts for known variabilities.
3. The synthesis of the speech signal.

The most widely used structural model is the hidden Markov model (HMM). HMMs are finite-state machines corresponding to phonetic contexts (different acoustic realizations of the same phoneme), which use two transitions between states to quickly search through a database (Doe, 1998). For each transition two probabilities are provided: the probability of going to the next state and the conditional probability that a word is correct. The HMM is estimated automatically with the aid of the speech text and the phonetic spellings of the words. If the ASR device is speaker-dependent or adaptive, HMM estimation is also achieved with the aid of training speech (text for which words have been transcribed and placed into a lexicon with appropriate phonetic spellings).

The type of speaker from whom voice input is accepted classifies speech recognition devices. For example, there are speaker-trained, speaker-adaptive, or speaker-independent devices. Devices can also be classified according to the type of voice input they will accept, ranging from isolated word/small vocabularies to continuous/large vocabularies. The current trend is towards the latter, as it most closely resembles natural conversational input. Impeding this progress is the large variability that can exist in the speech signal. Typical sources of variability are presented in Table 1-3.

Table 1-3: Sources of Variability in the Speech Signal

Source of Variation	Causes of Variation
Linguistic Variability	• phonetics, phonology, syntax, semantics
Intra- or Interspeaker Variability	• attributes of speaker: dialect, gender, age • manner of speaking: breath and lip noise, stress, rate, level, pitch, cooperativeness
Channel Variability	• additive noise: stationary and non-stationary • speech-correlated noise: reverberation and reflection • input equipment: microphone, filter transmission system, and recording equipment

Despite this variation, advances in ASR have taken place. Makhoul and Schwartz (1994) report that word error rates have been reduced by more than a factor of five and recognition speeds increased by several orders of magnitude. Facilitating these advances are faster recognition search algorithms and more powerful computers.

1.4.1.2 Speech Synthesis

Speech or voice synthesis is the conversion of textual information to speech using sets of rules for converting letters to phonemes and for converting phonemes to acoustic events (Kamm, 1994). In general, it is the process of transforming speech from text in such as way as to approximate how a human would read those same words (Schmandt, 1994). Synthesized speech that can emulate the quality of human speech is difficult to achieve due to the irregularity inherent to the orthography of the English language (i.e. the mapping from letters to sound are not one to one).

Irregularities require that speech synthesis be a two-step process. In the first step, text is converted to a less ambiguous representation: a string of phonemes. Phonemes are units of speech that together define all the sounds from which all words can be constructed. This process can be approached in two ways. One approach is to construct a pronunciation lexicon in which the appropriate phonemes are found by looking up the appropriate word. Alternatively, knowledge of spelling rules can be used to derive pronunciation from the text (Schmandt, 1994). In either case, it may be necessary to preprocess or normalize the text (ex. to convert symbols or abbreviations to their full form) and apply morphological

analysis. Consideration of such factors as lexical stress, coarticulation, intonation, and prosody is also critical to ensure that the appropriate sound is generated.

Sound generation is the second step of speech synthesis. Again there are two approaches by which this can be accomplished. The first approach, called parametric synthesis, continuously varies the parameters controlling a digital voice tract. The second approach, called concatenative synthesis, is discrete-based; that is, it generates speech by piecing together small segments of digitized speech.

Of importance to the user is the quality of the speech output. The intelligibility and naturalness of the speech can affect not only the user's subjective evaluation of the speech, but may also have implications with respect to performance (i.e. it may impose a greater cognitive load on the user than its lexical counterpart).

1.4.2 Design Considerations

The combination of speech recognition and speech synthesis technology into a voice interface enables that interface to support the two key components to human-human dialogue: speaking and listening. Human-human dialogue is a complex activity: it involves applying many layers of knowledge and sophisticated protocols (Schmandt, 1994). It is also characterized by certain behavioral habits, such as speaking in a continuous manner, using extraneous speech and hesitation sounds (ex. "uhh"), anticipating responses, and speaking at the same time as another talker (known as "talk-over") (Kamm, 1994). Intonation can also be used to denote meaning: falling intonation may be used to imply acknowledgement while rising intonation may signal uncertainty and/or potential error (Karis and Dobroth, 1994). These types of behavioral habits are difficult to change and hence must be supported if the interface is to be useable.

The factors upon which the usability of a voice interface depends are inherently different than those pertaining to a graphical user interface or web interface. Kamm (1994) has identified three factors that must be addressed if a useable interface is to be achieved: 1) the information requirements of the task and interdependencies of task demands; 2) the limitations and capabilities of the voice technology; and 3) expectations, expertise, and preferences of the user. The interface must also be designed to account for the temporal nature of speech output and its concomitant demands on user memory. Design strategies by which to address these factors have been developed (see Kamm, 1994). The effectiveness of applying these design strategies can be verified by means of usability evaluation.

1.4.3 Applications

The range of applications to which voice interfaces can be applied is expansive. As the underlying technology becomes more sophisticated and powerful, and increasing number of applications are becoming viable. A description of some of the more prevalent applications is provided below.

1.4.3.1 Screen Readers

The application that initiated the bulk of the original work in speech synthesis was the development of a reading machine for the blind (Schmandt, 1994). Now commonly referred to as screen readers, these “machines” are comprised of both software and hardware components (ex. sound card device and voice synthesizer) that together produce audio representations of text output from other programs (Richards, 1997). Specifically, filtering techniques embedded within the software transform the lexical structure of the user interface into auditory format one line at a time. As such they allow blind users to navigate through applications, determine the state of control, and read text without the need to look at a monitor (Bergman and Johnson, 1995).

Limiting the effectiveness of screen readers is the prevalence of graphical user interfaces: screen readers have difficulty accessing and expressing graphic information in words. A direct conversion often

results in the loss of structural cues inherent to visually displayed output or in their conversion to noise (Raman, 1996).

1.4.3.2 Voice Email Interfaces

Email is becoming a mission-critical tool for millions of people and it is predicted that email volume will grow by 60 % in the year of 1999 alone (Strathdee, 1999). Not only is email becoming more prevalent, but so too is the need for business professionals, students, and academic to maintain contact with the office from remote locations. In response to this need, voice interfaces that permit access to email have been developed. Typically these interfaces are designed such that a user dials in to the system and gains access to their email account by mean of an account number or password. Once accessed, the user issues specific voice commands to navigate through new and old email messages and to perform basic email functions (ex. send, reply, and forward).

The main advantage of this type of system is that it provides the user with the ability to gain access to email from a variety of locations, such as stores, hotel rooms, and payphones. This may be of particular interest to the mobile profession who is constantly separated from his or her computer. Mobile professionals are heavy email users, among the more than 21 million people who use Microsoft's Exchange software (Strathdee, 1999). Strathdee also reports that two thirds of business travelers use email extensively, the most prevalent reason being to maintain communication with the office. These statistics provide support for the development of voice access to email. Also advantageous is the telephone-based interaction upon which such a product is based: it is familiar, requires minimal physical effort, and leaves hands and eyes free for other activities, such as driving (Yankelovich, Levow, and Marx, 1995).

Of interest is what usability evaluation methods have been applied to voice email interfaces and with what success. Walker, Fromer, Di Fabbrizio, Mestel, and Hindle (1998) conducted an experiment to

evaluate the usability of ELVIS (EmaiL Voice Interactive System) and to compare the effectiveness of two different types of dialogue designs: system initiative (SI), in which the user is prompted at each stage of the dialogue, and mixed-initiative (MI), in which the user is responsible for knowing what to say. Experimental tasks were presented to SI and MI test participants via web pages. For each task, variables such as total time of the interaction and the number of timeout prompts, ASR rejections, help requests, system turns, and user turns were recorded. Usability was assessed by means of a survey soliciting task performance evaluation and system satisfaction data. Specifically, participants were asked to rate the ease of understanding and completing the task, appropriateness of interaction pace, speed of system response, ability to formulate plans of action, and appropriateness of behavior. The sum of all ratings generated a Cumulative Satisfaction score to which correlations with user performance could be ascertained.

Statistical analyses indicated that users' preferences were not determined by efficiency of use with the hypothesized explanation being that the qualitative aspects (especially automatic speech recognition performance) were of more influence on the user. Interestingly, the final recommendation proposed by Walker et al. (1998) is that a longer study of daily users *in the field* be conducted to confirm the results obtained.

Walker et al. (1998) used usability measures to assist in the determination of the effectiveness of one voice interface feature over another. Yankelovich et al. (1995) employed a different approach in their evaluation of SpeechActs. An experimental conversational speech system integrating speech recognition and synthesis with telephony and natural language processing, SpeechActs provides voice access to email, calendar, weather, and stock quote applications. Of interest to the designers of this product was to use usability testing as a means of supporting iterative redesign of the SpeechActs interface.

A formative evaluation study design was employed in which small groups of users were tested and a substantial amount of iterative design was conducted between groups. Participants were required to complete a set of tasks, the instructions for which were embedded in mail messages. The number of utterances taken, the total amount of time required to complete the tasks, the number of tasks completed, and the recognition rates were recorded. Problems with the interface were identified, presumably via the follow-up interviews conducted and through observation of the participants completing the tasks. However, a systematic procedure for identifying and recording events that greatly hindered or assisted task performance does not appear to have been employed. Like Walker et al. (1998), Yankelovich et al. (1995) state an intention to conduct a longer-term field study.

In summary, a variety of voice email systems exist for which usability testing has been applied. The premise upon which this testing was conducted can vary depending on the objectives of the study (ex. feature comparison versus product redesign). In the two studies reviewed, usability testing was conducted in a laboratory-based setting using a formal evaluation design. Although usability problems were noted, the critical incident method was not employed to systematically collect and record this data. Finally, there appears to be a pervasive awareness of the benefits of being able to evaluate voice interfaces in actual usage conditions and across an extended period of time. In other words, there is a recognized need for remote evaluation.

1.5 PROBLEM STATEMENT

Previous research efforts have demonstrated the feasibility of applying the critical incident technique to remote usability evaluation (ex. Hartson et al. (1996); Castillo (1997)). The validity of these results is limited by the fact that they are based on qualitative data extracted from case studies and exploratory studies. Furthermore, these studies used a pseudo-random approach, whereby remote conditions were

simulated in a laboratory environment. Quantitative measures are needed that can be used to substantiate case and exploratory findings and the implementation of a true remote condition, whereby users outside of the laboratory are engaged in critical incident reporting.

Previous research efforts have also been limited to evaluations of either web pages or graphical user interfaces. When an evaluation method is applied to only one type of interface or software application and the experimenter attempts to generalize the success/failure of this method to other interfaces types or software applications, there is a risk of mono-method bias (Gray and Salzman, 1998). Consequently, generalizations may be false, since the method may work differently depending on the software or system under investigation. Therefore, there is a need to apply the critical incident technique to the evaluation of different interface modalities to verify its feasibility overall.

This research seeks to address the aforementioned gaps in existing research. Specifically, the critical incident technique will be applied to the evaluation of a web interface and a telephone-based interface. Results from both interface types will be compared to assess common or dissimilar attributes. A true remote usability evaluation will be implemented in which users remote to the lab will be used and their performance compared to that of users working in a laboratory environment. Comparing their performance against that of usability experts will assess the feasibility of engaging users in the critical incident reporting process. Finally, changes in critical incident identification and reporting that take place with repeated exposure to the system will be investigated. A modified version of the user-reported critical incident technique developed by Castillo (1997) called the REmote Critical Incident TEchnique (RECITE), will be developed for this purpose.

In summary, the overall objective of this study is to evaluate the effectiveness of extending the critical incident technique to the remote evaluation of web and voice interfaces. Effectiveness is measured as the

extent to which critical incident data generated using RECITE is comparable in frequency and severity to those generated by traditional, laboratory-based applications of the critical incident. The overall objective of this research translates into five specific research goals:

1) Investigate the effectiveness of engaging users versus usability experts in the critical incident identification and reporting process.
2) Investigate the effectiveness of gathering critical incident data from users who are remote to the laboratory environment.
3) Investigate the nature of critical incidents reported during the evaluation of web- and voice-based interfaces.
4) Investigate the effects of repeated exposure to an interface on the critical incident identification and reporting process.
5) Investigate the feasibility of using web-based training and reporting tools to conduct remote usability evaluation via the critical incident technique.

1.6 RESEARCH QUESTIONS

In order to achieve the aforementioned research objectives, the critical incident technique was used to evaluate a web-based as well as a voice interface, and data was collected from users located in a remote setting. Results were compared to critical incident reports obtained from users in a laboratory setting and from those reported by experts observing users. The study was carried out over a period of five days so as to capture and characterize incidents related to the development of skill levels (ex. novice to intermediate user). The pursuit of these research questions will address the following research hypotheses:

1. Is there a difference between the critical incidents (type, frequency, severity[†]) reported by users and those reported through observation of users by usability experts?

2. Is there a difference between critical incidents (type, frequency, severity) reported by users in the field versus users in a laboratory setting?

3. Is there a difference between the critical incidences reported (frequency and severity) while using a web interface versus a voice interface?

4. Is there a difference in the critical incidents (type, frequency, severity) reported during first-time use of the system versus those reported after repeated use of the system?

1.7 RESEARCH HYPOTHESES

For each of the research questions outlined in the previous section, there is a corresponding research hypothesis, as indicated below:

1. There will be no significant difference between the type, number, and severity of critical incidents reported by expert versus trained users. Support for this hypothesis will imply that implementing the critical incident technique with user reporters is at least as valid as using an expert observer.

2. The number and severity of critical incidents reported by remote users will at least be equivalent to those reported by lab-based users. Ideally, the remote users will report a greater number of high severity critical incidents (those whose resolution will have the greatest impact on interface usability), thereby providing support for the benefits of remote evaluation versus lab-based evaluation.

[†] Note: In this study, severity was measured along several dimensions, including impact on task performance, impact on satisfaction, error severity, and task frequency.

3. There will be no interactions between the person who reports the critical incident (user versus expert) and the interface from which the critical incidents originated (web versus voice). Similarly, there will be no interactions between the location from which the critical incidents are being reported (lab versus remote) and the interface from which the critical incidents are originated. Differences in the number and severity of the critical incidents reporting using the web-based versus voice-based reporting tool cannot be evaluated directly since it cannot be ensured that the web and voice interfaces are equivalent with respect to usability problems. In other words, if a greater number of high severity critical incidents are identified for the voice interface, it cannot be determined whether this was due to voice interfaces being more conducive to critical incident technique or due to the fact that this interface had more severe usability problems. Instead, it will be verified that between location and reporter conditions, the number and severity of critical incidents reported is not significantly different.

4. The average number of different critical incidents reported will increase rapidly during the first few days of exposure and then drop off (i.e. fewer and fewer critical incidents are reported) as the number of days increases. The initial large number of critical incidents will reflect the difficulties associated with being a new user of the system and will decline as the user becomes more familiar with the interface and begins to exhaust the functionality of the system. This hypothesized relationship between exposure time and number of critical incidents reported is illustrated in Figure 1-3.

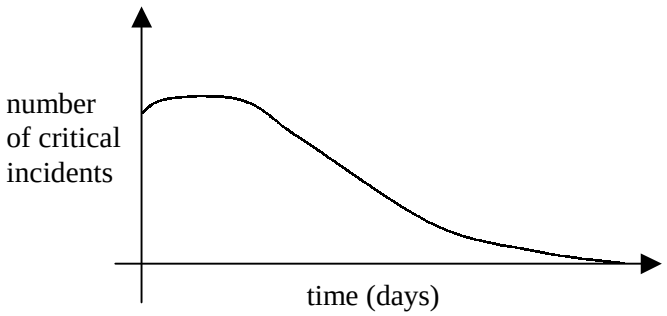

Figure 1-3: Hypothesized Relationship Between Exposure Time and Number of Critical Incidents

4. The satisfaction and task performance impact ratings and error severity ratings reported will be high during a participant's first exposure to the system, on account of a tendency to identify incidents that have the most impact and cause the most severe errors. As exposure time to the interface increases, ratings will decrease as the user begins to exhaust the functionality of the system and is able to extract out incidents have less impact or cause less severe errors. It is at this time that ratings are expected to approach some minimum rating that represents the criteria according to which an incident is considered "critical". This hypothesized relationship is illustrated in Figure 1-4.

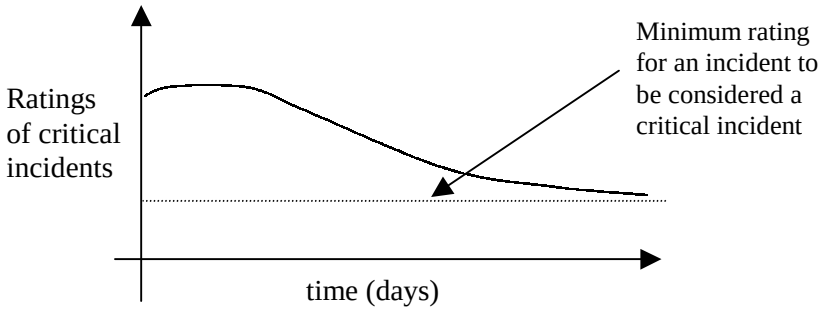

Figure 1-4: Hypothesized Relationship Between Exposure Time and Reported Ratings

CHAPTER 2. METHOD

2.1 EXPERIMENTAL DESIGN

The study used a 3x2x5 mixed-factorial design. The factors of interest were Treatment (T), Interface (I), and Day (D). The levels of each factor and its type are listed in Table 2-1 and were combined to form the data matrix shown in Table 2-2.

Table 2-1: Factor Types

Factor Name	Levels	Type
Interface Type (I)	Web Voice	Within-subject Fixed Effects
Treatment Condition (T)	Remote User Reporters Lab User Reporters Lab Non-Reporters (observed by Expert Reporters)	Between-subject Fixed Effects
Day (D)	Day 1 Day 2 : Day 5	Within-subject Fixed Effects
Subject (S)	S1 S2 etc.	Between-subject Random Effects

Table 2-2: Data Matrix

	Factor I INTERFACE TYPE							
	i1: web				**i2: voice**			
	Factor D DAY							
Factor T TREATMENT CONDITION	**d1: Day 1**	**d2: Day 2**	…	**d5: Day 5**	**d1: Day 1**	**d2: Day 2**	…	**d5: Day 5**
t1: Remote User Reporters	S1 S2 : S10	S1 S2 : S10	…	S1 S2 : S10	S1 S2 : S10	S1 S2 : S10	…	S1 S2 : S10
t2: Lab User Reporters	S1 S2 : S10	S1 S2 : S10	…	S1 S2 : S10	S1 S2 : S10	S1 S2 : S10	…	S1 S2 : S10
t3: Lab Non-Reporters (observed by Usability Experts)	S1 S2 : S10	S1 S2 : S10	…	S1 S2 : S10	S1 S2 : S10	S1 S2 : S10	…	S1 S2 : S10

2.1.1 Treatment Factor

The critical incident technique can be applied in several ways, depending on what individual is selected to report the incidents (reporter) and where that individual is located (setting) relative to the laboratory. Two types of reporters exist: usability expert reporters and user reporters. Usability expert reporters are individuals trained in usability evaluation who observe (directly or indirectly) users interacting with the interface being evaluated and report critical incidents that occur. User reporters are real users of the interface who are trained to identify and report critical incidents that occur during task performance. These critical incidents are then later interpreted by usability analyst(s) for the purpose of creating a list of usability problem descriptions. Reporters can be remote or local. Remote reporters are separated in time and/or space from the experimenter. In contrast, local reporters are located in the laboratory (a controlled environment) and in the direct or indirect presence of the experimenter.

The Treatment (T) factor represents the various applicable combinations of reporter and setting. In this study, comparisons of usability expert versus reporters and remote versus local usability evaluation were of interest. This resulted in three feasible combinations or levels of the Treatment factor: Remote User Reporters, Lab User Reporters, and Lab Non-Reporters. The latter treatment condition represented the case in which user subjects were asked to interact with the interface, but not required to report critical incidents. Video (in the form of screen footage) and audio were recorded onto videotapes, which were then distributed to usability experts for review. Therefore, different sets of data were generated for the same user participant group (i.e. user non-reporters). This is in contrast to the two other treatment levels wherein user participants reported their own critical incidents.

Table 2-3 presents a summary of each Treatment factor level, including the number of participants randomly assigned to each.

Table 2-3: Description of Treatment Condition Levels

Condition	Description	Number of Participants
Remote User Reporters	Users will interact with the experimental interface according to a given set of task scenarios. Users will be allowed to access the system from any location and will be required to identify and report critical incidents on their own.	6 male 4 female
Lab User Reporters	Users will interact with the experimental interface according to a given set of task scenarios. They will be required to come to the laboratory to perform these tasks, but will be responsible for identifying and reporting critical incidents on their own.	6 male 4 female
Lab Non-Reporters/ Usability Experts	Users will interact with the experimental interface according to a given set of task scenarios. Usability experts will review videotapes of these users to identify and report the critical incidents that took place.	6 male 4 female +2 usability experts

2.1.2 Interface Factor

The Interface (I) factor denotes the type of interface being evaluated. Mono-method bias can be avoided when the feasibility of a usability evaluation method is demonstrated across several different interface types. In this study, two interfaces were investigated: a voice or telephone-based interface and a web-based interface. This is a within-subject factor since participants were exposed to both interfaces.

2.1.3 Day Factor

Extending a usability evaluation over several days, weeks, or even months, can be done to capture data relevant to a user's exposure to a system. For example, as the participant gains experience with the interface, they may be more likely to explore the interface, experimenting with features that they would not otherwise have confidence to use. As a result the features of the interface to which the critical incidents apply may shift from basic to more advanced, and a more thorough evaluation of the interface may be achieved. It might also be possible that the number of critical incidents reported will vary with time, with the majority being reported at the start of the evaluation, when the user is unfamiliar with the interface.

In this study, a period of five days was selected for the usability evaluation. The Day (D) factor has five levels, corresponding to each of these days. It was expected that this time period would provide sufficient exposure to capture a transition from novice to intermediate skill levels. A longer time period was not feasible due to the perceived difficulties in recruiting university students able to make long-term time commitments. A formal pilot study to confirm the validity of this study duration was not possible due to time constraints. Instead, the number of critical incidents reported by each of the first seven user-reporting participants was monitored and summed across each day. The results obtained are presented in Table 2-4. As shown, there was a noticeable drop in the number of critical incidents reported on the fifth day, suggesting that benefits in extending the length of the study would be minimal. Hence, sufficient evidence was obtained to retain the five-day study duration.

Table 2-4: Total Number of Critical Incidents Reported by the First Seven Participants

Day 1	Day 2	Day 3	Day 4	Day 5
8	8	5	6	2

2.2 PARTICIPANTS

Thirty (18 males, 12 females) participants ranging in age from 18-30 were recruited to participate in this study. All participants were Virginia Tech volunteer users of the voice email service with no prior experience using the system. Participants were screened by means of a pre-test questionnaire to ensure that they met the following requirements:

- At least one year experience using a PC and an email server application.
- Able to write and speak fluent English.
- No graduate-level courses in human factors engineering, ergonomics, or human-computer interaction and no more than two undergraduate-level courses in any of these areas.
- Prior use of Microsoft Internet Explorer.
- Involvement (either as a participant or a researcher) in no more than 1-2 usability evaluations.

Participants were recruited by means of flyers posted in Virginia Tech campus buildings, advertisements posted on the ISE undergraduate and graduate list serve and from two summer session undergraduate-level ISE courses. Students recruited from the latter group were given extra credit for their participation. Once recruited and verified to meet all screening requirements, participants were randomly assigned to one of the three treatment conditions (remote/reporting, lab/reporting, lab/non-reporting). Six males and four females were assigned to each condition. Balancing across gender was not considered necessary since gender differences related to use of the critical incident technique have not been reported in the literature.

No user participants dropped out of the study. It was necessary to terminate participation of two subjects. In the first case, the participant did not have prior experience with Microsoft Internet Explorer. In the second case, the participant was unable to gain access into the voice email phone system upon four phone-in attempts. These difficulties were most likely attributable to the fact that the participant spoke with an accent making it difficult for the (speaker-independent) system to understand the participant's commands. Accordingly, the system was deemed incompatible with the user. In both cases, the participants were compensated accordingly.

Two additional participants were recruited from the Computer Science Department to act as the usability experts. A usability expert was defined as an individual who had conducted at least five human factors experiments and completed the equivalent of two semesters of human factors-related coursework. It was also required that the experts be familiar with the critical incident technique and have applied it in the context of at least one usability evaluation. It was ensured that the usability experts had no prior experience with the experimental application. This was done to ensure that all participants had similar experience with the interfaces, making it possible to separate the influence of participant background from variations in critical incident reporting strategies (i.e. user versus expert-reported). Lack of prior

experience with the voice email system also ensured that bias in the interpretation of observed user interactions and reporting of critical incidents was minimized.

2.3 EXPERIMENTAL APPLICATION

A voice email messaging service was system as a candidate interface for evaluation. For confidentiality purposes, the name of this service cannot be identified and will henceforth be denoted as VEMS (voice email messaging system).

VEMS is currently targeted towards mobile professionals, but new accounts can be purchased and activated on-line by any interested party for a monthly fee. It was selected on account of the fact that it is comprised of a voice and web interface. The primary interface is the voice interface, which is built on the Watson V2.1 speech engine produced by AT&T. This search engine combines automatic speech recognition and speech synthesis technologies. A system initiative dialogue format is employed, such that the user is provided with prompts at each phase of the interaction. The user responds verbally to these prompts via a limited set of commands. Alternatively, a keypad press may be used. The VEMS voice interface supports a variety of functions that support both standard email tasks, navigation, configuration of the synthesized speech output (ex. change speed or volume), and access to a help facility. The interface is designed to cater to both new and experienced users by providing more or less guidance depending on whether a certain feature is being used for the first time or is regularly used, respectively.

VEMS also includes a web site. The primary purpose of this web site is to allow configuration of the voice account. For example, a user can add or delete members from an address book, reply list, exclude list, or prioritize list, or modify email-handling information. An account number and passcode are required to access this web site.

2.4 APPARATUS

2.4.1 Hardware

In the laboratory, user participant testing was conducted on a Gateway 2000 E-4200 CPU with a 21" Sony Monitor and a standard touch-tone phone were available to interact with the web and voice interface, respectively. Participants' voices were recorded via a portable microphone (Sony Electrey Condenser Mircrophone ECM-T6), worn on a shirt lapel. Telephone interactions with the voice email service were recorded via a Hello Direct Universal Telephone Recorder (Model #TDI-5). Both microphone and telephone audio streams were combined onto a single channel using a Radio Shack SSM-60 Stereo Sound Mixer. A TVator Remote Scan Converter (Antec, Inc.) scan converter was used to capture screen footage. Figure 2-1 illustrates the set-up of the aforementioned apparatus.

Figure 2-1. User Participant Apparatus Set-up

Both audio and video data were recorded onto a VHS tape via a Panasonic SVHS Video Camera Recorder and monitored in real-time using a Sony Trinitron television, as illustrated in Figure 2-2.

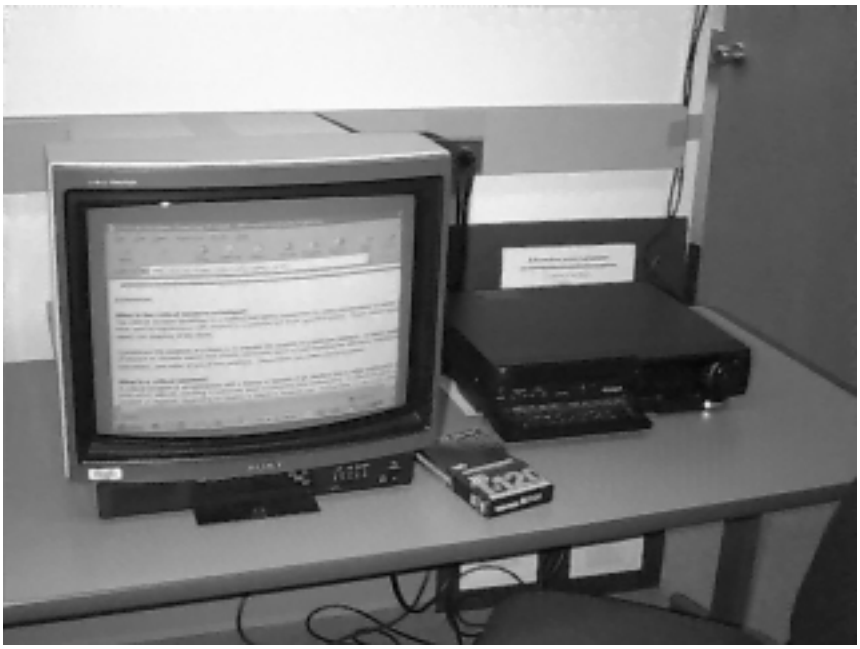

Figure 2-2. Equipment used to record and monitor audio and video data

Equipment to support remote data collection was not supplied. Instead, requirements were imposed by which participants were requested to abide. These requirements included simultaneous access to a telephone and an Internet-accessible computer with Microsoft Internet Explorer (version 4.0 or higher).

A separate workstation, shown in Figure 2-3, was set up for the usability experts to facilitate the review of videotapes. This workstation was comprised of a 14” color monitor connected to a Panasonic SVHS Video Camera Recorder and a Gateway 2000 PC desktop computer with a 17” monitor.

Figure 2-3. Usability Expert Workstation

2.4.2 Evaluation Support Tools

Prior to experimentation, a method of applying critical incidents to remote usability evaluation needed to be developed. Development began by establishing a working definition of the user-reported critical incident technique. Criteria established by Castillo (1997) led to the following definition: a usability evaluation method that involves real (and minimally trained) users performing tasks in their normal working environment reporting high-quality data centered on critical incidents that occur on a daily basis and cost-effectively such that interaction with the evaluator is not required and conversion into usability problems by an expert is easily accomplished. When this technique is applied to remote usability evaluation, an additional criterion is imposed: information transmission between user and evaluator must be equally viable and cost-effective when both parties are separated in space and/or time. According to these criteria, the critical incident technique must incorporate at least the following: 1) user training; 2) critical incident reporting; 3) remote access to information.

Castillo (1997) developed a user-reporter critical incident technique that addressed these criteria (see Section 1.3 for a description). A case study (Castillo, 1997) demonstrated the feasibility of this method, although the need for future refinements was determined. In this study, a method was developed that attempted to implement these refinements. This method, called RECITE (or the REmote Critical Incident TEchnique), required the development of training and reporting tools, each of which is described below.

Critical Incident Training Tool

A self-paced critical incident training tool was created on the premise that in order for users to be competent at applying the critical incident technique, they must be trained. The World Wide Web (WWW) was considered an ideal medium upon which to develop a training program because of its accessibility to an international network of users and ability to rapidly distribute software at low cost.

Training content was developed by means of a needs assessment and through the application of the minimalist approach, as described in Appendix A. The resulting program was comprised of two modules, each of which targeted skill development in a particular area: identifying a critical incident and reporting that critical incident such that a designer can reconstruct the critical incident and respond in a suitable manner. Originally, each module was comprised of a set of instructions followed by a series of hands-on exercises. A pilot study was conducted to determine the effectiveness of the training with hands-on exercise versus training with instructions only. Results of this pilot study are reported in Appendix A and indicate that there was no added benefit of the hands-on exercises, according to both objective and subjective measures. For this reason, hands-on exercises were eliminated from the training session, thereby reducing training time by one-half. Pilot testing also allowed for the usability of the training tool interface to be assessed, and modifications were made to resolve any usability problems before evaluation of the VEMS began. These modifications are described in detail in Appendix A.

Critical Incident Report Form

In this study, a mechanism was needed that could support the transmission of critical incidents from reporters to the experimenter. It was decided that a stand-alone reporting mechanism was the most feasible approach. Several factors motivated the selection of this approach over integrating the mechanism with the experimental interface [for which preference was reported (Castillo, 1997)]. First, it was interface-insensitive, unlike an integrated mechanism, thereby providing a consistent means of reporting critical incidents. Consistency ensures that the potential confounding effect of modality type is minimized. Finally, a stand-alone reporting tool permitted development to occur independent of the interface being evaluated.

Like the training tool, the critical incident report form was designed as a web-based tool in order to benefit from accessibility to remote and laboratory users. To accommodate and emphasize the distinction between positive and negative critical incidents, two separate report forms were created. A General Information box facilitated selection of the appropriate type, as well as providing a means by the reporter could identify him or herself. Figure 2-4 illustrates the design of the General Information box.

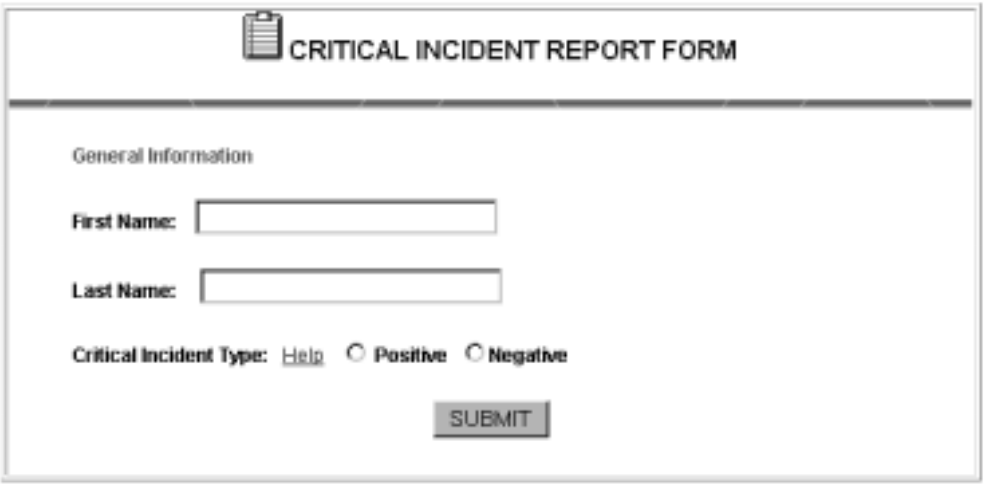

Figure 2-4. General Information Box

The critical incident report form design addressed the many factors associated with a critical incident [as identified by Castillo (1997)]. For instance, both report forms contained input fields for describing aspects of the task being performed and of the critical incident. Keywords were included within these input fields to serve as a reminder of topics to be addressed and as a guide for first-time users. Also available were information buttons whose activation caused a window to open with more detailed instructions and examples.

The critical incident report forms also provided an opportunity to rate the severity of a critical incident. In lieu of a single rating [which Castillo (1997) reported as being difficult to use due to a lack of granularity], four smaller ratings were provided: task frequency, impact on task performance and satisfaction, and error severity (negative critical incidents only). A Likert-type rating scale was not deemed suitable since a "neutral" position was not applicable for many of the ratings. Instead, a 5-point ordinal scale was adopted using verbal descriptors. To increase the speed with which ratings could be selected, a radio button format was used.

Each report form contained a SUBMIT button. Activation of this button initiated a routine to verify that all fields were filled out and submit the contents of the report to the experimenter in the form of an email message. Submitted along with this information was the name of the participant, the time taken to complete the report form, and the number of times an information button was accessed.

A formative evaluation was conducted in a pilot study to evaluate the usability of the critical incident report form interface. Results of this pilot study are presented in Appendix A of this report and were used to resolve any usability problems prior to the evaluation of the VEM system.

Usability Evaluation Web Site

A central means by which to access the training and critical incident reporting tools as well as any other information relevant to the usability evaluation was needed. This need led to the creation of an on-line resource independent of the interface being evaluated called the Usability Evaluation Web Site.

Each experimental condition was provided with its own customized version of the Usability Evaluation Web Site. For instance, user participants assigned to the lab/non-reporting condition did not require access to the critical incident report forms or the critical incident training tool, whereas those assigned to reporting conditions did. Furthermore, expert participants required specific information relating to the daily task scenarios to allow them to better interpret videotape footage. Successful entry of a password a web site login box determined to which web site a participant was taken.

In general, the Usability Evaluation Web Site was designed to provide quick access to the training and reporting tools. A navigation side bar, with picture buttons, was added to ensure access to these components at all times. The main page was designed to provide participants with an overview of the study, links to important information, and a description of all the tools available. A picture of the Usability Evaluation Web Site, with the main components identified, is presented in Figure 2-5.

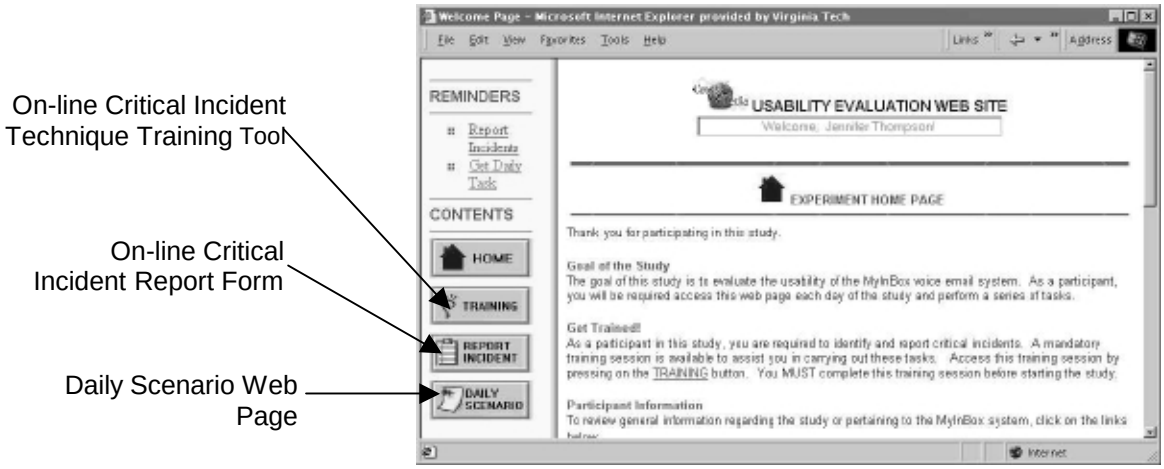

Figure 2-5. Usability Evaluation Web Site Home Page

2.5 TASK SCENARIOS AND EMAIL MESSAGES

Task Scenarios

On each day of the evaluation, user participants were required to complete a task scenario. A task scenario was typically comprised of 3-5 main tasks, broken down into step-by-step instructions for the participant to follow. Structuring via task scenarios ensured that all participants perform the same tasks and hence differences in the number, type, and severity of incidents are more easily attributed to differences in the setting (remote versus lab) or the type of reporter (user versus expert).

Scenarios were designed to provide a fair coverage of the voice and web interface functionalities and to reflect actions representative of a typical VEMS user. For instance, tasks relevant to a user's interaction with the web interface had to account for the main purpose of the web site: to allow the user to create and configure their voice email account. Also taken into account was that reading and responding to email messages are typically done on a daily (if not more frequent) basis and that VEMS only allows these activities to be conducted via the voice interface. Based on these considerations, the Day 1 Scenario was designed with a focus on the web interface and account registration, with miscellaneous tasks related to account configuration incorporated in subsequent task scenarios. Tasks associated with the voice interface were integrated throughout all task scenarios, with greater emphasis being placed after the VEMS account was set up and properly configured (i.e. Day 2 and onwards).

An on-line Daily Scenario home page was added to the Usability Evaluation Web Site to provide remote and laboratory-based participants with equal access to the daily scenarios. As shown in Figure 2-6, links on this home page directed participants to the scenario appropriate for the current evaluation day (ex. on the first day, participants could choose Day 1 Scenario).

Figure 2-6. Daily Scenario Home Page

Email Messages

Task scenarios were created on the basis that a participant would be sent a known set of email messages on each day of the evaluation. To add realism, it was necessary that these email messages come from a variety of senders and contain information relevant to the participant. Creating accounts through free web-based email services and assigning to each of these accounts a persona most readily met these requirements. A business theme was adopted to establish relationships among these personas and to give context to the messages. Other accounts were created for the purpose of sending out "junk email", such as daily weather and travel updates. Table 2-5 presents a summary of the email accounts created for the purpose of this study.

Table 2-5. Email Account Information Summary

Name	Email
Donna Hannun	donna_hannun@yahoo.com
Travel Info	travel_info_99@yahoo.com
Graham Roeburg	graham_roeburg@techpointer.com
Mark Lillehammer	your_supervisor@techpointer.com
Madeline Finch	madeline_finch@techpointer.com
Weather Info	weather_info@usa.com
Billy Bob	friend_billy@writeme.com
Brenda Donnaway	brenda_donnaway@hotmail.com
Your Best Friend	yourbest_friend@usa.net

On each day of the evaluation, email messages were sent directly to the participant's voice email account from one or more of the accounts listed above. Task scenarios required that all email messages be read, and if instructions to perform a certain task were embedded within the email, that these instructions be followed. Additional instructions were given via the task scenarios. Appendix B contains the task scenarios as well as copies of all email messages sent to the participant over the course of the evaluation.

2.6 PROCEDURE

All participants were required to participate on a daily basis (whether remote or laboratory-based) over five consecutive days. The way in which test sessions were conducted varied according to the condition to which the participant was assigned. A description of the testing procedures is provided below, with differences noted.

2.6.1 Introduction Session

All participants were required to participate in an introduction session conducted in the Human-Computer Interaction Laboratory. The exact nature of this session varied according to the condition to which participants were assigned, and included at most a pre-test questionnaire, on-line critical incident training, a system introduction, and the completion of the first task scenario.

2.6.1.1 Informed Consent

The Introduction Session began by presenting all participants with a written overview of the study, its purpose, objectives, and the requirements of their participation. Participants were then given an informed consent form to review, and if in agreement with the terms therein contained, asked to sign it. Consent forms for each participant type (ex. user-reporter/lab, user-reporter/remote, user-subject, usability analyst) are presented in Appendix C of this report.

2.6.1.2 Pre-test Questionnaire

If consent was given, participants were asked to complete a pre-test questionnaire (see Appendix C). The purpose of the questionnaire was to assess a participant's education, computer and web experience, and exposure to voice synthesis and automated speech recognition technologies.

2.6.1.3 System Introduction

Participants were then asked to read a written description of the two systems with which they would be interacting throughout the usability evaluation: VEMS and the Usability Evaluation web site. The description of VEMS was designed specifically to provide only that information which would typically be available to a new voice email service user. For example, participants were shown how to access the voice email service web page and dial into the phone service. A wallet-sized card was given to the participant to provide quick access to voice command names and keypad equivalents. On the back of this card was written important information, such as the voice email web site URL, the Usability Evaluation URL, and the participant's Usability Evaluation password.

The Usability Evaluation Web Site, its purpose, major components, and the method by which it could be accessed were also described. A copy of the system introduction overview is provided in Appendix C of this report. Following the system introduction, participants were brought into the computer room and asked to log into the Usability Evaluation Web Site using the appropriate password. Opportunity was given at this point for the participant to explore the contents of the web site.

2.6.2 Critical Incident Technique Training

Participants assigned to the user-reported conditions (laboratory and remote) were then requested to undergo critical incident training via the on-line critical incident training tool. At the completion of their training session, users were asked to evaluate the tool by filling out an on-line questionnaire. A copy of this questionnaire is provided in Appendix C.

2.6.2.1 First Task Scenario

All participants were then asked to return to the Usability Evaluation Web Site home page and access the Day 1 Scenario from the Daily Scenario web page. Major tasks included in this initial scenario were registering for a new voice email account, configuring the account, and dialing into the voice system to read a message. Participants were encouraged to read through the scenario and then follow the instructions sequentially and to the best of their ability. Assistance was only provided if requested and the participant was unable to complete a task. Participants indicated their completion of the scenario by pressing a "SCENARIO IS COMPLETE" button at the bottom of the scenario web page.

All participants in the reporting conditions were asked to report critical incidents using the on-line critical incident report form. A concern was that participants fill out a report as soon as the critical incident occurs, to ensure that all details of the event were remembered correctly. Although an emphasis on immediate reporting was made in the training program, additional reminders were deemed necessary. Therefore, messages were presented at completion of each daily scenario (i.e. upon activation of a "Scenario is Complete" button) that reminded participants to report all remaining critical incidents before quitting the session completely. Furthermore, reminders to report critical incidents were embedded directly into the scenario instructions, at the beginning of each new task.

2.6.3 Subsequent Test Sessions

Test sessions were scheduled for all laboratory-based participants on a daily basis for a period of five days. During these test sessions, whose length varied from five to thirty minutes, the participants were asked to log into the Usability Evaluation web page, access the appropriate daily scenario web page, and perform the tasks listed on that page. Participants assigned to the laboratory/user-reported condition were encouraged to report critical incident during their interaction with both of the web and voice interfaces. Audio and screen interactions of all participants were videotaped.

Participants assigned to the remote condition were not required to return to the lab for subsequent test sessions. Instead, they were requested to log into the Usability Evaluation web page, access the appropriate daily scenario web page, and perform the tasks listed on that page between 8:00 AM and 10:00 PM each day. Email reminders to complete the daily scenario were submitted to the participant's Virginia Tech email account on the first day of the evaluation and on subsequent days if the scenario was not completed before 9:00 PM. Participants were expected to identify and report any critical incident encountered during their interactions using the on-line report form. In the event that any difficulties or questions arose, participants were encouraged to contact the experimenter (either by phone or email).

2.6.4 Post-test Questionnaire

All participants were directed to an on-line post-test questionnaire form at the completion of their Day 5 Scenario tasks (see Appendix B for a copy). The purpose of this questionnaire was to solicit subjective information from the participant based on their experiences with the voice email service interfaces and their participant in the study. The exact questions varied according to the condition to which the participant was assigned, but in general addressed attitudes towards the following:

- the web and voice interfaces provided by the voice email service and their associated features
- continued use of the voice email service
- the act of reporting a critical incident report form (reporting participants only)
- the questions contained within the critical incident report form (reporting participants only)
- the degree to which the scenario tasks reflected typical emailing tasks

Participants were also asked to identify positive and negative aspects of the voice and web interfaces, as well as of the on-line critical incident report tool.

2.6.5 Compensation

Following submission of the questionnaire, participants were thanked for their participation in the study and reimbursed appropriately. All participants were compensated for their participation. Non-usability expert participants were given $10/hour for their participation in the first test session (which lasted approximately 1-1.5 hours) and a fixed amount of $5 per session thereafter for a total approximate amount of $30-$40.

2.7 PROCEDURES FOR EXPERT EVALUATOR SUBJECTS

When at least half of the non-reporting user participants had completed their 5-day usage sessions, the usability expert participants were invited to individually meet with the experimenter for an Introductory Session.

2.7.1 Introductory Session

During the Introductory Session, expert participants were asked to sign an informed consent form, which provided a detailed introduction to the study and explained the participant's responsibilities with respect to the study. A pre-test questionnaire was then filled out to assess knowledge and experience in human-factors and usability evaluation, as well as experience in the use of web browsers, voice synthesis technology, and speech recognition systems. A copy of this questionnaire is provided in Appendix C of this report. Usability experts were then shown how to access the Usability Evaluation web site and given a brief description of each of its main components.

2.7.2 Critical Incident Training

It was recognized that critical incident training should not be restricted to user-subjects only. For this reason, usability experts were requested to undergo training, albeit using a modified version from that given to the user-subjects. This modified version used a different practice exercise to emphasize skills involved in identifying a critical incident based on observation of another user's interactions. In this practice exercise, a series of narrated screen shots were presented that contained a critical incident. The

task of identifying the critical incident and reporting it using the critical incident report form was assigned. A sample critical incident report form was given at the end of the practice exercise as a form of feedback regarding the participant's performance. Wording changes were also implemented to better reflect the skills and knowledge inherent to a usability expert.

2.7.3 System Introduction

It was important that usability experts be given a thorough introduction to the system to facilitate their interpretation of user interactions and to better understand any critical incidents that occurred. First, a written overview of the basics was given to the usability experts to review. Demonstrations coupled with more thorough explanations of the VEMS voice and web interfaces were then given. An existing voice demonstration was used to highlight features of the voice interface. A toll-free number was dialed and usability expert participants were asked to listed to the first portion of the demonstration, which included a recording of a person interacting with the voice interface to send, read, remove, and forward messages. Participants were then given a short walkthrough of the web interface. A booklet containing all relevant information concerning VEMS was given to the experts for reference purposes. Also provided were paper copies of all email messages sent to user participants on each day of their participation.

2.7.4 Equipment Demonstration

Finally, usability experts were given a brief demonstration of the workstation equipment. This equipment included a commercial S-VHS VCR, a television monitor, a desktop PC computer with Internet access, and the set of ten videotapes containing screen interaction recordings of non-reporting user interactions with the voice email service. With this equipment the usability expert could review the videotapes and report any critical incidents observed.

2.7.5 Data Collection

Usability experts were given approximately two weeks to complete the review of the videotapes. All reviews had to be conducted in the Human-Computer Interaction Lab (Whittemore 320) to ensure confidentiality of the data and to eliminate possible confounding effects of setting and test equipment.

Videotape review involved two activities. First, the experts were required to watch screen footage interactions and identify any critical incidents that occurred during interaction with either the voice or web interface. Second, the experts were required to report the incident. All reporting was done via the on-line critical incident report form (CIRF), accessible from the Usability Evaluation Web Site and designed specifically for use by usability experts. For instance, usability expert report forms did not contain impact on satisfaction ratings, were worded to reflect a usability expert's perspective (i.e. that of an observer), and include an additional rating question pertaining to the overall criticality of the incident. This latter rating was included because it was felt that experts, unlike user-subjects, would have the knowledge and skills necessary to provide accurate criticality ratings. Otherwise, usability expert reports forms were equivalent in formatting, layout, and operation as those designed for user-subject use. All critical incident report form contents were emailed to the experimenter for analysis. Feedback regarding the quality of the reports was given if it was determined that the usability expert was not properly interpreting a question or providing sufficiently descriptive information.

2.7.6 Post-Test Questionnaire

At the end of the two-week period, usability experts were requested to complete a post-questionnaire. The purpose of this questionnaire was to solicit subjective information, such as the ease with which they were able to identify critical incidents and their attitude towards the critical incident form (including the applicability of the questions).

2.7.7 Compensation

Both participants then were thanked and reimbursed for their participation in the study. Usability experts were paid $20 per tape reviewed, for a maximum compensation of $200.

2.8 DEPENDENT VARIABLES

In this study, the effect of the experimental factors and their interactions on the following measures was of interest:

- number of critical incident reported (total, positive, and negative);
- impact on task performance ratings;
- impact on satisfaction ratings;
- error severity ratings;
- time required to report a critical incident;
- number of times help was accessed during reporting; and
- pre-test and post-test questionnaire responses

With the exception of the final item in this list, data for each measure was collected from critical incident reports generated through use of the on-line critical incident report form.

2.9 CLASSIFICATION OF CRITICAL INCIDENT REPORTS

According to Flanagan (1954), data collection of critical incidents must be followed by an inductive process whereby the critical incidents are grouped into areas and sub-areas according to some general frame of reference. In this research, two different processes were used. The first process was based on a bottom-up classification approach, whereby task and critical incident descriptions were summarized into usability success and problem descriptions, and then similar descriptions grouped together. The second process used the Usability Problem Investigator. Based on content and structure of the User Action Framework, the UPI is designed specifically to conduct highly focused and top-down inspection,

resulting in a list of usability problems and successes. Results of each classification process are presented in the sections below.

2.9.1 Classification of Critical Incident Data

The classification of critical incidents into usability problems is a transformation process by which data can be summarized in a manner appropriate for use by designers and developers (or any other stakeholder in the interface designer). Two classification approaches were adopted in this research to summarize critical incident reports. The first approach was a bottom-up classification approach, the purpose of which was to draw out the usability problem or success in a form that could be easily conveyed to designers of the voice email product (i.e. that could be directly transformed into a solution strategy). Once a base-level description was formulated, an attempt was made to group similar problems or successes into higher-level classifications. Nielsen's heuristics were used as a basis for formulating these higher-level classifications. These classifications provided a means of summarizing the problems, and served a descriptive more than a functional purpose.

The second evaluation method used was the Usability Problem Inspector (UPI). The UPI is a part of a set of structured methods and tools used for organizing usability concepts, issues, design features, usability problems, and design guidelines. Unifying these methods and tools is the User Action Framework (UAF), which was developed at Virginia Tech (Hartson, Andre, Williges, and van Rens, 1999). The UAF is based on Norman's stages of action model, which describes the various stages that a user undergoes while interacting with machines (referred to as the Interaction Cycle), and on the premise that during each stage of task-based interaction, the user is affected by the interaction design. An illustration of the interaction cycle and the interaction stages therein contained is provided in Figure 2-7.

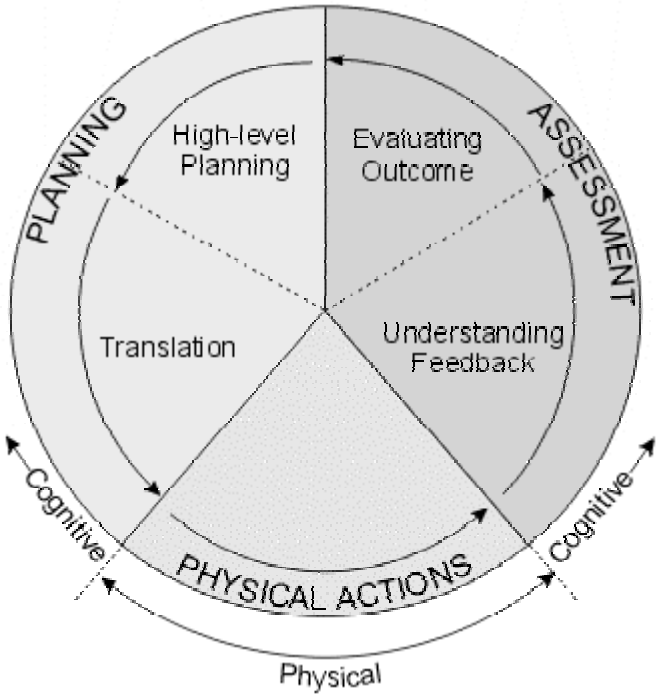

Figure 2-7: The User Action Framework (Hartson et al., 1999)

The meaning and structure of the UAF is mapped to the UPI, albeit expressed in term of problems (and successes) to look for in the system application being evaluated.

The UPI uses a top-down classification method, whereby the investigator must first determine in what interaction activity the reported incident occurred. For example, an incident that occurred while the user was determining what to do and how to do it would be classified under Planning whereas an incident that arose while the user was assessing the outcome of his or her action(s) would be considered an Assessment issue. Once the relevant interaction activity is selected, the investigator proceeds down though a series of hierarchical levels until arriving at a final classification. The number of vertical levels ranges from 2 to 5, each one increasing it its level of specificity. In summary, the UPI is a tool that facilitates highly focused inspection resulting in a list of usability problems or successes.

2.9.2 Justification

Using the UPI to perform a second classification was done for several reasons. First, the UPI provides a user-centered approach to classification by considering what happens to the user during interaction. Trends emerge that can depict where in the interaction users are experiencing a greater number of problems. Investigation of whether critical incident reporting is more amenable to the discussion of problems in one stage than another can also be pursued. These issues, while interesting from a research perspective, are less aligned with the needs of interface designers and software developers who are looking for ways in which to improve the interface. This particular audience requires concrete information, stated in user-specified terms (and not the language of an abstracted model) from which solutions can easily be derived and a means of prioritizing these problems is provided. It is for this reason that both a bottom-up and UPI classification were considered equally important.

The UPI was also implemented for validation purposes. The UPI is primarily developed based on *problems* encountered during interaction with a *graphical user interface.* The data collected in this study provided a unique opportunity to validate this model with usability problems *and* successes pertaining to *web* and *voice* interfaces. This study also presented the opportunity to make a contribution to help build upon the existing UAF knowledge base (comprised of usability concepts, issues, and guidelines) and to demonstrate the possibility of integrating two research endeavors currently being addressed in the Human-Computer Interaction Lab at Virginia Tech.

There was also a practical reason for using the UPI. Classification of problems is an arbitrary process, and it is often to the discretion of the classifier to determine the exact nature of the problem and in what classification group it should be placed. Bias can lead to classification choices that are not valid or repeatable. By making the classification process more systematic, the UPI may help eliminate the arbitrariness inherent to classification, such that repeated classifications of the same data could lead to

near identical groupings. Moreover, it may provide a common means of communicating results obtained from different studies and form a basis for their comparison.

A final reason for using the UAF was to determine its applicability to the classification of critical incidents that may arise from interaction with a variety of different interface modalities and to assess its potential for integration within the critical incident technique. Research efforts are currently being expended to develop the UAF as a web-based mechanism. As such, it lends itself to one of the efforts of this research – the development of a set of web-based reporting and training tools to support remote evaluation via the critical incident technique.

In summary, the UAF was used to capture trends regarding user interaction with VEMS; the bottom-up classification approach was used to capture trends regarding the problems and successes with the features, functions, and operations of the interface.

CHAPTER 3. RESULTS AND DISCUSSION

This chapter is divided into six major sections. The first section presents a summary of negative and positive usability issues, as determined through two systematic classifications of the critical incident report descriptions. Pre-test and training questionnaire data is analyzed in Sections 4.2 and Section 4.3, respectively. Sections 4.4 and 4.5 address the frequency with which positive and negative critical incidents were reported and the severity ratings allocated to these incidents, respectively. The final section presents results and discussions related to the analysis of post-test questionnaire data. A level of significance of α=0.05 was used, unless otherwise noted.

3.1 BOTTOM-UP CLASSIFICATION RESULTS

Bottom-up classification first involved dividing all critical incident reports according to type (positive versus negative) and to the interface that they addressed (web versus voice). Task and critical incident descriptions were then reviewed and a succinct description of each usability problem or success was derived. Usability Problem Descriptions (UPD) and Usability Success Descriptions (USD) are clear and succinct statement that captures what was wrong or right about the interface, respectively, by making direct references to the feature or function involved. An example UPD is: "Account number/password had to be repeated before successful log-in". A UPD does not incorporate possible solutions. The development of solution strategies was considered to fall under the responsibility of the designers, and was hence outside of the scope of the analysis.

UPDs and UPSs were recorded as they were created so that in the event of the same problem or success being reported on multiple occasions, it would be consistently classified under the same description. If a critical incident description was deemed to be a function of the interface, yet contained a description of more than one unique critical incident (ex. two voice recognition errors were reported in a single critical

incident report), the critical incident was classified under multiple descriptions as appropriate. Consequently, the sum of instances across all UPDs and UPSs is higher than the total number of critical incident reports.

In total, 166 unique usability descriptions were reported. Table 3-1 provides a breakdown of these descriptions with respect to the interface (voice or web) and type of description (success or failure) to which they pertain. A list of all usability descriptions, broken down according to interface type and description type, is provided in Appendix D.

Table 3-1: Breakdown of Unique Usability Descriptions

Interface Type	Description Type	Number of Unique Usability Descriptions	Percentage of Usability Descriptions
Voice Interface	Problem	82	49.1 %
	Success	29	17.4 %
Web Interface	Problem	35	21.0 %
	Success	21	12.6 %
Total		**167**	**100 %**

The figures given in Table 3-1 represent the number of unique UPDs and USDs, but not the total number of times each one was reported. Frequency data is provided in Column 3 of the tables in Appendix D. It should be noted that this data combines critical incident reports that were common among the usability experts. This was done to ensure that common critical incidents were counted only once.

Also calculated was the severity of each usability problem or success. In this study, severity was measured along several dimensions, including task frequency, impact on task performance and satisfaction, error severity (negative critical incidents only), and overall criticality (reported by usability experts only). Ratings corresponding to each of these dimensions were averaged across all instances of a usability problem or success and are presented in Column 4-7 of the tables in Appendix D.

Frequency and severity rating data can be used to determine which usability problems or successes should be allocated a higher or lower degree of priority relative to others. The way in which a designer prioritizes problems or successes is likely to vary depending on the objectives of the re-design. For example, it may be of interest to ensure that the most frequently reported usability descriptions be addressed, if minimizing the total number of problems encountered is a usability objective. Or, it may be of interest to give priority to usability descriptions that have the greatest impact on satisfaction, if increasing user satisfaction is the main usability objective. The critical incident data collected in this study provides flexibility with respect to priority criteria.

3.1.1 Usability Descriptions Ranking

In this section, successes and failures pertaining to both voice and web interfaces are sorted according to the following criteria:

- Frequency (all participants)
- Task Frequency (all participants)
- Impact on task performance (all participants)
- Impact on satisfaction (non-experts only)
- Severity of errors (all participants)
- Overall criticality (experts only)

Indicated by the parentheses in the above list are the sources from which the usability descriptions were collected. A discrepancy exists since usability expert and non-expert versions of the critical incident reports were slightly different (i.e. usability experts were not required to report effect on user performance and were asked to evaluate the overall severity of the critical incidents).

Ranking Across a Single Criterion

Due to the large number of usability problems and successes reported for each interface type, it is impractical to present the complete list of problems sorted according to each of the aforementioned criteria. Instead, the top five usability descriptions are presented in Table 3-2 to Table 3-5 for each interface and description type. These usability descriptions have been sorted according to the criteria identified in the top row of each table. Frequency of occurrence was used as a secondary sorting criterion for all severity-related ratings (task frequency, impact on task performance, etc.).

Ranking Across All Criteria

Due to the large number of UPDs and UPSs reported for each interface type, it is impractical to expect that designers address every one. A more practical solution would be to first determine what are the main objectives of the re-design and then prioritize the UPDs and USDs accordingly. Alternatively, if all criterion are deemed important, they could be themselves be ranked (ex. 1st, 2nd, etc.), and then the usability descriptions could be sorted first by the top-ranked criterion, then by the second-ranked criterion, etc.

Another option would be to review the lists of the top five UPDs and USDs generated by sorting according to each individual criterion (see Table 3-6 to Table 3-9) and identify the descriptions that appear multiple times. Occurrences in multiple lists may indicate the most critical UPDs and UPSs, and hence narrow down the number of problems or successes that could be addressed by the designer. The tables below identify the descriptions that occurred in the top five of at least two criterion-sorted lists. The rank of the description in each list is also given.

Table 3-2. Top Five Voice Interface Usability Problems

	CRITERIA					
	Frequency	Task Frequency	Impact on Task Performance	Impact on Satisfaction	Severity of errors	Overall Criticality
1	Account number/ password had to be repeated before successful log-in. 11	VR Error - Next message vs. read message 5	Old messages difficult to locate because not where expected (stacked above new messages). 5	System had difficulty recognizing Dictate New Message command. 5	System had difficulty recognizing Remove Messages More than 1 day old command. 5	Commands in general are not understood by system. 3
2	Cannot remove all messages from a particular sender. 9	System had difficulty recognizing "Read message" command. 5	System had difficulty recognizing Remove Messages More than 1 day old command. 5	VR Error - Read message vs. get 1st new message 5	VR Error - Command (not specified) misinterpreted as "remove message" 5	Command to send a dictation once dictation complete is not obvious/ clear - user unsure of how to proceed. 2.33
3	Command confusion - send reply vs. dictate reply 9	System had difficulty recognizing Goodbye command. 5	Help provided in exclude sender mode does not guide the user in what to do (just says Exclude sender means you will exclude this sender from your mail list") 5	VR Error - Read message vs. next message 5	Ability to restore message not obvious. 5	Cannot add a reply while on the phone. 2
4	System does not allow user to list messages from a non-address book member. 7	VR Error - Read message vs. get 1st new message. 5	VR Error - Command (not specified) misinterpreted as "remove message". 5	System had difficulty recognizing Remove Messages More than 1 day old command. 5	VR Error - Next message vs. get last new message. 5	System had difficulty recognizing "No" command during recipient name confirmation in Dictate New Message mode. 2
5	VR Error - Next message vs. read message. 7	VR Error - Read message vs. next message. 5	Ability to restore message not obvious. 5	VR Error - Command (not specified) misinterpreted as "remove message". 5	Having to go through new messages to get to old messages is time-consuming. 5	VR Error - Dictate new message vs. get first new message. 2

Table 3-3. Top Five Web Interface Usability Problems

	CRITERIA					
	Frequency	Task Frequency	Impact on Task Performance	Impact on Satisfaction	Severity of errors	Overall Criticality
1	Personal Profile Home link difficult to locate (due to size and location of link - expected to be one of the main icons). 8	Log-in information on home page is lost if user accesses Hint page and then returns. 5	Concept and implication of "Prioritize Sender" function unclear. 5	List entry windows should have a toolbar available (for copy and paste functionality). 5	Concept and implication of "Prioritize Sender" function unclear. 4	Personal Profile Home link difficult to locate (due to size and location of link - expected to be one of the main icons). 1.2
2	Requirement that reply name must contain at least 2 words is not stated/obvious. 5	List entry windows should have a toolbar available (for copy and paste functionality). 5	Changes made to an entry in one list are not reflected in other lists containing that same entry (lists are not linked). 5	Concept and implication of "Prioritize Sender" function unclear. 4	"Priority notification is not setup" message is misinterpreted to mean that a priority list entry was not successfully added. 4	Requirement that reply name must contain at least 2 words is not stated/obvious. 1.125
3	Hint Option mis-interpreted as referring to passcode (hint option must match with passcode). 3	MyAddressBook icon not apparent/easily located. 4	Polling set-up feature is not easily found. 4	Unable to prioritize someone directly from address book link (lists should be linked) 4	MyAddressBook icon not apparent/easily located. 4	Account Number log-in box is mistaken for location in which corporate account number is entered. 1
4	ExpressLane is difficult to locate 3	MyPriorityList icon not apparent/easily located. 4	MyAddressBook icon not apparent/easily located. 4	"Priority notification is not setup" message is misinterpreted to mean that a priority list entry was not successfully added. 4	Personal Profile Home link difficult to locate (due to size and location of link - expected to be one of the main icons). 3.13	Polling set-up feature is not easily found. 1
5	Account Number log-in box is mistaken for location in which corporate account number is entered. 2	The number and types of fields that must be specified while creating a new MyPriorityList entry are not clear/obvious. 4	Unable to prioritize someone directly from address book link (lists should be linked) 4	Personal Profile Home link difficult to locate (due to size and location of link - expected to be one of the main icons). 3.33	Account Number log-in box is mistaken for location in which corporate account number is entered. 3	Protocol for deleting list entries not intuitive. 1

Table 3-4. Top Five Voice Interface Usability Successes

	CRITERIA				
	Frequency	Task Frequency	Impact on Task Performance	Impact on Satisfaction	Overall Criticality
1	Remove messages more than X days old permits quick and easy bulk message removal. 8	System correctly recognized commands. 5	System correctly recognized commands. 4.5	Forward message function is quickly and easily executed. 5	System correctly recognized commands. 2
2	Forward message function is quickly and easily executed. 7	Keypad option useful. 5	Restore Message command useful. 4	System correctly recognized commands. 5	List messages from X command reduces the amount of search time. 2
3	Extra assistance provided during first-time use of a command is helpful and reduces confusion. 7	Example account number helpful in log-in. 5	Exclude sender function allows junk mail to be quickly eliminated. 4	Review Dictation command in Dictate New Message mode useful. 5	Remove messages more than X days old permits quick and easy bulk message removal. 1.2
4	Send Reply function is quickly and easily executed. 5	Easier to log-in with practice. 5	Keypad option useful. 4	System easier to use with practice. 5	Forward message function is quickly and easily executed. 1.17
5	Exclude Sender function is quickly and easily executed. 5	Dictate Reply function is easily executed. 5	Example account number helpful in log-in. 4	Voice has good inflection. 5	Send Reply function is quickly and easily executed. 1

Table 3-5. Top Five Web Interface Usability Successes

	CRITERIA				
	Frequency	Task Frequency	Impact on Task Performance	Impact on Satisfaction	Overall Criticality
1	Size and position of main icons (ex. for MyReplyList) & location of managing dialog box make them easily recognized and accessible. 12	Confirmation of new entry good - indicates add entry action successful + allows user to check for errors. 4.33	Protocol for interacting and navigating through web site intuitive/quick and easy. 4.5	Protocol for interacting and navigating through web site intuitive/quick and easy. 5	Confirmation of new entry good - indicates add entry action successful + allows user to check for errors. 2
2	Size and placement of Add icon (in management dialog box) makes it easily recognized and accessed. 11	Size and placement of Add icon (in management dialog box) makes it easily recognized and accessed. 4	Protocol for adding a new priority list member is intuitive/quick and easy. 4.	Protocol for adding new replies intuitive/quick and easy. 5	Size and placement of Add icon (in management dialog box) makes it easily recognized and accessed. 1.09
3	ExpressLane allows for quick and easy addition of new entries (provides shortcut). 7	ExpressLane allows for quick and easy addition of new entries (provides shortcut). 4	Protocol for deleting list entries intuitive/quick and easy. 3.5	Confirmation of new entry good - indicates add entry action successful + allows user to check for errors. 4	Size and position of main icons (ex. for MyReplyList) & location of managing dialog box make them easily recognized and accessible. 1
4	Protocol for deleting list entries intuitive/quick and easy. 6	Protocol for interacting and navigating through web site intuitive/quick and easy. 4	Good feedback for specifying too many prioritize fields in new Priority List entry. 3.5	Protocol for adding a new priority list member is intuitive/quick and easy. 4	ExpressLane allows for quick and easy addition of new entries (provides shortcut). 1
5	Protocol for interacting and navigating through web site intuitive/quick and easy. 4	Separate window for list entry creation/management good. 4	Protocol for adding new replies intuitive/quick and easy. 3.5	Ability to add new entries consecutively (I.e. from confirmation screen) good (provides a short-cut). 4	Protocol for deleting list entries intuitive/quick and easy. 1

Table 3-6. Voice Interface Usability Problems that Occurred in Multiple Criterion-sorted Lists and their Corresponding Ranks

	Frequency	Task Frequency	Impact on Task Performance	Impact on Satisfaction	Severity of errors	Overall Criticality
VR Error - Next message vs. read message	5	1	--	--	--	--
System had difficulty recognizing Remove Messages More than 1 day old command.	--	--	2	4	1	--
VR Error - Read message vs. next message.	--	5	--	3	--	--
VR Error - Command (not specified) misinterpreted as "remove message".	--	--	4	5	2	--
Ability to restore message not obvious.	--	--	5	--	3	--

Table 3-7. Web Interface Usability Problems that Occurred in Multiple Criterion-sorted Lists and their Corresponding Ranks

	Frequency	Task Frequency	Impact on Task Performance	Impact on Satisfaction	Severity of errors	Overall Criticality
Personal Profile Home link difficult to locate (due to size and location of link expected to be one of the main icons).	1	--	--	5	4	1
Requirement that reply name must contain at least 2 words is not stated/obvious.	2	--	--	--	--	2
Account Number log-in box is mistaken for location in which corporate account number is entered.	5	--	--	--	5	3
List entry windows should have a toolbar available (for copy and paste functionality).	--	2	--	1	--	--
My AddressBook icon not apparent/easily located.	--	3	4	--	3	--
Concept and implication of “Prioritize Sender’ function unclear.	--	--	1	2	1	--
Polling set-up feature is not easily found.	--	--	3	--	--	4
Unable to prioritize someone directly from MyAddressBook (lists should be linked).	--	--	5	3	--	--
"Priority notification is not setup" message is misinterpreted to mean that a priority list entry was not successfully added.	--	--	--	4	2	--

Table 3-8. Voice Interface Usability Successes that Occurred in Multiple Criterion-sorted Lists and their Corresponding Ranks

	Frequency	Task Frequency	Impact on Task Performance	Impact on Satisfaction	Overall Criticality
Remove messages more than X days old permits quick and easy bulk message removal.	1	--	--	--	3
Forward message function is quickly and easily executed.	2	--	--	1	4
Send Reply function is quickly and easily executed.	4	--	--	--	5
System correctly recognized commands.	--	1	1	2	1
Keypad option useful.	--	2	4	--	--
Example account number helpful in log-in.	--	3	5	--	--

Table 3-9. Web Interface Usability Successes that Occurred in Multiple Criterion-sorted Lists and their Corresponding Ranks

	Frequency	Task Frequency	Impact on Task Performance	Impact on Satisfaction	Overall Criticality
Size and position of main icons (ex. for MyReplyList) & location of managing dialog box make them easily recognized and accessible.	1	--	--	--	3
Size and placement of Add icon (in management dialog box) makes it easily recognized and accessed.	2	2	--	--	2
ExpressLane allows for quick and easy addition of new entries (provides shortcut).	3	3	--	--	4
Protocol for deleting list entries intuitive/quick and easy.	4	--	3	--	5
Protocol for interacting and navigating through web site intuitive/quick and easy.	5	4	1	1	--
Confirmation of new entry good - indicates add entry action successful + allows user to check for errors.	--	1	--	3	1
Protocol for adding a new priority list member is intuitive/quick and easy.	--	--	2	4	--
Protocol for adding new replies intuitive/quick and easy.	--	--	5	2	--

3.1.2 Discussion of Bottom-up Classification Results

In total, 167 unique usability descriptions were reported. A breakdown of these descriptions by interface indicates that, of the total number of usability problem descriptions derived from the critical incident reports, those pertaining to problems with the voice interface were the most numerous. Reducing the 167 usability descriptions into UPDs and USDs that are ranked according to some criterion may prove useful to the designer for assessing the impact of the critical incidents. Caution in prioritizing UPDs and USDs

according to average rating is advised. If a particular critical incident was reported only once, average ratings will reflect the opinion of only ONE user and may not be representative of what the user population as a whole would say in response to that same incident. Furthermore, there were differences in the characteristics of the critical incident that were rated by usability experts and non-experts. For example, only usability experts assessed overall criticality. Unless both experts and non-experts reported the same UPD or USD, sorting according to overall criticality would draw only from the pool of UPSs and UPDs reported by experts. The sorted list would consequently be biased, since it possible that users and experts may disagree on if the incident was truly critical or on how to rank the various characteristics of that incident.

3.2 UPI CLASSIFICATION RESULTS

The UPI is tool designed to help determine where in the UAF a critical incident occurred. The appropriate location is determined by identifying the way in which the design fails to support what the user wants and needs and the resulting effect on the user. This identification process is facilitated by a hypertext design, which enables the inspector to jump from one location in the UPI to the next. For example, the cause of the design is determined by following a sequence of appropriate links downward through the vertical layers (i.e. web pages) of the UPI. Each link selection updates a right-hand frame that contains Effect on User descriptions. An example screenshot of the UPI is provided in Figure 3-1.

The UPI was used to classify each critical incident report, using the task and critical incident descriptions. Difficulties encountered in the use of this tool limited the classification to the Interaction Activity level (i.e. the top-most layer of the UAF).

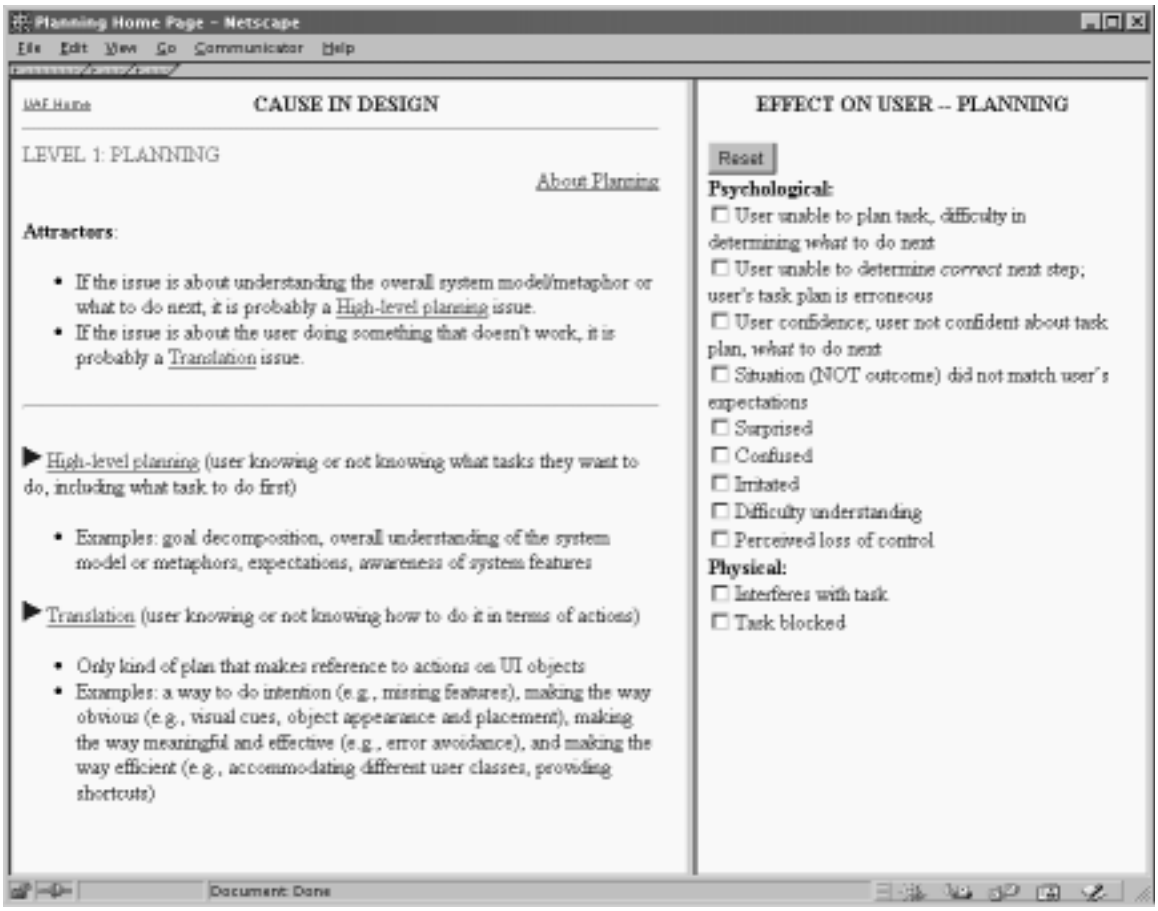

Figure 3-1: Sample UPI Screen Depicting Causes and Effects Relevant to the Planning Interaction Activity (Andre, 1999)

In total, 370 critical incidents were classified using the UPI. Table 3-10 provides a breakdown of these incidents with respect to the interface type, treatment group, and Interaction Activity to which they pertain. Negative and positive incidents have been grouped together since the UPI does not distinguish explicitly between usability problems and successes. A day-by-day breakdown of incidents was not considered practical due to the small number of incidents of each Interaction Activity type reported per day.

Table 3-10. Breakdown of Critical Incidents By Interaction Activity, Interface Type and Treatment Group

	Planning		**Physical**		**Outcome**		**Assessment**		
	Web	Phone	Web	Phone	Web	Phone	Web	Phone	**Total**
Remote	16	22	2	6	1	11	2	6	**66**
Lab	27	34	4	7	2	23	1	11	**109**
Expert 1	54	45	7	25	2	40	5	3	**181**
Expert 2	7	9	0	0	0	16	2	0	**34**
Total	**109**	**117**	**13**	**38**	**5**	**98**	**10**	**20**	**370****

*Note: This total is based on the number of UNIQUE critical incidents reported by each expert (ex. Expert 1 = 161 and Expert 2 = 14)

It was of interest to investigate whether there were differences in the types of critical incidents reported by each treatment group for each interface type. A Chi-Square Test of Goodness was selected to investigate these differences at the level of Interaction Activity. These tests were conducted for each treatment group and interface separately, based on the expectation that an equivalent number of reports would be generated in each Interaction Activity. The following test hypotheses and decision rules were applied:

H_o: O = E (observed frequency is equal to the expected frequency)

H_a: O ≠ E (observed frequency is not equal to the expected frequency)

$\alpha = 0.05$

Decision Rule: I reject H_o if $\chi^2_{observed} > \chi^2_{tabled}$

Test results are presented in Appendix D. In cases for which goodness of fit was not observed (i.e. when there was an unequal distribution of critical incidents across Interaction Activity types), significance was determined by looking for the largest difference. Table 3-11 summarizes the findings.

Table 3-11. Summary of Chi-Square Goodness of Fit Test Results

Treatment Group	Interface	Conclusion	Significant Difference
Remote/Reporting	Web	Reject H_o – there is evidence that the numbers of web-related critical incidents reported by Remote/Reporting treatment were unequally distributed across the interaction activity types.	A larger number of planning incidents were reported than expected by the remote users when interacting with the web interface
	Voice	Reject H_o – the numbers of voice-related critical incidents reported by the Remote/Reporting treatment were unequally distributed across the interaction activity types.	A larger number of planning incidents were reported than expected by the remote users when interacting with the voice interface.
Lab/Reporting	Web	Reject H_o – the numbers of web-related critical incidents reported by the Lab/Reporting treatment were unequally distributed across the interaction activity types.	A larger number of planning incidents were reported than expected by the laboratory-based users when interacting with the web interface
	Voice	Reject H_o – the numbers of voice-related critical incidents reported by the Lab/Reporting treatment were unequally distributed across the interaction activity types.	The laboratory-based users interacting with the voice interface reported a larger number of planning incidents than expected. This condition also reported far fewer Physical Action critical incidents than expected.
Expert 1	Web	Reject H_o – the numbers of web-related critical incidents reported by Expert 1 were unequally distributed across the interaction activity types.	A larger number of planning incidents were reported than expected by Expert 1 when interacting with the web interface
	Voice	Reject H_o – the numbers of voice-related critical incidents reported by Expert 1 were unequally distributed across the interaction activity types.	Expert 1 reported far fewer Assessment incidents for the voice interface than expected.
Expert 2	Web	Reject H_o – the numbers of web-related critical incidents reported by Expert 2 were unequally distributed across the interaction activity types.	A larger number of planning incidents were reported than expected by Expert 2 when observing interactions with the web interface.
	Voice	Reject H_o – the numbers of voice-related critical incidents reported by Expert 2 were unequally distributed across the interaction activity types.	Expert 2 reported far fewer Assessment and Physical Action incidents for the voice interface than expected.

The results summarized in the table above indicate a general lack of equality in the distribution of critical incidents across the various Interaction Activity categories, regardless of the reporter or the interface being evaluated. A common finding among all treatment groups was that a larger number of planning

incidents related to the web interface were reported than expected. This finding may be indicative of a greater ability to detect and articulate incidents that occur during this phase of the Interaction cycle. It also may indicate that the strengths and/or weaknesses of the web page design are currently associated with features and affordances that support the planning phase.

The analysis of voice interface data revealed differences among the treatment conditions. User reporting treatment groups reporter a higher than expected number of critical incidents related to the Planning Interaction cycle, with laboratory-based users also reporting fewer than expected incidents related to Physical Action. Expert 1 and Expert 2 reported fewer than expected Assessment incidents, with the latter two also reporting fewer than expected Physical Action incidents. This lack of consistency can be related to differences in reporting strategies rather than to the design of the interface, since the same one was used across all treatment groups. For example, it may be more difficult for an observer of end-users to identify Assessment or Physical Actions incidents that occur during interaction with a voice interface.

Difficulties were encountered in the application of the UPI to the classification of critical incident data. Some of these difficulties stem from the fact that the UPI was designed with respect to graphical user interfaces, making it difficult to classify incidents pertaining to the voice interface. For example, it was not clear from the way in which the UPI defined feedback if it encompassed synthesized voice commands (which are issued in response to a command, whether that command was successfully or unsuccessfully used).

Additional difficulties arose due to fact that UPI classification was sensitive to the way in which the critical incident description was worded and the extent to which the reporter detailed how the critical incident occurred. As a result, critical incidents grouped under same USD or UPD were occasionally classified into different interaction activities. To illustrate this phenomenon, consider the UPD for old

messages being difficult to locate using the voice interface. Some users cited the problem as being caused by non-intuitive navigation commands. This description is best classified in the Planning Interaction activity, and specifically in the Translation and Meaning & Effectiveness category. Other users described this problem in terms of the lack of usefulness of a prompt received in the process of trying to access saved messages. According to this description, the more appropriate Interaction Activity classification is Assessment, under Clarity of Meaning. This discrepancy may be due to an inability of the bottom-up classification process to distinguish between what are really different UPDs or USDs. Alternatively, it may be due to the extreme degree of granularity inherent to the UPI, which gives so many possible classification routes that two identical critical incidents may never be similarly classified. In either case, the UPI did not completely eliminate subjectivity from the classification process.

3.3 PRE-TEST QUESTIONNAIRE DATA

A pre-test questionnaire was distributed in order to obtain data relevant to each participant's background to assess participant status eligibility and to demonstrate equality among experimental conditions. A summary of responses gathered from each participant group is provided in Table 3-12.

Table 3-12. Summary of Pre-Test Questionnaire Responses

Question	Possible Response	Lab/ Reporting (n=10)	Lab/Non-Reporting (n=10)	Remote/ Reporting (n=10)	Usability Experts (n=2)
How many university-level courses have you taken that addressed human factors evaluation methodologies, usability evaluation, or human-computer interaction?	none	7	5	6	--
	1 - 2	3	5**	4**	--
	3- 5	--	--	--	2
	5+	--	--	--	--
How many usability evaluations have you conducted in the past (ex. as a participant, for a course project, for research purposes, in industry)?	None	8	5	8	--
	1-2	2	4	1	--
	3-5	--	1	1	1
	6-10	--	--	--	1
	10+	--	--	--	--
For how long have you been using computers?	> 6 months	--	--	--	--
	6 – 12 months	--	--	--	--
	1 – 3 years	2	1	1	--
	3+ years	8	9	9	2
Rate your level of expertise using Microsoft Internet Explorer	Very	2	1	4	--
	Moderate	7	8	5	2
	Minimal	1	1	1	--
	none	--	--	--	--
Rate your level of expertise with automatic speech recognition	Very	--	--	--	--
	Moderate	--	--	1	1
	Occasional	2	4	1	--
	none	8	6	8	1
Rate your level of expertise with synthesized voice	Very	--	1	--	--
	Moderate	--	1	1	1
	Occasional	4	3	5	1
	none	6	5	4	--
I am very familiar with the critical incident technique	5-point Likert Rating Scale*	4.5 (average)	N/A	4.7 (average)	1, 1
I am very experienced at applying the critical incident technique	5-point Likert Rating Scale*	4.5 (average)	N/A	4.7 (average)	1, 1
Have you ever used an on-line instructional tool?	Yes	6	5	3	1
	No	4	5	7	1

*1=strongly agree, 2=agree, 3=neutral, 4=disagree, 5=strongly disagree
**Undergraduate-level courses

3.3.1 User Participant Data

It is of interest to demonstrate equivalence among user participant groups to ensure that prior experience was not a potential confounding variable. Due to their potential influence on user participant performance, a more detailed investigation of the familiarity and experience of user participants was

warranted. A one-way Analysis of Variance (ANOVA) was conducted to investigate this issue across each of the reporting treatment groups (this data was not collected from the lab/non-reporting group since they were not required to learn or apply the critical incident technique). An ANOVA was considered a suitable analysis approach since a true Likert rating scale was used to collect both familiarity and experience rating data. In this analysis, a result of no difference is of interest. Accordingly, a larger level of significance is desired; in this analysis a level of 0.2 is used. Results of the analyses are presented in Table 3-13 and Table 3-14.

Table 3-13. ANOVA Summary Table for Familiarity with the CIT Prior to Training

Source	DF	SS	MS	F	P
Treatment	1	0.10	0.10	0.08	0.784
Error	8	10.00	1.25		
Total	9	10.10			

Since p=0.784>> 0.2, it can be concluded that there was no difference in the level of familiarity of the reporting training groups.

Table 3-14. ANOVA Summary Table for Experience with the CIT Prior to Training

Source	DF	SS	MS	F	P
Treatment	1	0.400	0.400	1.60	0.242
Error	8	2.000	0.250		
Total	9	2.400			

Since p=0.242>>0.2, it can be concluded that there was no difference in the level of experience of the reporting training groups.

In summary, the results suggest that, with respect to experience and familiarity with the critical incident technique, the reporting participants groups (lab/reporting and remote/reporting) were not statistically different from one another. This finding increases the certainty to which differences in dependent measures (ex. number of critical incidents reported) can be attributed to the independent variables.

3.3.2 Usability Expert Participant Data

Usability experts were recruited based on several criteria, including prior experience in conducting usability evaluations, successful completion of at least two semesters of human-factors related courses, and in applying the critical incident technique. Pre-test questionnaire data presented in Table 3-15 indicate that these screening mechanisms were generally successful in recruiting participants with similar experience and skill sets. Some slight discrepancies in usability experts' prior experience with speech recognition, voice synthesis technologies, and on-line instructional tools were found. However, since the role of the usability experts was to observe interactions of user participants with the voice email service versus themselves interacting with it, their potential as confounding variables is considered negligible.

It is important to demonstrate that the usability experts had significantly greater awareness and experience with the critical incident technique than the non-usability experts (hence confirming their status as usability experts). Due to the non-equality of group sizes (two expert usability participants versus thirty user participants), a formal analysis of variance was not conducted. However, as Table 3-15 indicates, the average user participant ratings were in all cases much lower than those of the usability expert participants.

Table 3-15. Comparison of Usability Expert versus User Participant Familiarity and Experience with the Critical Incident Technique

Treatment Group	Average Familiarity with the Critical Incident Technique	Average Experience in Applying the Critical Incident Technique
Lab/Reporting	4.5	4.5
Lab/NonReporting	N/A	N/A
Remote/Reporting	4.7	4.7
Usability Expert	1	1

3.4 USABILITY EXPERT PERFORMANCE COMPARISON

Before analyses of critical incident data could be conducted, a supplemental analysis was required to account for the fact that two separate usability experts generated critical incident reports from the same set of data (i.e. from observation of the lab/non-reporting participants). Their performance was compared

to determine whether or not they observed the same thing; that is, whether their performance was similar with respect to the number of critical incidents reported and the severity ratings of these incidents. Since experts generated critical incidents from the same set of data, it was expected that the critical incidents reported would be the same.

To test this hypothesis, the critical incidents reported by the usability experts were reviewed. . In some cases, a critical incident was classified as inadmissible and removed from the data set. This was done if the incident was not a function of the interface being evaluated. Examples included critical incidents that pertained to the experimental protocol (ex. user omitted a required task or user was sent a duplicate message) (5 cases) or to an Internet Explorer internal error (1 case). It is interesting to note that Expert 1 was the source of all inadmissible critical incidents.

Admissible critical incidents reported for each participant were then compared to identify those that were common between both experts and those that were unique. Common critical incidents were then grouped into a separate data set, called the Common Expert. The total number of unique critical incidents per non-reporting user participant is presented in Table 3-16.

Table 3-16. Total Number of Critical Incident Reports Generated by Usability Experts

Participant	Expert 1	Expert 2	Common
1	16	3	5
2	13	1	2
3	16	0	0
4	16	2	2
5	15	0	2
6	10	1	3
7	17	0	1
8	18	2	2
9	11	2	3
10	12	0	3
AVERAGE	14.4	1.1	2.3

A t-Test for a Paired Two Sample for Means was conducted to determine whether the average number of unique critical incidents generated by usability experts was distinct. Expected was that a finding of non-significance (equivalent usability expert performance) would be found, allowing only the common critical incidents to be retained and used to conduct ANOVAs for each of the dependent variables. Results in Table 3-17 prove otherwise. Since $p<0.05$, the null hypothesis (no difference) was rejected. Therefore, there was a significant difference in the average number of critical incidents reported by the two experts ($\alpha = 0.05$).

Table 3-17. T-test Results for the Comparison of the Unique Expert-Reported Critical Incidents

	Variable 1	*Variable 2*
Mean	14.6	1.3
Variance	6.711	1.1222
Observations	10	10
Pearson Correlation	-0.0324	
Hypothesized Mean Difference	0	
df	9	
t Stat	14.8595	
P(T<=t) one-tail	6.1206E-08	
t Critical one-tail	1.8331	
P(T<=t) two-tail	1.2241E-07	
t Critical two-tail	2.2622	

A more detailed comparison of usability expert performance can be achieved by assessing the correlation between the numbers of critical incidents reported <u>for each participant</u>. The correlation coefficient is used to measure the strength of the relationship between two sets of data. A correlation coefficient of 0.3136 was calculated for the total number of critical incidents reported by each usability expert. A test for significance of this correlation coefficient is presented below:

H_0: $r = 0$
$\alpha = 0.05$
Test statistic: $t^* = R\frac{\sqrt{n-2}}{1-R^2} = 0.3136\frac{\sqrt{10-2}}{1-0.3136^2} = 0.9341$
Decision rule: I reject H_o if $|t^*| \geq t_{\alpha,n-2} = 1.833$ (one-tailed)
Conclusion: Since $t^* < 1.8333$, fail to reject H_o.

Therefore, the number of critical incidents reported by Expert 1 does not correlate with number of critical incidents reported by Expert 2. That is, large numbers of critical incidents reported for a particular participant by Expert 1 are not associated with large numbers given by Expert 2. Gray and Salzman (1998) refer to this phenomenon as the Wildcard Effect, whereby two participants were recruited that were extremely above and extremely below average.

Having determined that usability experts were not consistent with respect to the number of critical incidents observed for the 10 participants, it was then of interest to examine performance when the same critical incident was reported. The correlation coefficient between the ratings assigned to each severity measure (task frequency, impact on task performance, and overall criticality) for each common critical incident was calculated (r = 0.751). A test for significance of this correlation coefficient is presented below:

H_0: r = 0

Test statistic: $t^* = R\frac{\sqrt{n-2}}{1-R^2} = 0.751\frac{\sqrt{81-2}}{1-0.751^2} = 10.237$

$\alpha = 0.05$

Decision rule: I reject H_o if $|t^*| \geq t_{\alpha,n-2} = 2.0$ (one-tailed)

Conclusion: Since t* >> 2.0, reject H_o.

Severity ratings assigned by the two usability experts for common critical incidents were thus positively related. That is, ratings given by Expert 1 were associated with large ratings given by Expert 2. Therefore, while usability experts disagreed on the total number of critical incidents that occurred for each participant, they were in agreement on severity ratings in the event that they both identified the same critical incident.

Based on t-test and correlation results, it was determined that expert data could not be combined. Instead, it was necessary to analyze Expert 1 and Expert 2 data separately. Separate analyses limit the generalizations regarding expert performance to only those measures for which Expert 1 and Expert 2 compared to user reporters gives the same result (ex. both experts report higher impact on task performance ratings than user-reporters). The "Common Expert" data set, comprised of critical incidents reported common among the experts, was not analyzed. Preliminary analyses indicated that this data set was primarily just an artifact set of results of two very divergent experts and contributed little to the understanding of usability expert performance.

3.5 CRITICAL INCIDENT FREQUENCY DATA

The number of critical incidents reported by the participants and experts were analyzed to make comparisons of performance among treatment groups over each day of the evaluation and for each interface type (voice and web). A total of 365 critical incident reports were submitted. Table 3-18 indicates the distribution of these critical incidents across treatment condition, interface type, and critical incident type. It should be noted that these totals are based on the number of critical incidents *as reported*. That is, a count of one is assigned to each critical incident report submitted, regardless if that report actually described multiple critical incidents.

Table 3-18. Total Number of Critical Incidents Reported Per Interface and Critical Incident Type

	Web Interface		Voice Interface		
	Negative	Positive	Negative	Positive	**Total**
Lab/Reporting	17	13	46	25	**101**
Remote/Reporting	12	9	35	6	**62**
Expert 1	20	43	74	31	**168**
Expert 2	8	2	23	1	**34**
Total	**57**	**67**	**178**	**63**	**365**

ANOVAs were carried out with respect to the following dependent measures:

- total number of critical incidents
- total number of negative critical incidents

- total number of positive critical incidents

In each case, three separate ANOVAs were required to accommodate the two different sets of usability expert data (ex. Expert 1 and Expert 2). Significant main effects and interactions are investigated to isolate the significant differences using the Newman-Keuls test. The Newman-Keuls method is based on the studentized range distribution, which differs from the t-distribution only in that it takes into account the number of means under consideration. Specifically, the greater the number of means, the larger the critical value of the studentized t, accounting for the fact that there is a higher likelihood that at least some differences between pairs of means will be large due to chance alone (Lane, 1999).

3.5.1 Total Number of Critical Incidents Reported

Analyses were conducted to investigate the effect of the interface type, day, and treatment conditions, as well as their interactions, on the total number of critical incidents reported by each participant. ANOVA results for the total number of critical incidents reported using Expert 1 and Expert 2 data are shown in Table 3-19 and Table 3-20. In each analysis, the main effects of Treatment, Day, and Interface were significant at $p < 0.001$, as was the interaction of Interface and Day at $p < 0.001$.

Table 3-19. ANOVA Table of Total Critical Incidents Using Expert 1 Data

Source	DF	SS	MS	F	P
Between					
Treatment (T)	2	57.4867	28.7433	27.32	0.000*
S/T	27	28.4100	1.0522	**	
Within					
Day (D)	4	30.3800	7.5950	9.51	0.000*
D*T	8	9.7800	1.2225	1.53	0.155
D*S/T	108	86.2400	0.7985	**	
Interface (I)	1	30.0833	30.0833	29.94	0.000*
I*T	2	2.2867	1.1433	1.14	0.335
I*S/T	27	27.1300	1.0048	**	
I*D	4	62.6333	15.6583	19.87	0.000*
I*D*T	8	16.2467	2.0308	2.58	0.013
I*D*S/T	108	85.1200	0.7881	**	
Total	**299**	**435.7967**			

Table 3-20. ANOVA Table of Total Critical Incidents Using Expert 2 Data

Source	DF	SS	MS	F	P
Between					
Treatment (T)	2	20.8067	10.4033	11.12	0.000*
S/T	27	25.2600	0.9356	**	
Within					
Day (D)	4	15.7333	3.9333	7.95	0.000*
D*T	8	2.4267	0.3033	0.61	0.765
D*S/T	108	53.4400	0.4948	**	
Interface (I)	1	18.2533	18.2533	19.48	0.000*
I*T	2	4.2467	2.1233	2.27	0.123
I*S/T	27	25.3000	0.9370	**	
I*D	4	26.2133	6.5533	11.84	0.000*
I*D*T	8	5.1867	0.6483	1.17	0.323
I*D*S/T	108	59.8000	0.5537	**	
Total	299	256.6667			

Significant main effects and interactions of the overall analyses were isolated using Newman-Keuls post hoc analyses, the results of which are presented in Appendix D.

Main Effect of Treatment Group

The number of critical incidents reported by Expert 1 was significantly higher than that reported by each of the two user reporting groups. In contrast, Expert 2 reported a significantly lower number of critical incidents than did the laboratory-based users. These findings illustrate the extreme variability among the usability experts and suggest that Expert 1 reported a larger than average number of critical incidents overall. Figure 3-2 illustrates these differences.

Another finding generated from post hoc comparisons of the Treatment main effect is that the number of critical incidents reported by the users in the lab is significantly higher than those reported by remote users. This finding challenges the hypothesis that remote and laboratory-based usability evaluations are comparable. It could be that the laboratory environment helped foster intrinsic motivation to report incidents on account of being under observation by the experimenter. Differences in the equipment used by laboratory-based users and remote users may have also contributed to the differences, if one particular set of equipment was more or less amenable to the occurrence or detection of a critical incident.

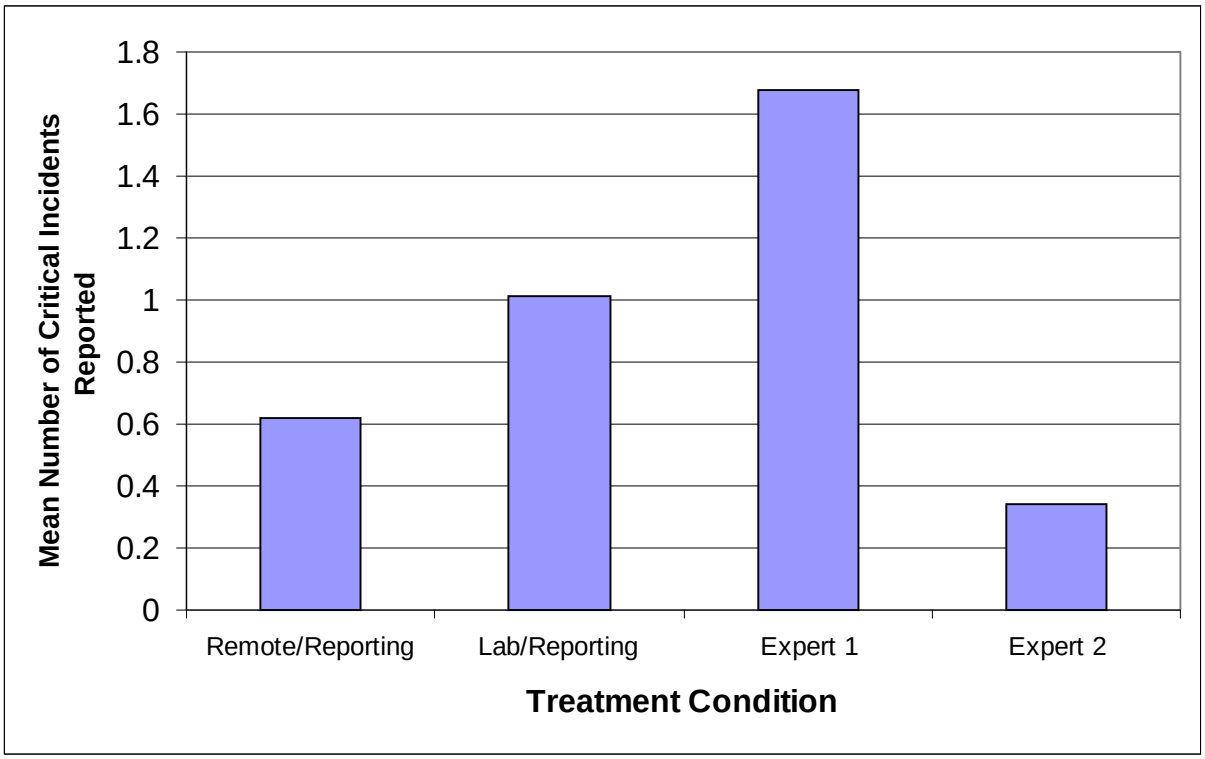

Figure 3-2. Mean Number of Critical Incidents Reported Per Treatment Condition

Main Effect of Interface Type

Since the Interface factor has two levels (voice and web), isolation of the main effect was conducted through a comparison of the mean number of critical incident reports generated for each type. Mean numbers corresponding to Expert 1 and Expert 2 analyses are presented in Table 3-21 and Table 3-22, respectively.

Table 3-21. Mean Number of Critical Incidents Reported Per Participant Per Interface (Expert 1 Data)

Interface	N	Mean	St. Dev
Web	150	0.760	0.981
Voice	150	1.447	1.314

Table 3-22. Mean Number of Critical Incidents Reported Per Participant Per Interface (Expert 2 Data)

Interface	N	Mean	St. Dev
Web	150	0.4067	0.7866
Voice	150	0.9067	0.9854

In both cases, a significantly higher number of critical incidents were reported for the voice interface than for the web interface, most likely due to differing levels of experience and familiarity with each interface type. Pre-test questionnaire results, for instance, indicated that while most participants had experience in the use of a web browser, very few were familiar with speech recognition or voice synthesis technologies. The ability of the critical incident technique to discern differences in skill levels provides support to the overall effectiveness of this evaluation method.

Main Effect of Day

Newman-Keuls results that isolated the main effect of Day differed depending on whether Expert 1 or Expert 2 data were used.

Expert 1 Data

The number of critical incidents reported on Day 1, Day 2, and Day 4 was significantly higher than those reported on Day 3 and Day 5. Furthermore, those reported on Day 3 were significantly higher than those reported on Day 5. Figure 3-3 illustrates the differences between the mean numbers of critical incidents reported per day. These results support the hypothesis that over the course of the usability evaluation, the number of critical incidents reported will decrease over time, although an initial rise does not appear to occur. The secondary peak observed on Day 4 was unexpected. Several factors may have contributed to the occurrence this peak. For example, the tasks assigned on that day may have been novel (increasing the likelihood of encountering a problem) or may have capitalized on a skill and knowledge base developed over the course of the evaluation (increasing the likelihood of encountering a success). Motivational factors may also have contributed to the peak: participants may have felt more inclined to report incidents on account of the upcoming end of the evaluation period.

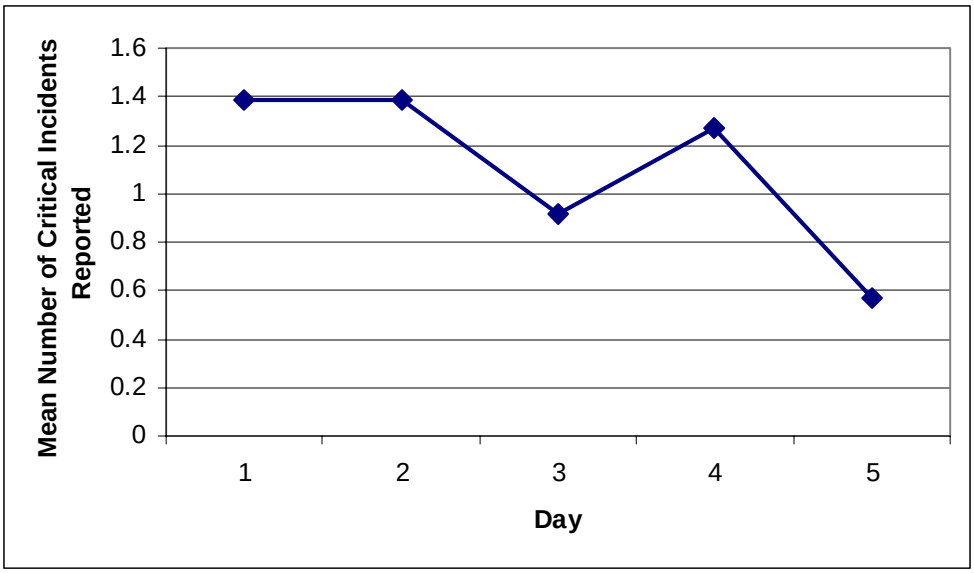

Figure 3-3. Mean Number of Critical Incidents Reported Daily Per Participant (Expert 1 Data)

Expert 2 Data

Inspection of post-hoc comparison test results reveals that the number of critical incidents reported on Day 1 and Day 2 are each significantly higher than those reported on Day 3, Day 4, and Day 5. Figure 3-4 illustrates the differences between the mean numbers of critical incidents reported per day, showing a progressive decrease in numbers, as hypothesized. These results, while in agreement with research hypotheses, are partially consistent with those generated using Expert 1 data. The main difference is that the rise in critical incidents reported on Day 4 is not significantly higher than those reported on Day 3. This suggests that the rise may be linked to Expert 1 performance rather than to any particular attribute of user performance.

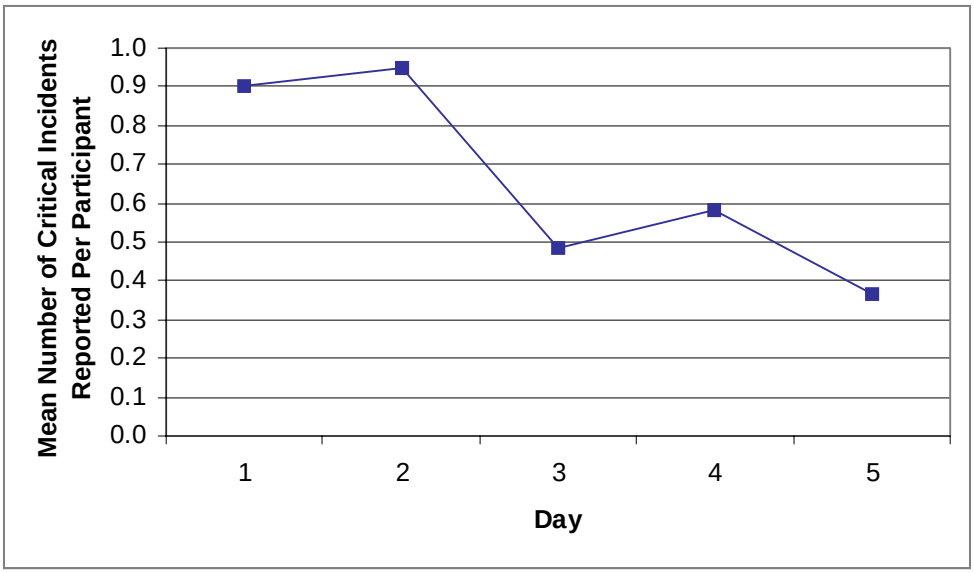

Figure 3-4. Mean Number of Critical Incidents Reported Daily Per Participant (Expert 2 Data)

Interaction Between Interface and Day

An interaction occurs when the relationship of one independent variable and the number of critical incidents reported depends on the level of a second dependent variable. It is of interest to determine at which level(s) this relationship exists. Newman-Keuls results for the Interface and Day interaction were similar across Expert 1 and Expert 2 data sets, as illustrated in Figure 3-5 and Figure 3-6, respectively.

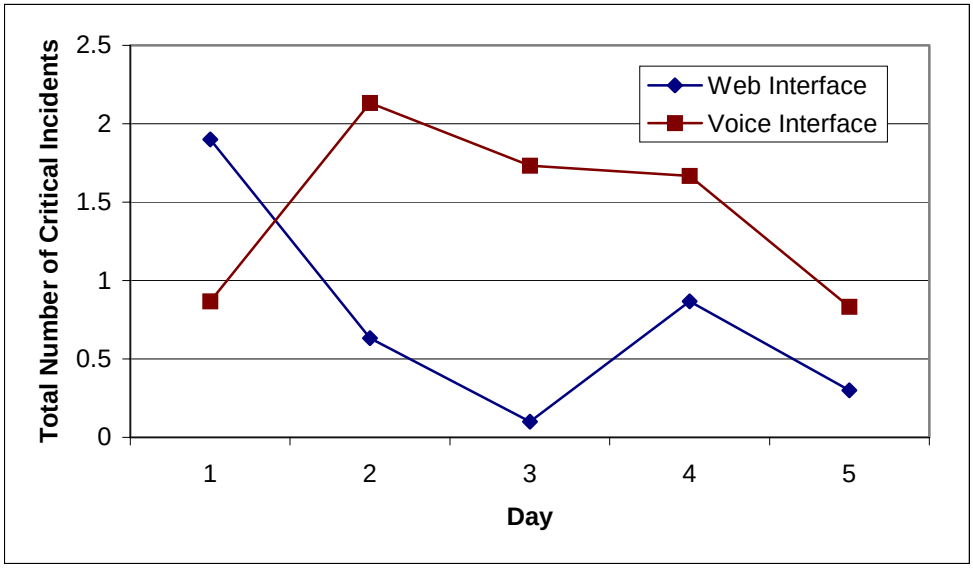

Figure 3-5. Interface x Day Interaction for Total Number of Critical Incidents Reported

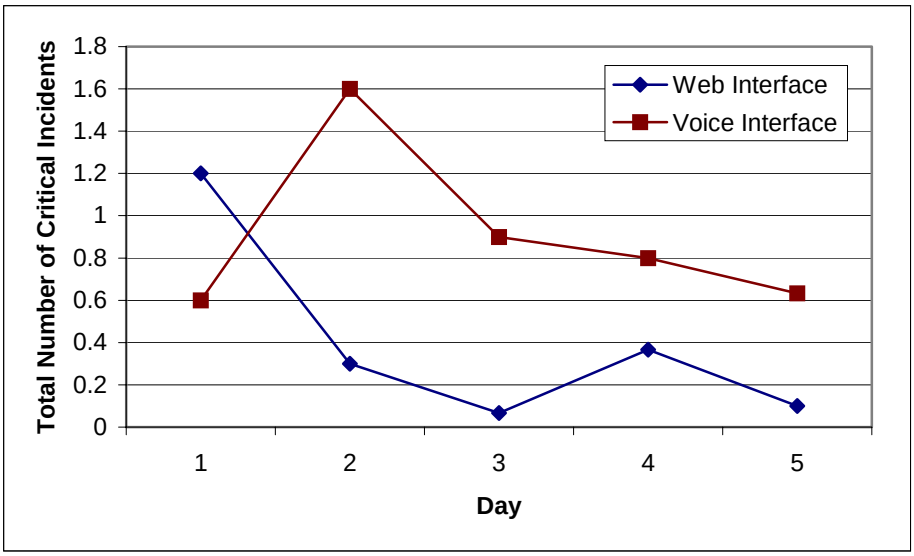

Figure 3-6. Interaction of Total Number of Critical Incidents Reported (Expert 2 Data)

As shown in the graphs above, the number of critical incidents reported for the voice interface was higher than that reported for the web interface on each day except the first day. This may be attributable to the fact that tasks assigned on the first day were focused on the web interface, allowing for more opportunity for a critical incident to occur. The difference between critical incidents pertaining to voice and web interface is greatest on Day 2 and Day 3 than on remaining days. An explanation requires consideration of the task scenarios assigned on those particular days. On Day 3, for example, no web interface tasks were assigned. On Day 2, tasks related to both interfaces were assigned. The larger number of incidents reported for the voice interface may be due to a differential familiarity and experience with voice versus web interfaces. Consequently, users may have experienced more difficulties with the voice interface initially but with increasing exposure their aptitude with both interfaces may have become progressively equivalent, causing the difference in numbers of critical incidents reported for both interfaces to decrease. This interpretation presupposes that the number of critical incidents reported accurately reflects the nature of the interaction (i.e. very good or very bad). Another plausible interpretation is that

participants became progressively more likely to generate reports for both the web and voice interfaces, for reasons irrespective of task performance.

3.5.2 Total Negative Critical Incidents

Analyses were conducted to investigate the effect of the interface type, day, and treatment conditions, as well as their interactions, on the total number of negative critical incidents reported by each participant. ANOVA results corresponding to Expert 1 and Expert 2 data are shown in Table 3-23 and Table 3-24, respectively.

Table 3-23. ANOVA Table of Total Negative Critical Incidents Using Expert 1 Data

Source	**DF**	**SS**	**MS**	**F**	**P**
Between					
Treatment (T)	2	11.4200	5.7100	8.35	0.002*
S/T	27	18.4600	0.6837	**	
Within					
Day (D)	4	18.2133	4.5533	7.93	0.000*
D*C	8	9.1467	1.1433	1.99	0.054
D*S/T	108	62.0400	0.5744	**	
Interface (I)	1	37.4533	37.4533	52.29	0.000*
I*T	2	5.4067	2.7033	3.77	0.036*
I*S/T	27	19.3400	0.7163	**	
I*D	4	32.3467	8.0867	15.78	0.000*
I*D*T	8	6.0933	0.7617	1.49	0.171
I*D*S/T	108	55.3600	0.5126	**	
Total	**299**	**275.2800**			

Table 3-24. ANOVA Table of Total Negative Critical Incidents Using Expert 2 Data

Source	DF	SS	MS	F	P
Between					
T	2	5.1200	2.5600	4.24	0.025*
S/T	27	16.3100	0.6041	**	
Within					
D	4	13.2467	3.3117	10.18	0.000*
D*T	8	4.4133	0.5517	1.70	0.108
D*S/T	108	35.1400	0.3254	**	
I	1	14.9633	14.9633	23.29	0.000*
I*T	2	0.9867	0.4933	0.77	0.474
I*S/T	27	17.3500	0.6426	**	
I*D	4	17.8867	4.4717	13.45	0.000*
I*D*T	8	3.4133	0.4267	1.28	0.260
I*D*S/T	108	35.9000	0.3324	**	
Total	**299**	**164.7300**			

In the analysis using Expert 1 data, the main effects of Treatment, Day, and Interface were all significant at $p = 0.002$, $p < 0.0001$, and $p < 0.0001$, respectively. Also significant was the two-way interaction of Interface x Treatment at $p = 0.036$ and the two-way interaction of Interface x Day at $p < 0.0001$.

Significant main effects of Treatment, Day, and Interface were also found in the analysis of Expert 2 data, albeit at different p-values ($p = 0.025$, $p < 0.0001$, and $p < 0.0001$, respectively), as was a significant two-way interaction of Interface x Day ($p < 0.0001$). A significant interaction of Interface x Treatment was not found.

Significant main effects and interactions of the overall analyses were isolated using Newman-Keuls post hoc results presented in Appendix D.

Main Effect of Treatment

The number of negative critical incidents reported by Expert 1 was significantly higher than that reported by each of the two user reporting groups. In contrast, Expert 2 reported significantly fewer negative critical incidents than did the laboratory-based users. Similar findings were found in the analysis of the total number of critical incidents and illustrate the variability among usability experts.

Interestingly, the number of negative critical incidents reported by lab-based users was not significantly different than that reported by remote-based users. This finding lends support to the hypothesis of comparable data collected using remote versus laboratory-based evaluation. Figure 3-7 illustrates the differences between the mean numbers of negative critical incidents reported per treatment condition.

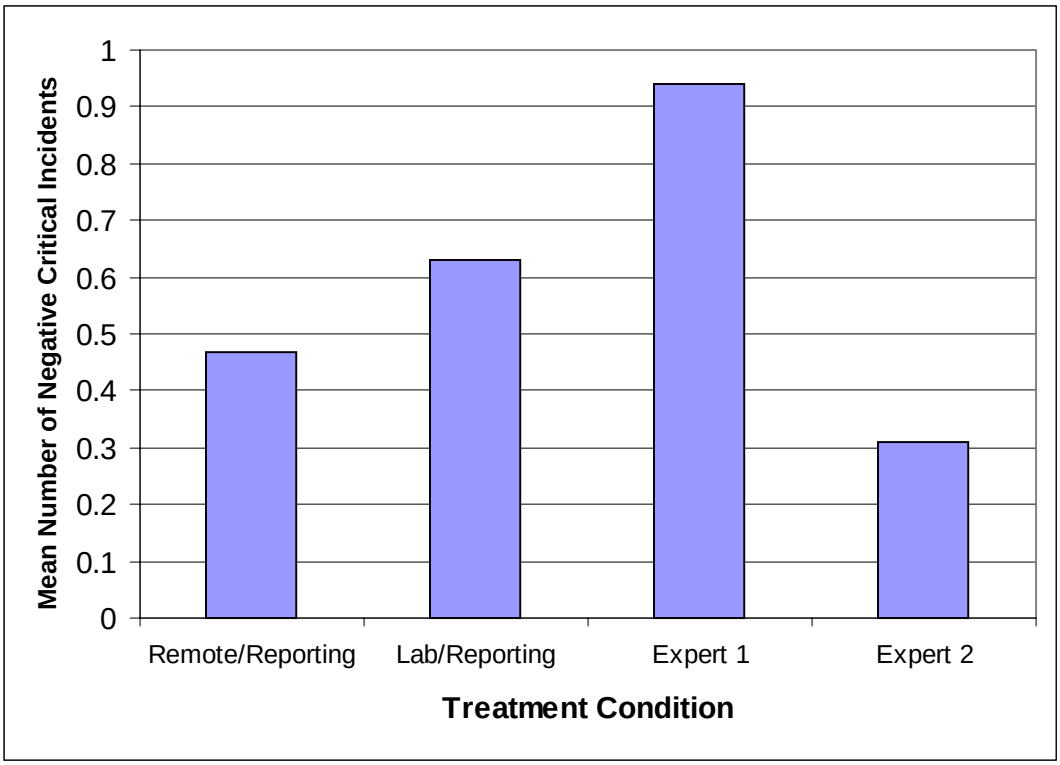

Figure 3-7. Mean Number of Negative Critical Incidents Reported Per Treatment Condition

Main Effect of Interface Type

The mean number of negative critical incidents reported per interface is presented in Table 3-25 for Expert 1 data and in Table 3-26 for Expert 2 data.

Table 3-25. Mean Number of Negative Critical Incidents Reported Per Participant Per Interface (Expert 1 Data)

Interface	N	Mean	St. Dev
Web	150	0.3267	0.5964
Voice	150	1.0333	1.1138

Table 3-26. Mean Number of Negative Critical Incidents Reported Per Participant Per Interface (Expert 2 Data)

Interface	N	Mean	St. Dev
Web	150	0.2467	0.5900
Voice	150	0.6933	0.8106

It can be concluded that a significantly higher number of negative critical incidents were reported for the voice interface than for the web interface and that this finding holds true across analyses of both usability expert data sets. A similar result was found in the analysis of the total number of critical incidents, and suggests that differing skill levels may have led to a greater number of problems encountered in the use of the two interfaces.

Main Effect of Day

Newman-Keuls results differed depending on whether Expert 1 or Expert 2 data sets were used.

Expert 1 Data

The results indicate that the numbers of critical incidents reported on Day 1, Day 2, Day 3, and Day 4 were each significantly higher than those reported on Day 5. In addition, those reported on Day 2 were significantly higher than those reported on Day 3. Figure 3-8 illustrates changes in the mean number of critical incidents reported over the duration of the evaluation. A general decrease in numbers is observed, as was hypothesized, with a slight deviation noted on Day 4. This rise may be attributed to the fact that some tasks assigned on Day 4 were novel, and hence skills and knowledge developed from previous interactions could not be applied. As a result, more problems may have been encountered.

Expert 2 Data

In the analysis of Expert 2 data, the number of negative critical incidents reported on Day 1 and Day 2 was found to be significantly higher than that reported on Day3, Day 4, and Day 5. As shown in Figure 3-9, a general decrease in problems appeared to take place as exposure to the interface increased, thereby lending support to the research hypothesis. Discrepancies between Figure 3-8 and Figure 3-9 can be attributed to variability among the usability experts with respect to the number of negative critical incidents they reported per day.

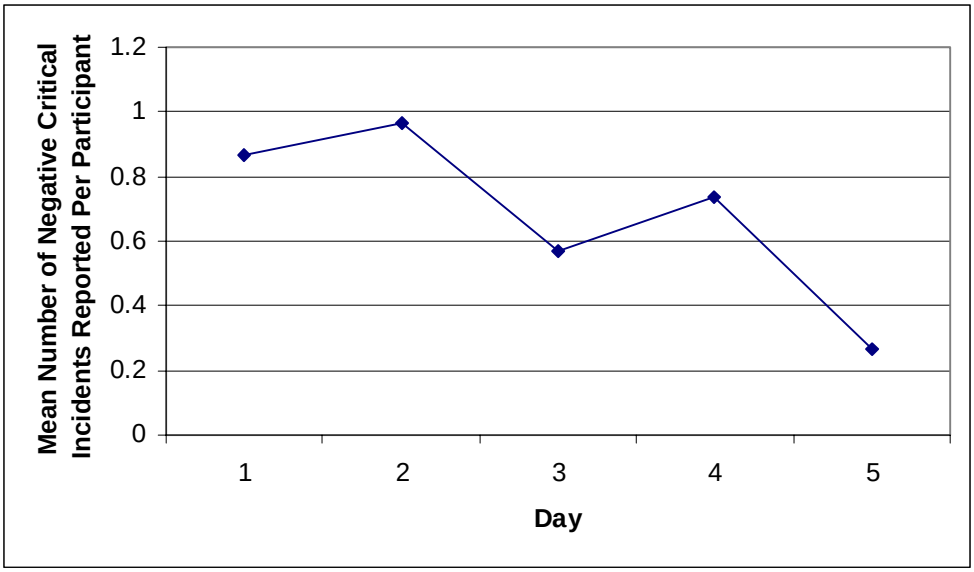

Figure 3-8. Mean Number of Negative Critical Incidents Reported Daily (Expert 1 Data)

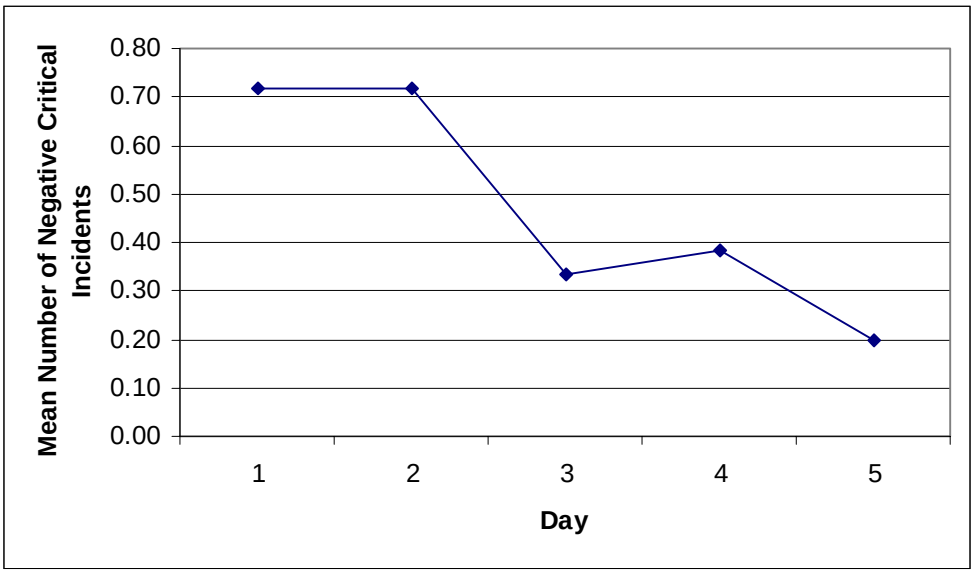

Figure 3-9. Mean Number of Negative Critical Incidents Reported Daily (Expert 2 Data)

Interaction of Interface and Treatment

A significant interaction of the Interface and Treatment factors was found in the analysis of Expert 1 data only. As illustrated in Figure 3-10, Expert 1 reported a significantly larger number of problems, and specifically voice-related problems, than did any other treatment group. It is unlikely that this result is

attributable to the non-reporting participants experiencing more difficulties on average than the other user participant groups. Rather, the differences may reflect different reporting strategies used by the usability expert versus the reporting user participants, with the latter reporting only those incidents that were personally problematic. The usability expert, not being privy to this information, may have falsely identified several critical incidents.

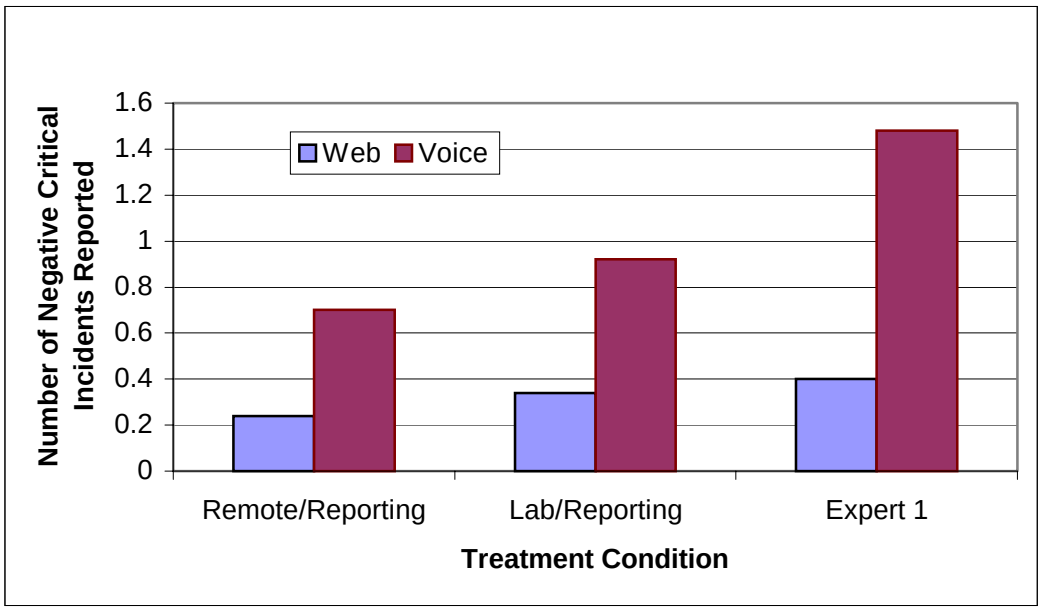

Figure 3-10. Interaction of Treatment and Interface on Number of Negative Critical Incidents (Expert 1 Data)

Interaction Between Interface and Day

Similar results for the Interface and Day interaction were found in the Newman-Keuls analysis of the usability expert data sets. As shown in Figure 3-11 and Figure 3-12, the number of negative critical incidents for the voice interface exceeded those for the web interface on each day of the evaluation, as expected by the significant main effect of interface. The data, however, also reveals differences in the way in which the number of negative critical incidents reported for the voice and web interface change over the 5-day evaluation period. Problems with the web interface decrease dramatically with time while those encountered with the voice interface appear to peak during the second and fourth day. An explanation of these differences requires consideration of the tasks performed on each day of the

evaluation. On the first day, participants were assigned the task of creating and configuring a new VEMS account, all of which required interaction with the web interface. In contrast, only a short exposure to the voice interface was given, providing a greater opportunity for problems to occur with the web interface than the voice interface. On the second day, participants were given their first opportunity to carry out more complex interactions with the voice interface, accounting for the significantly larger number of problems reported on this day. It is most likely that this number was greater due to a general lack of familiarity with voice versus web interfaces. The second peak in voice-related problems on Day 4 was unexpected, but likely attributable to the nature of the scenario assigned on that day. On Day 4, users were required to perform some novel tasks, such as removing messages from a particular sender. Negative critical incidents may be more likely to occur during the performance of novel tasks than practiced tasks, resulting than average number of negative critical incidents being reported. It is interesting to note that the Day 3 scenario was comprised entirely of tasks for the voice interface and yet the number of problems reported is lower than that on the previous and following day. This suggests that either the task requirements capitalized on knowledge and skills already developed or that a drop in motivation to report a critical incident occurred towards the middle of the evaluation period.

Changes in web-related problems over time are also of interest. Results show that number of web-related critical incidents reported on Day 1 is significantly greater than that reported on Day 3 and that in general, this number decreases with time. This result provides support for the notion that with increasing exposure to the (web) interface, user familiarity and proficiency increase, resulting in a fewer number of problems encountered. However, the difference could equally be attributed to a decrease in motivation with respect to reporting a web-related interface problem, although the absence of a similar observation for voice-related critical incidents makes this explanation difficult to justify.

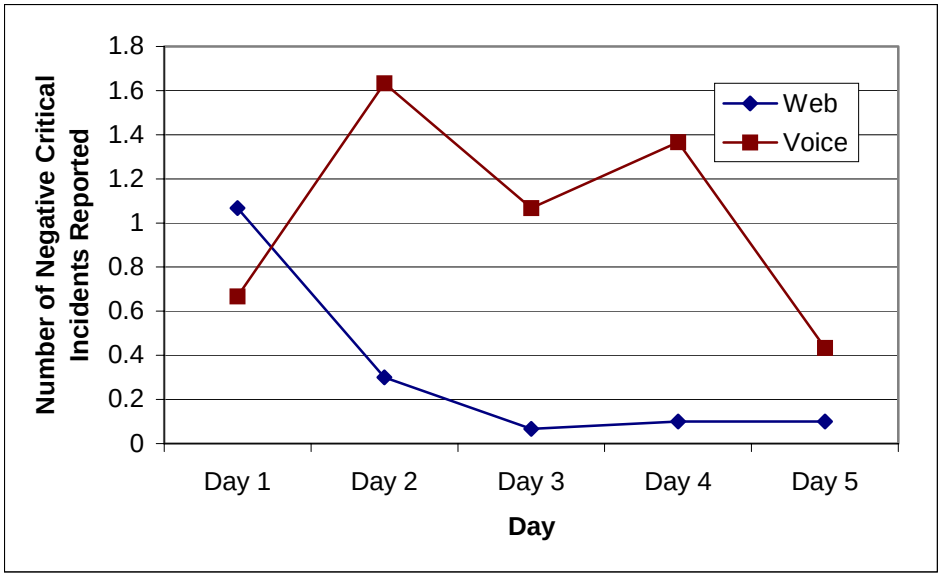

Figure 3-11. Interaction Plot for Interaction of Interface and Day on Number of Negative Critical Incidents (Expert 1 Data)

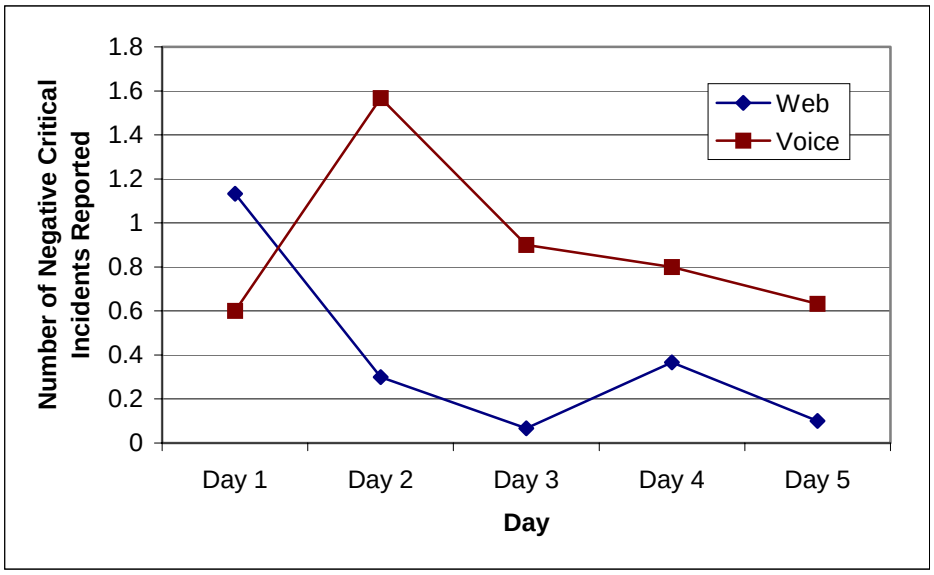

Figure 3-12. Interaction Plot for Interaction of Interface and Day on Number of Negative Critical Incidents (Expert 2 Data)

3.5.3 Total Positive Critical Incidents

Analyses were conducted to investigate the effect of the interface type, day, and treatment conditions, as well as their interactions, on the total number of positive critical incidents reported by each participant.

ANOVA results corresponding to Expert 1 and Expert 2 data are shown in Table 3-27 and Table 3-28, respectively.

Table 3-27. ANOVA Table of Total Positive Critical Incidents Using Expert 1 Data

Source	**DF**	**SS**	**MS**	**F**	**P**
Between					
T	2	17.6867	8.8433	24.24	0.000*
S/T	27	9.8500	0.3648	**	
Within					
D	4	2.4867	0.6217	2.36	0.058
D*T	8	5.3133	0.6642	2.53	0.015*
D*S/T	108	28.4000	0.2630	**	
I	1	0.0300	0.0300	0.05	0.823
I*T	2	2.9400	1.4700	2.49	0.102
I*S/T	27	15.9300	0.5900	**	
I*D	4	16.2867	4.0717	14.50	0.000*
I*D*T	8	7.9933	0.9992	3.56	0.001*
I*D*S/T	108	30.3200	0.2807	**	
Total	**299**	**137.2367**			

Table 3-28. ANOVA Table of Total Positive Critical Incidents Using Expert 2 Data

Source	DF	SS	MS	F	P
Between					
T	2	6.32667	3.16333	12.17	0.000*
S/T	27	7.02000	0.26000	**	
Within					
D	4	0.24667	0.06167	0.47	0.760
D*T	8	0.67333	0.08417	0.64	0.745
D*S/T	108	14.28000	0.13222	**	
I	1	0.21333	0.21333	1.14	0.295
I*T	2	1.32667	0.66333	3.54	0.043*
I*S/T	27	5.06000	0.18741	**	
I*D	4	3.95333	0.98833	7.39	0.000*
I*D*T	8	2.00667	0.25083	1.88	0.071
I*D*S/T	108	14.44000	0.13370	**	
Total	**299**	**55.54667**			

In both analyses, the main effect of Treatment and the two-way interaction of Interface x Day were significant each at $p < 0.0001$. A significant two-way interaction of Day x Treatment and three-way interaction were found in the analysis of Expert 1 data only ($p = 0.015$ and $p=0.001$, respectively). The two-way interaction of Interface and Treatment was found to be significant at $p=0.043$ in the analysis of Expert 2 data only.

Significant main effects and interactions of the overall analyses were isolated using Newman-Keuls post hoc analyses. The results of these analyses are presented in Appendix D.

Main Effect of Treatment Group

The results indicate that the number of positive critical incidents reported by Expert 1 was significantly higher than that reported by either of the two user reporting groups. In contrast, Expert 2 reported significantly fewer positive critical incidents than the laboratory-based user group. These findings are consistent with analyses of the total number of critical incidents and the number of negative critical incidents and are indicative of expert variability with respect to the reporting of interface successes. Figure 3-13 illustrates the differences between the mean numbers of negative critical incidents reported per treatment condition.

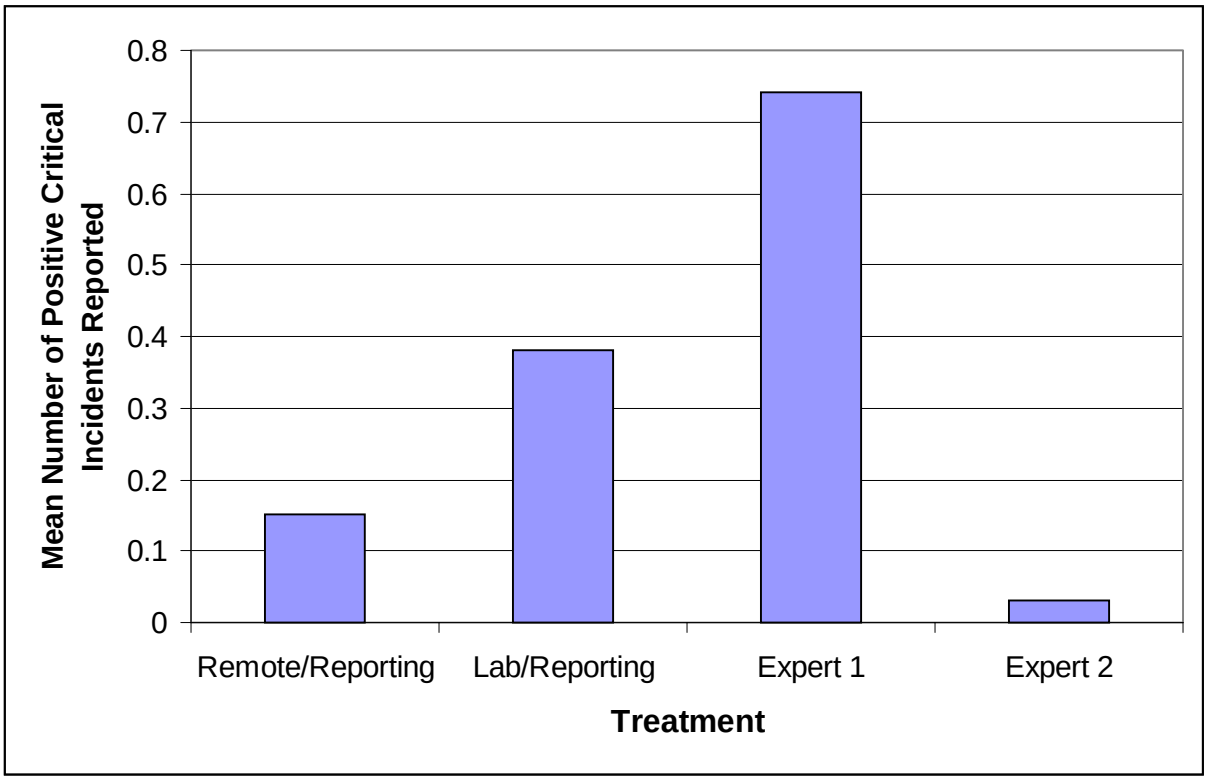

Figure 3-13. Mean Number of Positive Critical Incidents Reported Per Treatment Condition

Another finding was that the number of positive critical incidents reported by lab-based users was significantly different than that reported by remote-based. Specifically, users in a laboratory environment were more likely to report a positive critical incident than remote users. This difference may be linked to the more idealistic environment conditions available in the laboratory (i.e. minimal ambient noise, no distractions, fast computer) or to the increased motivation to balance positive with negative reports while in the presence of the experimenter.

Interaction of Interface and Treatment

A significant interaction of the Interface and Treatment factor was found in the analysis of Expert 2 data only. Figure 3-14 provides a means by which to interpret the results. As shown, a significantly larger number of positive critical incidents were reported by laboratory-based users with respect to the voice interface than for any other combination of treatment and interface levels. Moreover, this treatment group reported a significantly higher number of successes for the voice than the web interface. Other treatment groups reported similar or slightly lower numbers. These findings may be attributed to the fact that the laboratory environment was specifically conducive to the use of a voice email messaging system. For instance, interactions with a voice interface may be highly sensitive to the amount of ambient noise and other audio distractions, both of which were minimized in the lab. The presence of the experimenter may have also acted as a motivator to report more positive critical incidents.

It is interesting to note that the number of successes reported for the web interface were not significantly different between the two user reporting groups. This result negates the presupposition that differences in the number of successes reported by the lab-based versus the remote users were related to differences in the speed of the Internet connection. This presupposition is based on the assumption that remote users likely relied upon modem connections to access the VEMS web interface. As such, they experienced slower connection and loading times than those encountered using the Ethernet connection in the lab,

which may have affected task performance. However, as the results indicate, the nature of the Internet connection does not appear to have adversely affected the number of successes encountered while using the web interface.

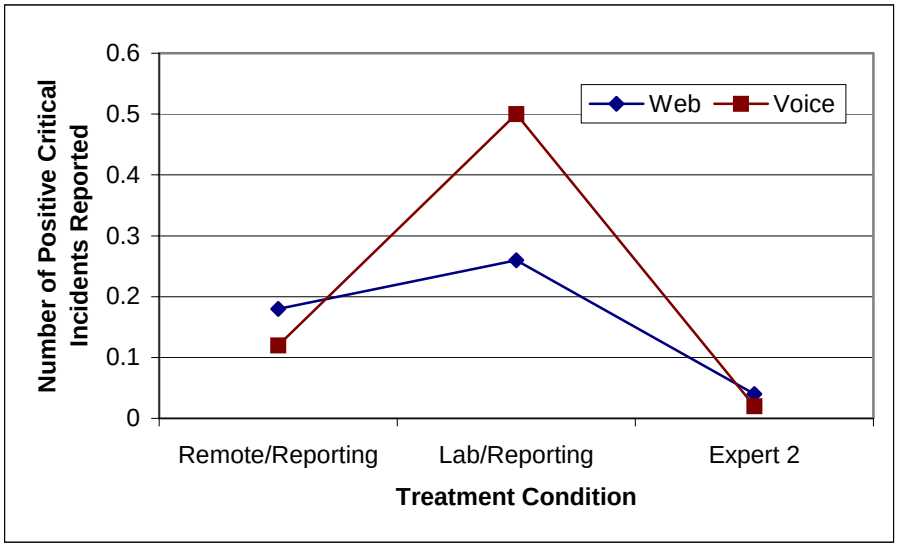

Figure 3-14. Interaction of Interface x Treatment on Number of Positive Critical Incidents (Expert 2 Data)

A second finding is that the laboratory-based users reported a greater number of positive incidents for the voice interface versus for the web interface. This result is surprising since all participants had experience using computers and web browsers, but no or minimal experience with voice email services and voice interfaces in general (suggesting a greater likelihood of success with the web interface). One possible explanation may be that users, having no prior experience with the voice interface, were more aware of successful interactions because they were unexpected. In contrast, prior experience and exposure to web interfaces may have led to the expectation that task performance should be successful, resulting in higher standards for what comprised a critical incident. In addition, users may have developed prior expectations of what comprised a well-designed interface or task structure, making them a more critical user of the web interface than of the voice interface.

Interaction of Day and Treatment

A significant interaction of the Day and Treatment factor was found in the analysis of Expert 1 data only. Figure 3-15 illustrates the changes in the number of positive critical incidents per treatment group over time. On Days 1 and 4, Expert 1 generated a significantly greater number of positive critical incidents than on any other day and by any other treatment group. This particular combination of factor levels may have been conducive to positive critical incidents due to Expert 1 reporting behavior or to actual performance of the non-reporting users being observed. In the former case, the expert's motivation to report a success may have changed over time, with peaks at the start and end of the evaluation. However, since similar changes were not observed for negative critical incidents, there is likely another, more plausible, explanation. For instance, it is possible that users experienced several successes on the first day since they were primarily using the more familiar web interface. As their skill levels with both interfaces increased, so too did the quality and efficiency of their task performance. The expert, via her birds eye view of the user's interactions, may have been able to better able to detect and associate these improvements with a particular aspect of the interface than the user (for whom successful performance was expected). If this explanation is true, it may indicate a possible progression from novice to intermediate skill level after four days of exposure to VEMS.

Interaction of Interface and Day

Newman-Keuls results for the Interface and Day interaction differed slightly across Expert 1 and Expert 2 data sets, and for this reason, are discussed separately.

Expert 1 Data

Figure 3-16 indicates a differential change in the number of negative critical incidents reported for the voice and web interface over the 5-day evaluation period. Whereas successes with the web interface significantly outnumber those with the voice interfaces on the first and fourth day of the evaluation, on

all other days (and especially on Day 3), the relationship is reversed. Furthermore, the overall change in positive critical incidents reported for each interface type differs. Web successes decline over the first three days, and then suddenly increase on the fourth day. In contrast, voice interface successes increase over the first three days and then drop to lower numbers during the remainder of the evaluation.

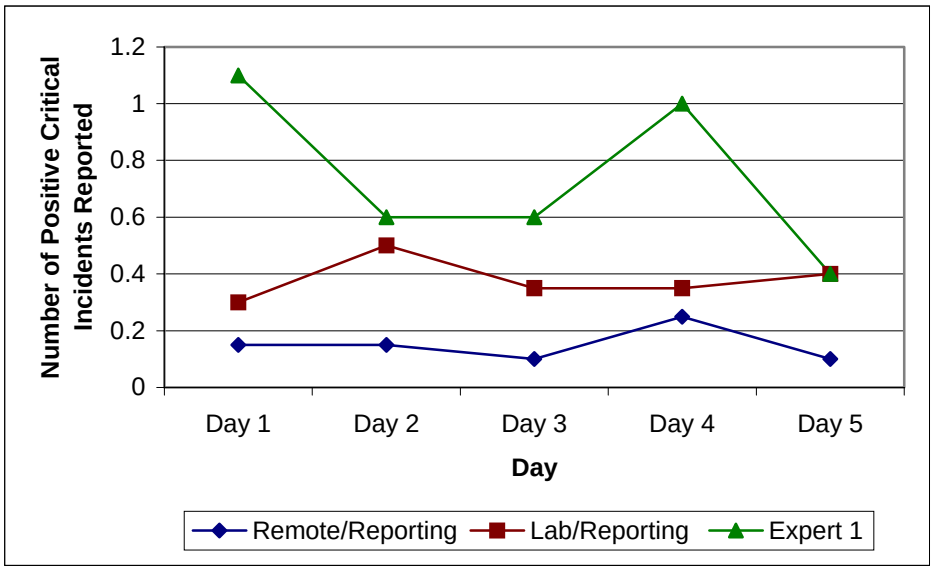

Figure 3-15. Interaction of Treatment x Day on the Number of Negative Critical Incidents (Expert 1 Data)

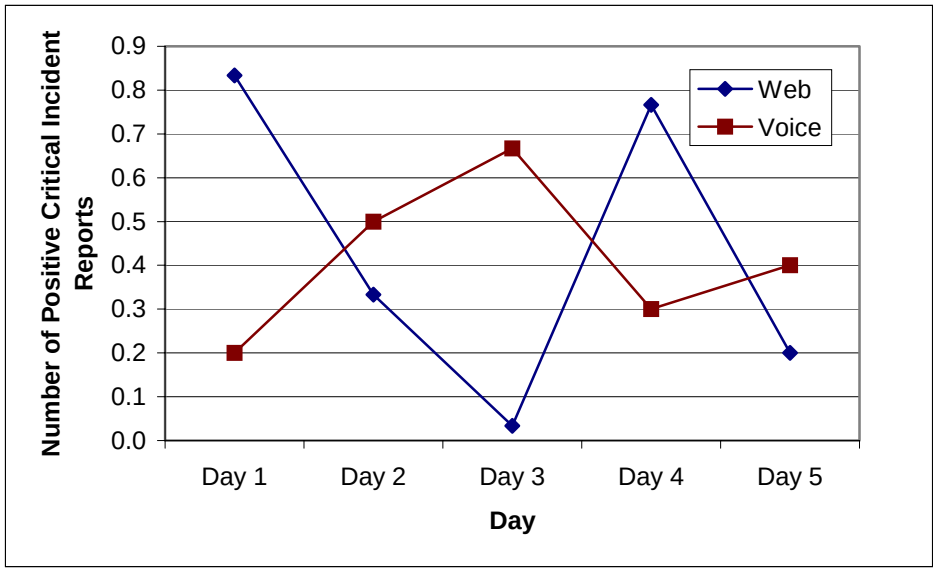

Figure 3-16. Interaction of Interface x Day on Number of Positive Critical Incidents (Expert 1 Data)

An explanation for these differences may relate to participant skill levels prior to the experiment. Pre-test questionnaire data indicated that all participants had at least one year of computer experience and experience using at least one web browser. This past experience may have facilitated interactions with the web interface, increasing the opportunity for positive critical incidents. In contrast, user participants cited minimal experience with voice synthesis and automated speech recognition technologies. Consequently, interactions with the voice interface did not have an existing resource of knowledge or skills upon which to draw, decreasing the likelihood of success. Repeated exposure may have allowed a skill set to develop, resulting in increasingly successful task performance over the course of the evaluation.

Another explanation is based upon the task scenario assigned on each day of the evaluation. On the first day, the majority of tasks required the use of the web interface with only minimal exposure allotted for the voice interface. Therefore, there was a greater opportunity for web-related versus voice-related successes to occur. Over the next two days of the experiment, the tasks focused more heavily on the voice interface, with Day 3 being comprised entirely of tasks related to the voice interface. This greater focus may account for the decline and rise of web and voice successes, respectively. The second peak in web-related successes is unexpected, but may suggest that tasks assigned on Day 4 were successful at capitalizing on existing skills, knowledge, and abilities.

Expert 2 Data

According to the Expert 2 results, a significantly higher number of positive critical incidents were reported on Days 2 and 3 for the voice interface and on Day 1 and Day 4 for the web interface than on Day 3 for the web interface. Figure 3-17 illustrates these differences. The results suggest that throughout the first half of the evaluation period there was an increasingly higher likelihood that a success would be reported for the voice interface versus the web interface until Day 4, when the

relationship reversed. Again, this finding is most likely related to the ratio of tasks assigned to the web versus the voice interface.

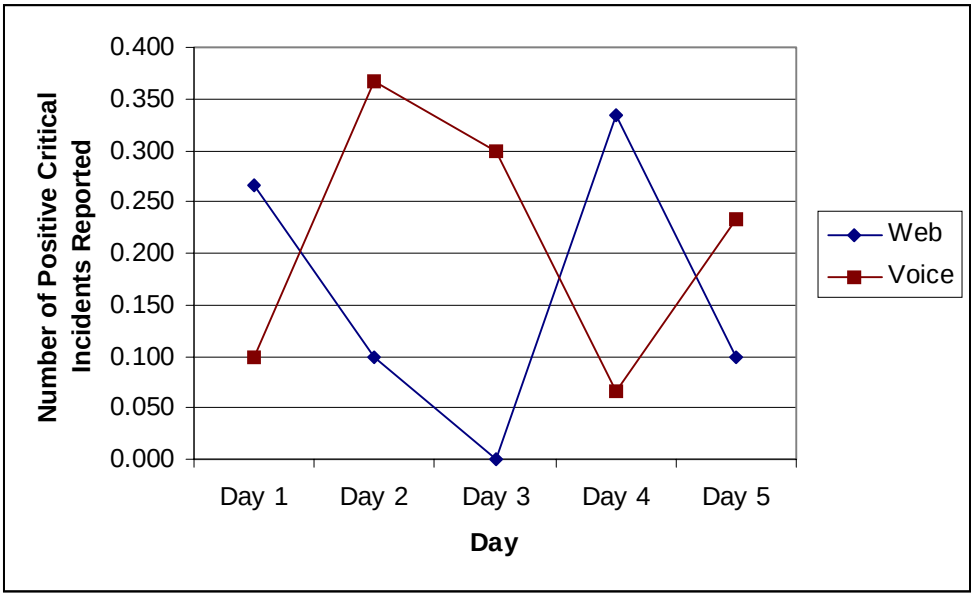

Figure 3-17. Interaction of Interface x Day for Number of Positive Critical Incidents (Expert 2 Data)

Interaction of Treatment x Interface x Day

A significant interaction of the Treatment, Interface and Day factors was found in the analysis of Expert 1 data only. Figure 3-18 provides graphically depicts how the number of positive critical incidents changed for each treatment condition and interface type over time. As shown, there are two significant treatment combinations: Expert 1 reporting on Day 1 and 4 for the web interface and on Day 3 for the voice interface. These results are not unexpected. As previously explained, the design of the task scenarios was such that there was a greater likelihood of critical incidents occurring for the web and voice interface on Day 1 and Day 3, respectively. In addition, isolation of the Treatment main effect showed that Expert 1 reported a greater number of positive critical incidents overall. Finally, variation in the expert's motivation to report a success for the voice interface may account for the significant rise on Day 4.

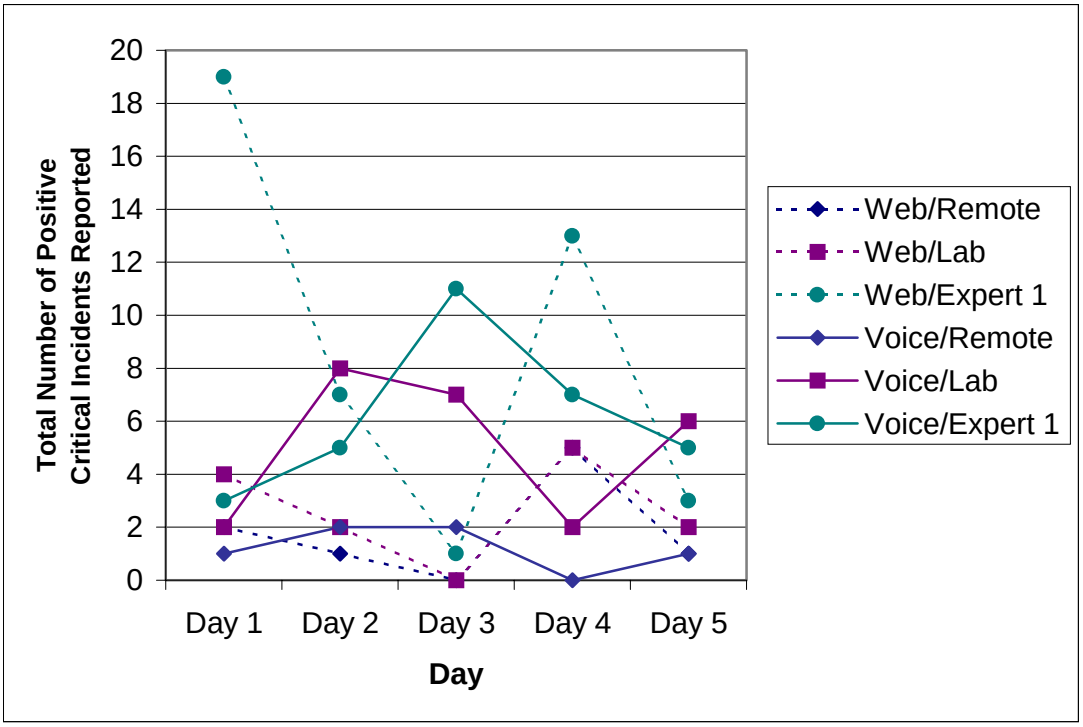

Figure 3-18. Plot of Interaction of Day x Treatment x Interface for Number of Positive Critical Incidents (Expert 1 Data)

3.6 CRITICAL INCIDENT SEVERITY

In this experiment, critical incident severity was measured along several dimensions, including task frequency, impact on task performance and satisfaction, and error severity (negative critical incidents only). Critical incident reporters rated each of these dimensions using an appropriate rating scale. Review of the rating score data set indicated that the statistical analysis approach used for the frequency data was not appropriate. Participants did not necessarily report both negative and positive critical incidents on each day, and only very rarely reported critical incidents pertaining to both the web and voice interface on a daily basis. Consequently, at least one critical incident satisfying every unique combination of factors typically did not exist, resulting in a high number of empty cells. For this reason, it was decided that rating scores generated over each day and for each interface type would be averaged together and compared across Treatment condition only. This approach has several implications.

Namely, it prevents the investigation of changes in severity ratings over the duration of the evaluation period and of differences in severity ratings corresponding to the two interface types.

Analyses were conducted to investigate average rating scores across the treatment conditions. Since rating scores were ordinal data, the Kruskal-Wallis test was selected as an appropriate analysis tool. This test is a non-parametric equivalent to the one-way analysis of variance. It performs a hypothesis test of the equality of population medians for a one-way design (two or more populations). Assumed by this test is that the samples from the different populations are independent random samples from continuous distributions, with the distributions having the same shape.

The Kruskal-Wallis test was performed to test the equality of medians for each of the following average rating score among the treatment conditions indicated in parentheses:

- Task Frequency - Negative and Positive Critical Incidents (all)
- Impact on Task Performance - Negative and Positive Critical Incidents (all)
- Impact on Satisfaction – Negative and Positive Critical Incidents (lab/reporting and remote/reporting)
- Error Severity – Negative Critical Incidents Average rating of severity of error (all)

In each case, separate tests were conducted using Expert 1 and Expert 2 data, respectively. Only one analysis (comparing user reporting participant groups) was required for Impact on Satisfaction Rating since this data was not obtained from the usability experts. It should be noted that all tests were based on the number of critical incidents *as reported*. That is, a count of one was assigned to each critical incident report submitted, whether or not it was determined during classification (see Section 2.8) that a critical incident description contained multiple critical incidents.

3.6.1 Task Frequency

Reporters were required to describe the task that was being carried out at the time of the critical incident. This description included a rating of how frequently the task was typically performed. One of five possible responses could be selected: Very Frequently, Frequently, Occasionally, Rarely, or First and Only Time Anticipated Performing This Task. Task frequency provides a measure of critical incident severity: a critical incident associated with a routine task (ex. reading a message) may be of greater severity than one that occurs during a singular activity (ex. registering for a new account). Analyses for Task Frequency ratings were conducted separately for positive and negative critical incidents. Results are discussed below.

3.6.1.1 Task Frequency Ratings for Negative Critical Incidents

The Kruskal-Wallis test was implemented to analyze the Task Frequency ratings allocated to negative critical incidents. Results differed depending on whether Expert 1 or Expert 2 data was used, as indicated by the data presented in Table 3-29 and Table 3-30, respectively.

Table 3-29. Kruskal-Wallis Test Results for Negative Critical Incident Task Frequency Ratings Using Expert 1 Data

Condition	N	Median	Ave Rank	Z
1	10	3.400	13.3	-0.97
2	10	3.565	14.1	-0.64
3	10	3.805	19.1	1.61
Overall	30		15.5	

H = 2.61 DF = 2 P = 0.271
H = 2.63 DF = 2 P = 0.268 (adjusted for ties)

Table 3-30. Kruskal-Wallis Test Results for Negative Critical Incident Task Frequency Ratings Using Expert 2 Data

Condition	N	Median	Ave Rank	Z
1	10	3.400	11.3	-1.70
2	10	3.565	10.9	-1.88
3	9	4.330	23.7	3.68
Overall	29		15.0	

H = 13.53 DF = 2 P = 0.001
H = 13.65 DF = 2 P = 0.001 (adjusted for ties)

A conclusion of no significance was drawn when task frequency ratings from Expert 1 were compared against those of the user reporting groups. However, the opposite was concluded when task frequency ratings from Expert 2 were considered. Specifically, the null hypothesis of no difference was rejected in favor of the alternative hypothesis of at least one difference among the treatment groups (of which Expert 2 was a member). Post hoc Z-tests were needed to determine between which pairs of treatment groups a significant difference in task frequency ratings existed. Results of this post hoc test are presented in Table 3-31, with significant differences between pairs of treatment groups indicated in bold typeface.

Table 3-31. Post Hoc Comparison of Negative Critical Incident Task Frequency Ratings Using Expert 2 Data

u	v	Ru-Rv	Z-statistic
1	2	0.4	9.425
1	3	**12.4**	9.683
2	3	**12.8**	9.683

According to the post hoc comparison results, Expert 2 reported higher task frequency ratings for negative critical incidents than did either of the two user reporting groups. This result may be attributed differences in Expert 2 versus user reporter performance. For example, Expert 2 may have been more selective in what problems were reported, choosing only those that affected common tasks. It is also plausible that Expert 2 was biased towards critical incidents associated with frequently performed tasks, since they were more salient than those related to rare or frequently occurring tasks. The differences in ratings may also be attributable to the inability of user participants to accurately predict task frequency without knowledge of future task scenarios. Expert 2 not only had access to this information, but could also use prior observations to accurately determine how often a particular task would occur.

3.6.1.2 Task Frequency Ratings for Positive Critical Incidents

Analysis of task frequency ratings allocated to positive critical incidents groups is limited to Expert 1 data. An insufficient number of positive critical incidents were reported by Expert 2 to permit analysis. Table 3-32 presents the results of a Kruskal-Wallis test of the effect of treatment condition on task frequency ratings.

Table 3-32. Kruskal-Wallis Test Results for Positive Critical Incident Task Frequency Ratings for Expert 1 Data

Condition	N	Median	Ave Rank	Z
1	9	4.000	13.6	-0.61
2	10	4.000	16.1	0.53
3	10	3.880	15.2	0.07
Overall	29		15.0	

H = 0.44 DF = 2 P = 0.801
H = 0.46 DF = 2 P = 0.795 (adjusted for ties)

The results provide insufficient evidence that the medians are not all equal ($\alpha = 0.05$). In other words, treatment groups did not significantly differ with respect to the way in which they rated the frequency of tasks during which positive critical incidents occurred.

3.6.2 Impact on Task Performance

Impact on task performance is a second measure of critical incident severity: the greater the impact, the greater the perceived severity of the incident. Users were required to specify the extent to which the critical incident impaired or facilitated task performance by selecting one of five possible options. The options provided differed depending on whether a negative or positive critical incident was being reported, as shown in Table 3-33.

Table 3-33. Impact on Task Performance Rating Options

Negative Critical Incident Rating Options	Positive Critical Incident Rating Options
• Unable to complete the task • Task completed with major difficulties • Task completed with minor difficulties • Task completed with negligible difficulties • Critical incident had no effect on my task performance	• Extremely large increase in speed/accuracy/ease. • Large increase in speed/accuracy/ease. • Moderate increase in speed/accuracy/ease. • Negligible increase in speed/accuracy/ease. • Critical incident had no impact on my satisfaction.

3.6.2.1 Impact on Task Performance Ratings for Negative Critical Incidents

The Kruskal-Wallis test was carried out to analyze Impact on Task Performance ratings allocated to negative critical incidents using Expert 1 and Expert 2 data. The results of each of these analyses are presented in Table 3-34 and Table 3-35, respectively.

Table 3-34. Kruskal-Wallis Test Results for Impact on Task Performance Ratings (Negative Critical Incidents) Using Expert 1 Data

Condition	N	Median	Ave Rank	Z
1	10	2.890	12.9	-1.14
2	10	2.730	8.7	-2.99
3	10	3.420	24.9	4.14
Overall	30		15.5	

H = 18.24 DF = 2 P = 0.000
H = 18.27 DF = 2 P = 0.000 (adjusted for ties)

Table 3-35. Kruskal-Wallis Test Results for Impact on Task Performance Ratings (Negative Critical Incidents) Using Expert 2 Data

Condition	N	Median	Ave Rank	Z
1	10	2.890	13.2	-0.85
2	10	2.730	8.8	-2.82
3	9	3.500	23.9	3.77
Overall	29		15.0	

H = 15.50 DF = 2 P = 0.000
H = 15.61 DF = 2 P = 0.000 (adjusted for ties)

Results from each analysis indicate that the null hypothesis (no difference) can be rejected at levels higher than 0 in favor of the alternative hypothesis of at least one difference among the treatment groups. Post hoc Z-tests were conducted to isolate the significant differences for each expert data set. Results of the post hoc tests using Expert 1 and Expert 2 data are presented in Table 3-36 and Table 3-37, respectively. Significant differences between pairs of treatment groups are denoted by bold typeface.

Table 3-36. Post Hoc Comparison of Impact on Task Performance Ratings Using Expert 1 Data

u	v	Ru-Rv	Z-statistic
1	2	4.2	9.425
1	3	**12**	9.425
2	3	**16.2**	9.425

Table 3-37. Post Hoc Comparison of Impact on Task Performance Ratings Using Expert 2 Data

u	v	Ru-Rv	Z-statistic
1	2	4.3	9.425
1	3	**10.9**	9.683
2	3	**15.2**	9.683

Post hoc comparison result indicate that expert ratings of the impact on task performance were significantly higher than those reported by both the remote and laboratory-based user participants, irregardless of whether data from Expert 1 or Expert 2 was used. This result provides support against the research hypothesis that user reporter critical incident data is comparable that of user experts. The discrepancy may be a result of experts overestimating the true impact on task performance, perhaps due to a lack of appropriate information or feedback from the user. Alternatively, users may tend to underestimate impact on task performance due to overconfidence in their ability to address the problem. Another explanation may be that the users observed by the two experts encountered more impactful incidents than did other users groups, although control mechanisms implemented (including fixed task scenarios and screening mechanisms) makes this the less likely explanation of the three.

3.6.2.2 Impact on Task Performance Ratings for Positive Critical Incidents

Analysis of impact on task performance ratings allocated to positive critical incidents groups is limited to Expert 1 data. An insufficient number of positive critical incidents were reported by Expert 2 to permit analysis. Table 3-32 presents the results of a Kruskal-Wallis test of the equality of impact on task performance ratings.

Table 3-38. Kruskal-Wallis Test Results for Positive Critical Incident Impact on Task Performance Ratings for Expert 1 Data

Condition	N	Median	Ave Rank	Z
1	9	4.000	21.2	2.62
2	10	3.500	17.2	1.01
3	10	3.000	7.2	-3.56
Overall	29		15.0	

H = 13.67 DF = 2 P = 0.001
H = 14.11 DF = 2 P = 0.001 (adjusted for ties)

Results show a significant main effect of treatment condition at p=0.001. Therefore, it can be concluded that there was a significant difference in the ratings for impact on task performance between treatment groups, of which Expert 1 was a member. A post hoc Z-test was conducted to isolate the significant

differences. Results of the post hoc tests are presented in Table 3-39 with significant differences between pairs of treatment groups denoted by bold typeface.

Table 3-39. Post Hoc Comparison of Positive Critical Incident Impact on Task Performance Ratings for Expert 1 Data

u	v	Ru-Rv	Z-statistic
1	2	4	9.425
1	3	**14**	9.425
2	3	**10**	9.425

Post hoc comparisons show that the ratings assigned by Expert 1 were significantly lower than those assigned by the laboratory-based and remote users. The opposite was found in the analysis of task impact ratings for negative critical incidents, in which experts were shown to give significantly higher ratings than user reporters. The lower ratings may reflect difficulties in having an outside observer accurately rate the impact of a successful aspect of the interface. Insufficient information or the inability to distinguish what elements of good task performance are attributable to the success of the interface may exacerbate these difficulties. Furthermore, there may be a difference in the ways in which users and experts perceive the utility of a critical incident report or interpret the rating scale. Users may see the report as a means by which to voice comments regarding the interface and hence may be more likely to be praiseworthy towards the interface (ex. “The interface was very easy to use and really helped me perform the task”). In other words, user ratings may capture the affective aspect of the interface. Experts, however, may perceive the report as a systematic means by which to document or characterize user versus system performance and hence base ratings on objective measures only.

3.6.3 Impact on Satisfaction

Not only may task performance be influenced by the occurrence of a critical incident, but so too may be the user’s overall impression or satisfaction with the interface. For this reason, satisfaction becomes another useful means by which to evaluate the severity of a critical incident, assuming that the more problematic or successful an incident, the greater its impact on satisfaction. Reporters were asked to rate

impact on satisfaction as extreme, high, moderate, negligible, or none. The same rating scale was used for positive and negative critical incidents. Due to the lack of objective and observable means by which to rate user satisfaction, impact on satisfaction ratings were not collected from usability experts. For this reason, analyses are limited to the comparison of ratings across user participants treatment groups (remote user reporters and lab user reporters). Separate analyses were conducted for satisfaction ratings associated with positive and negative critical incidents. Results are presented below.

3.6.3.1 Impact on Satisfaction Performance Ratings for Negative Critical Incidents

The Kruskal-Wallis test was carried out to analyze Impact on Satisfaction Performance ratings allocated to negative critical incidents by laboratory-based and remote users. Results are presented in Table 3-40.

Table 3-40. Kruskal-Wallis Test Results for Positive Critical Incident Impact on Task Performance Ratings for Expert 1 Data

Condition	N	Median	Ave Rank	Z
1	10	3.285	9.6	-0.68
2	10	3.460	11.4	0.68
Overall	20		10.5	

H = 0.46 DF = 1 P = 0.496
H = 0.46 DF = 1 P = 0.496 (adjusted for ties)

The test statistic has a p-value of 0.496, both unadjusted and adjusted for ties, indicating that the null hypothesis cannot be rejected in favor of the alternative hypothesis of a difference among the user reporting groups. This finding of no significance provides support for the notion that data collected from remote locations does not differ significantly from that collected in the lab, thereby providing support for the effectiveness of remote evaluation.

3.6.3.2 Impact on Satisfaction Performance Ratings for Negative Critical Incidents

A similar analysis was conducted for Impact on Satisfaction Performance ratings allocated to positive critical incidents, the results of which are presented in Table 3-41.

Table 3-41. Kruskal-Wallis Test Results for Positive Critical Incident Impact on Task Performance Ratings for Expert 1 Data

Condition	N	Median	Ave Rank	Z
1	9	4.000	9.8	-0.12
2	10	4.000	10.2	0.12
Overall	19		10.0	

H = 0.01 DF = 1 P = 0.903
H = 0.02 DF = 1 P = 0.900 (adjusted for ties)

The test statistic has a p-value of 0.903 unadjusted and of 0.900 adjusted for ties, indicating that the null hypothesis cannot be rejected in favor of the alternative hypothesis of a difference between the user reporting groups. This finding is consistent with the analysis of negative critical incident satisfaction ratings analysis, providing further support for the effectiveness of remote evaluation.

3.6.4 Error Severity

Error severity ratings were allocated to negative critical incidents only and provided a measure of the degree to which error recovery strategies were required. Reporters were encouraged to rate error severity according to the costs incurred (with respect to time, effort, productivity, safety, etc.) from extremely high cost to no associated cost. Table 3-42 and Table 3-43 present Kruskal-Wallis test results of the equality of Error Severity ratings for the three treatment conditions using Expert 1 and Expert 2 data, respectively.

Table 3-42. Kruskal-Wallis Test Results for Error Severity (Expert 1 Data)

Condition	N	Median	Ave Rank	Z
1	10	3.050	15.0	-0.22
2	10	2.815	12.8	-1.19
3	10	3.275	18.7	1.41
Overall	30		15.5	

H = 2.29 DF = 2 P = 0.318
H = 2.30 DF = 2 P = 0.317 (adjusted for ties)

Table 3-43. Kruskal-Wallis Test Results for Error Severity (Expert 2 Data)

Condition	N	Median	Ave Rank	Z
1	10	3.050	12.8	-1.01
2	10	2.815	11.6	-1.58
3	9	3.750	21.3	2.66
Overall	29		15.0	

H = 7.20 DF = 2 P = 0.027
H = 7.23 DF = 2 P = 0.027 (adjusted for ties)

Results differed depending on whether Expert 1 or Expert 2 data was used. There is insufficient evidence of a significant difference in Error Severity ratings between Expert 1 and the user reporter groups ($\alpha = 0.05$). The opposite conclusion is drawn from analysis of Expert 2 data. Results show a significant main effect of Treatment at p=0.027. A post hoc Z-test was conducted to isolate the significant differences, the results of which are presented in Table 3-44 with significant differences between pairs of treatment groups indicated in bold typeface.

Table 3-44. Post Hoc Comparison of Error Severity Ratings (Expert 2 Data)

u	v	Ru-Rv	Z-statistic
1	2	1.2	1.954572
1	3	**8.5**	2.008132
2	3	**9.7**	2.008132

Results show that Expert 2 reported higher error severity ratings than the remote and lab user reporters. Differences in the reporting strategies adopted by Expert 2 may account for this difference. For instance, Expert 2 may have been more selective in what problems were reported, choosing only those that resulted in very severe errors. Another possible explanation is that screen interactions provided insufficient cues to accurately determine the nature of the errors caused by a particular critical incident, causing Expert 2 to overestimate severity. Alternatively, non-usability experts may have underestimated severity due to a lack of foresight regarding the consequences of a critical incident or due to a lack in ability to accurately distinguish between more or less severe errors.

3.7 REPORTING PERFORMANCE DATA

Reporting performance is defined as the ease with which reporters were able to interact with the critical incident report form. It was measured objectively in two ways: by the amount of time required to complete each critical incident and by the numbers of times a participant accessed an information button located on the critical incident report.

A one-way ANOVA was implemented to evaluate the main effect of Treatment on reporting performance data using a level of significance of 0.05. Implementation of a 3x2x5 mixed factor ANOVA was not feasible to an insufficient number of critical incidents that satisfied all combinations of factor levels. The ANOVA results are presented below, using Expert 1 and Expert 2 data as appropriate.

3.7.1 Average Critical Incident Reporting Time

Time to report an incident was measured from the point at which the appropriate critical incident report was opened to the time at which it was submitted by pressing the Submit button. Times to report negative and positive critical incidents were analyzed separately due to differences in the types and number of input fields they contained.

3.7.1.1 Average Time to Report a Negative Critical Incident

The results of a one-way ANOVA for the average time to report a negative critical incident are presented in Table 3-45 and Table 3-46 corresponding to analyses using Expert 1 and Expert 2, respectively.

Table 3-45. ANOVA Results for the Average Time to Report a Negative Critical Incident Using Expert 1 Data

Source	DF	SS	MS	F	P
Condition	2	46918	23459	1.62	0.217
Error	27	391363	14495		
Total	29	438282			

Table 3-46. ANOVA Results for the Average Time to Report a Negative Critical Incident Using Expert 2 Data

Source	DF	SS	MS	F	P
Condition	2	24112	12056	0.76	0.476
Error	26	410333	15782		
Total	28	434445			

There is insufficient evidence at $\alpha = 0.05$ of differences among the treatment means when either Expert 1 or Expert 2 is included. This implies that the time required to report a critical incident differs neither between users and experts nor between remote and laboratory-based users, lending support to research hypotheses.

3.7.1.2 Average Time to Report a Positive Critical Incident

Analysis of the average time to report a positive critical incident is limited to comparisons with Expert 1 data only. An insufficient number of positive critical incidents were reported by Expert 2 to permit analysis. Table 3-47 presents results of the one-way ANOVA.

Table 3-47. ANOVA Results for the Average Time to Report a Negative Critical Incident Using Expert 1 Data

Source	DF	SS	MS	F	P
Condition	2	614	307	0.13	0.881
Error	26	62934	2421		
Total	28	63548			

The F-test p-value of 0.881 indicates that the null hypothesis cannot be rejected – there is insufficient evidence at $\alpha = 0.05$ of differences among the treatment means (of which Expert 1 is a member). This implies that the time required to report a critical incident differs neither between user reporters and Expert 1 nor between remote and laboratory-based users. These findings support the hypotheses that critical incident data collected by users versus experts and by users remote versus local to the laboratory are comparable.

3.7.2 Average Number of Times Help Accessed

Help was available to reporters in the form of Information Buttons located directly on the report form. When a button was pressed, a counter was incremented and a window containing additional instructions and examples was presented to the user. The value of the counter at the time of report submission was sent to the experimenter along with the contents of the critical incident report form. Due to differences in the types and number of input fields contained within the positive and negative critical incident report forms, they were analyzed separately.

3.7.2.1 Average Number of Times Help was Accessed During Negative Critical Incident Reporting

The results of a one-way ANOVA for the average number of times help was accessed during the reporting of a negative critical incident are presented in Table 3-48 and Table 3-49 corresponding to analyses using Expert 1 and Expert 2, respectively.

Table 3-48. ANOVA Results for the Help Accessed Count Accumulated During Negative Critical Incident Reporting Using Expert 1 Data

Source	DF	SS	MS	F	P
Condition	2	0.3306	0.1653	1.88	0.173
Error	26	2.2844	0.0879		
Total	28	2.6150			

Table 3-49. ANOVA Results for the Help Accessed Count Accumulated During Negative Critical Incident Reporting Using Expert 2 Data

Source	DF	SS	MS	F	P
Condition	2	0.2446	0.1223	1.23	0.309
Error	25	2.4806	0.0992		
Total	27	2.7252			

The F-test p-values of 0.173 and 0.309 indicate that the null hypothesis cannot be rejected or that there is insufficient evidence at $\alpha = 0.05$ of differences among the treatment groups when either Expert 1 or Expert 2 is included. This implies that the number of times help was accessed during the reporting of a critical incident differs neither between user reporters and experts nor between remote and laboratory-

based users. These findings support the hypotheses that critical incident data collected by user reporters versus experts and by users remote versus local to the laboratory are comparable

3.7.2.2 Average Time to Report a Positive Critical Incident

Analysis of the average number of times help was accessed during the reporting of positive critical incidents is limited to Expert 1 data. An insufficient number of positive critical incidents were reported by Expert 2 to permit analysis. Table 3-50 presents results of the one-way ANOVA.

Table 3-50. ANOVA Results for the Help Accessed Count Accumulated During Positive Critical Incident Reporting Using Expert 1 Data

Source	DF	SS	MS	F	P
Condition	2	0.2595	0.1298	1.34	0.281
Error	25	2.4244	0.0970		
Total	27	2.6839			

The F-test p-value of 0.281 indicates that the null hypothesis cannot be rejected; that is, there is insufficient evidence at $\alpha = 0.05$ of differences among the treatment means (of which Expert 1 is a member). This implies that the number of times help was accessed during the reporting of a critical incident differs neither between user reporters and Expert 1 nor between remote and laboratory-based users. Again, support for the research hypotheses is gained.

In summary, it can be concluded that the ease with which reporters were able to interact with the critical incident report form in order to report a critical incident, as measured by the average reporting time and the number of times help was accessed, was equivalent across treatment groups.

3.8 QUESTIONNAIRE DATA

Two questionnaires were administered to participants using an on-line form: a training questionnaire and a post-test questionnaire. The purpose of these questionnaires was to solicit subjective ratings regarding the usability of a particular interface. An analysis and discussion of ratings gathered from each questionnaire are presented below.

3.8.1 Training Questionnaire Results

The training questionnaire was administered to all participants that were asked to undergo critical incident training. It was comprised of a series of questions that required the participant to rate some aspect of the training program. Ratings allocated by each participant are presented in Appendix D and summarized in Table 3-51. Participants were also asked to describe the most negative and positive aspects of the training tool. The responses obtained are summarized in Table 3-52 and Table 3-53, respectively, with the frequency of each response indicated.

Since all participants received the same training program, the questionnaire data is not amenable to the comparison of potential training performance differences across the different groups. Rather, its analysis can be used to conduct a formative evaluation of the training tool design, whereby key usability problems can be identified and design recommendations proposed. A first iteration of evaluation and redesign was conducted based on results of the pilot study. A second iteration is beyond the scope of this experiment.

Table 3-51. Summary of Training Questionnaire Rating Data

Question	Average (n=20)
It was simple to use the training tool.	4.35
The training was easy to follow.	4.30
The training tool provided sufficient information.	4.65
The information was easy to understand.	4.55
The organization of information was clear.	4.40
I liked interacting with the training tool.	3.75
The training helped me learn to **identify positive** critical incidents.	4.55
The training helped me learn to **identify negative** critical incidents.	4.60
The training helped me learn to **report positive** critical incidents.	4.55
The training helped me learn to **report negative** critical incidents.	4.60
The material covered by the training tool was sufficient.	4.40
I feel better prepared to **identify** critical incidents after going through the training tool.	4.55
I feel better prepared to **report** critical incidents after going through the training tool.	4.50

Table 3-52. Negative Aspects Reported for the Critical Incident Training Tool

Description	Frequency (n = 29)
Excessive scrolling required	4
Too many browser windows opened at once; windows need to be closed manually - should be automatic	4
Information was repetitive	3
Unclear whether following a link embedded in the body of the training material is optional or mandatory	2
Too many links - difficult to keep track of location in web site; several links resulted in a dead end	2
Long; no indication given to indicate how much training material remains.	2
Unclear whether sample form must be filled out.	2
More (positive) examples are needed.	2
Unclear whether practice exercise was optional or mandatory.	1
Help should be provided on the same page as critical incident report form.	1
Help was not clear.	1
Example should proceed the practice reporting exercise.	1
Difficult to remember a critical incident from the past.	1
More color would be helpful.	1
Font was too large.	1
Example critical incident was not obvious.	1

Table 3-53. Positive Aspects Reported for the Positive Critical Incident Training Tool

Description	Frequency (/36)
Good design - simple layout; good organization; easy to navigate	15
Information is clear and concise; easy to understand.	6
Example was good	3
Review of information helped reinforce the key concepts.	2
Step by step approach is good.	2
No time limit	1
Easy to use	1
Concepts are straightforward	1
Sample critical incident report form was helpful	1
Help was available at all times.	1
Adequate information provided	1
Helped learn how to report a critical incident.	1
Presentation of information is simple and logical	1

3.8.2 Post-Test Questionnaire Data

Supplemented analyses were conducted on data collected from the post-test questionnaires. Post-test questionnaires were administered to all participants and solicited subjective feedback regarding the voice email service voice and web interfaces, the Usability Evaluation Web Site, and the critical incident report form (reporting user participants only). The purpose of collecting this data was two-fold: to determine if there were any differences in the subjective evaluation of the voice email program and critical incident

form across treatment groups and 2) to identify usability problems (or successes) with which to drive a redesign of the training and reporting tools.

The post-test questionnaire contained two types of questions: close-ended questions in the form of Likert rating scales (where 1=strongly disagree and 5=strongly agree) and open-ended questions in the form of text boxes, within which the user was given unrestricted freedom to comment as necessary. Responses to the close-ended ratings questions are summarized in Table 3-54.

Table 3-54. Summary of Post-Test Questionnaire Ratings

	TOTAL		PER TREATMENT GROUP		
Question	**Total Average**	**St. Dev.**	**T1 Average**	**T2 Average**	**T3 Average**
It was simple to use the VEMS web site.	4	0.791	4.1	4.4	4
I was able to complete my tasks quickly using the VEMS web site.	4	0.913	3.7	4.6	4.2
I am able to efficiently complete my tasks using the VEMS web site.	4	0.923	3.7	4.4	4.2
It was easy to learn to use the VEMS web site.	5	0.572	4.1	4.7	4.7
The VEMS web page gives error messages that clearly tell me how to fix problems.	4	0.961	3.75	3.33	3.63
Whenever I make a mistake using the VEMS web site, I recover easily and quickly.	4	1.177	3.7	3.9	3.78
The information (ex. on-line help, on-screen messages) provided by the VEMS web pages is clear.	4	0.718	3.7	3.9	4.3
It is easy to find the information I need using the VEMS web pages.	4	0.819	3.5	4	4.1
The information provided by the VEMS web site is easy to understand.	4	0.596	3.9	4.5	4.5
The organization of information on the VEMS web pages is clear.	4	0.960	3.7	4.2	4.4
I like using the VEMS web site.	4	0.785	3.7	4.2	4.3
The VEMS web site has all the functions and capabilities that I expect it to have.	4	0.845	3.5	4.3	3.9
I am able to complete my emailing tasks quickly using the VEMS voice system.	3	1.202	2.4	3.6	3.2
I am able to efficiently complete my emailing tasks using the VEMS voice system.	3	0.980	2.5	3.3	3.4
It was easy to learn to use the voice system.	4	1.083	3.2	4.5	4.3
The voice system gives error messages that clearly tell me how to fix problems.	3	1.117	3	3.4	2.67
The information (ex. help or prompts) provided by the VEMS voice system is clear.	4	1.015	3	4	3.9
Whenever I make a mistake using the voice system, I recover easily and quickly.	3	1.213	2.8	3.3	3.2
It was easy to navigate within the voice email	3	1.114	2.5	3.2	3.3

	TOTAL		PER TREATMENT GROUP		
Question	**Total Average**	**St. Dev.**	**T1 Average**	**T2 Average**	**T3 Average**
system.					
It is easy to find the information I need using the VEMS voice system.	3	0.858	2.9	3.3	3.5
The information provided by the voice system is easy to understand.	4	0.860	3.7	4.11	3.9
The organization of email messages is clear.	4	0.974	3	3.9	3.6
I like using the VEMS voice system.	3	1.326	2.7	3.8	3.6
The VEMS voice system has all the functions and capabilities that I expect it to have.	3	0.964	3.3	3.5	3.3
It was easy to access the critical incident report form.	4	0.663	4.11	4.2	N/A
It was easy to report critical incidents using the report form.	5	0.607	4.3	4.7	N/A
The questions on the report form were easy to understand.	4	0.671	4	4.3	N/A
The questions on the report form covered sufficient detail concerning the critical incident.	4	0.686	3.8	4.1	N/A
I was motivated to report negative critical incidents.	4	0.852	3.8	4.4	N/A
I was motivated to report positive critical incidents.	4	0.718	3.8	4.4	N/A
It was easy to describe the task I was performing using the Critical Incident Report Form.	3	0.883	3	3.8	N/A
It was easy to describe the critical incident using the Critical Incident Report Form.	4	0.768	3.8	3.8	N/A
It was easy to rate the impact of the critical incidents on task performance.	4	0.562	4	4	N/A
It was easy to rate the impact of the critical incident on satisfaction.	4	0.865	3.6	3.8	N/A
It was easy to rate the severity of errors (negative critical incidents only).	4	0.602	3.7	4	N/A
The rating scales used were appropriate.	4	0.834	3.8	3.8	N/A
I consider the critical incident technique an effective way of evaluating an interface.	4	0.641	4	3.8	N/A
It was easy to navigate through the Usability Evaluation web site.	4	0.737	3.89	3.9	N/A
It was easy to learn to use the Usability Evaluation web site.	4	0.616	4.11	4	N/A
The information provided by the Usability Evaluation web pages is clear.	4	0.885	3.6	4.1	4.6
It is easy to find the information I need.	4	0.651	3.8	4.5	4.6
The information that is provided by the Usability Evaluation web site is easy to understand.	4	0.535	4	4.5	4.4
The organization of information on the Usability Evaluation web pages is clear.	4	0.803	3.7	4.3	4.3
I liked using the Usability Evaluation web site.	4	0.490	4.2	4.5	4.4
The emailing tasks that I was required to perform were realistic.	4	0.699	3.9	4.1	4.5

Of interest was to determine if significant differences in usability ratings of the voice email product, the critical incident report form, and the Usability Evaluation Web Site existed across the different treatment conditions. A one-way ANOVA was carried out (Factor T = Treatment; 3 levels: Remote/Reporting, Lab/Reporting, Lab/Non-reporting) for each question, with the exception of questions pertaining to the Critical Incident Report Form. For these questions, only the Lab/Reporting and Remote/Reporting treatment groups were compared. A parametric analysis was considered a valid approach since 5-point Likert scales were used, which approximate interval scale data. Due to an extra return character contained within the <HIDDEN> HTML tag for order of presentation of the input fields, the responses for certain questions were not submitted and hence not available for analysis. These questions are as follows:

- Question 15: It was easy to use the VEMS voice system.
- Question 32: I like the idea of reporting critical incident information to developers (reporting user participants only).
- Question 48: It was simple to use the Usability Evaluation web site.

The ANOVA tables corresponding to the analysis of each question are provided in Appendix D of this report. Significant main effects of interface were found for 7 out of the 10 questions at p-values indicated in Table 3-56. The mean rating scores are also provided per interface type. The significantly higher rating is in bold typeface (note that 1=strongly disagree and 5=strongly agree).

Table 3-55. Summary of Main Effects of Interface on Post-test Questionnaire Rating Scores

		Treatment Condition			
Question	**p-value**	**T1: Remote/ Reporting**	**T2: Lab/ Reporting**	**T3: Lab/Non-Reporting**	**Newman-Keuls Results**
It was easy to learn to use the VEMS.	p = 0.020	4.1	4.7	4.7	T3 > T1 T2 > T1
The information provided by the VEMS web site is easy to understand	p = 0.028	3.9	4.5	4.5	T3 > T1 T2 > T1
It was easy to learn to use the voice system	p = 0.01	3.2	4.5	4.3	T3 > T1 T2 > T1
The information (ex. help or prompts) provided by the VEMS voice system is clear	p = 0.05	3.0	4.0	3.9	T3 > T1 T2 > T1
It was easy to navigate through the Usability Evaluation web site	p = 0.035	3.6	4.1	4.6	T3 > T1
It was easy to learn to use the Usability Evaluation web site	p=0.007	3.8	4.5	4.6	T3 > T1 T2 > T1
The emailing tasks that I was required to perform were realistic	p=0.041	3.9	4.4	4.7	T3 > T1
I will continue to use VEMS now that I have completed the study	p<0.0001	2.8	4.2	4.5	T3 > T1 T2 > T1
I was motivated to report positive critical incidents" (comparison between lab/reporting and remote/reporting treatment groups only)	p=0.039	3.0	3.8	N/A	T2 > T1

Laboratory-based users (reporting and non-reporting) allocated higher ratings (i.e. indicated stronger agreement) to statements than did remote users, in all but two cases. These results differ from analyses of critical incident report frequency and rating data, which generally failed to show differences among user reporting groups. This discrepancy may be related to differences associated with participating in a remote versus a laboratory environment. For example, users in the lab had the opportunity to make informal comments to the experimenter or request help in the event of a problem. These opportunities may have improved the user's subjective impression of the interfaces. Laboratory based users may also have felt more compelled to give favorable ratings simply by being under indirect observation of the experimenter.

Not only did the post-test questionnaire facilitate comparisons across treatment groups, but also across interface type. Interface comparisons were feasible provided a similar question was asked of both the web interface and voice interface. There were ten such questions, including:

1. I was able to complete my tasks quickly using the [interface type].
2. I am able to efficiently complete my tasks using the [interface type].
3. It was easy to learn to use the [interface type].
4. The [interface type] gives error messages that clearly tell me how to fix problems.
5. Whenever I make a mistake using the [interface type], I recover easily and quickly.
6. The information provided by the [interface type] is clear.
7. It is easy to find information I need using the [interface type].
8. The information provided by the [interface type] is easy to understand.
9. I like using the [interface type].
10. The [interface type] has all the functions and capabilities that I expect it to have.

A one-way ANOVA was conducted to evaluate the effect of interface type (a within-subject factor with two levels) on ratings allocated across all treatment groups for each of the above questions. Results of these ANOVAs are presented in Appendix D.

Significant main effects of interface were found for 7 out of the 10 questions at p-values indicated in Table 3-56. The mean rating scores are also provided per interface type. Significantly higher ratings are denoted by bold typeface. As shown, a significantly higher (more favorable) rating was allocated to the web interface in all cases. These results are in agreement with earlier findings that a significantly higher number of negative critical incidents (unfavorable incidents) were reported for the voice versus the web interface.

Table 3-56. Summary of Main Effects of Interface on Post-test Questionnaire Rating Scores

Note: 1=strongly disagree and 5=strongly agree

		Mean Rating Scores	
Question	**p-value**	**Web Interface**	**Voice Interface**
I am able to quickly complete my tasks using the [interface type].	p < 0.0001	**4.2**	3.1
I am able to efficiently complete my tasks using the [interface type].	p < 0.0001	**4.1**	3.1
It was easy to learn to use the [interface type].	p = 0.007	**4.5**	4.0
Whenever I make a mistake using the [interface type], I recover easily and quickly	p = 0.005	**3.8**	3.1
It is easy to find information I need using the [interface type]	p = 0.001	**3.9**	3.2
I like using the [interface type].	p = 0.004	**4.1**	3.4
The [interface type] has all the functions and capabilities that I expect it to have	p = 0.016	**3.9**	3.4

3.8.3 User Comments

Questionnaire respondents were given the opportunity to describe two aspects of the voice email service (web and voice interface) and the on-line reporting tools (Critical Incident Form and Usability Evaluation Web Site) that they liked and disliked the most. While this data is not amenable to a formal analysis, it provides additional input towards the usability evaluation of the voice email service, as well as supporting a second iteration of on-line reporting tool redesign. Lists of positive and negative aspects reported for the voice email service web interface and voice interface and for the Critical Incident Report Form are presented in Appendix D of this report. Participants were also asked to make comments in response to the statement: "I will continue to use VEMS" and to discuss any issues relevant to their use of the Usability Evaluation web site. These comments are presented in Appendix D of this report.

3.9 SUMMARY OF RESULTS

In this research, the effects of treatment condition, day, and interface type and their interactions on the number of positive and negative critical incidents reported were analyzed. Analyses of the effect of treatment condition on critical incident severity ratings, reporting performance, and post-test questionnaire results were also conducted. A summary of results directly applicable to the research hypotheses and for which consistency across expert performance was found is presented in Table 3-57.

Table 3-57. Summary of Important Results

	RESULTS		
MEASURE	**Effect of Treatment**	**Interaction of Interface x Treatment**	**Effect of Day**
Number of Negative Critical Incidents	**Significant** • Expert 1 reported a significantly greater number of incidents. • Expert 2 reported a significantly lower number of critical incidents. • No difference between remote or laboratory-based users.	**Not Significant**	**Significant** The number of critical incidents decreased over time.
Number of Positive Critical Incidents	**Significant** • Expert 1 reported a significantly greater number of incidents. • Expert 2 reported a significantly lower number of critical incidents. • Laboratory-based users reported a greater number of critical incidents than remote users.	**Significant** Laboratory-based users reported a significantly higher number of positive critical incidents for the voice interface than did remote users.	**Not Significant**
Task Frequency Ratings	**Significant** Expert 2 gave significantly higher ratings than user reporters.		
Impact on Task Performance Ratings	**Significant** Usability experts gave significantly higher ratings for negative critical incidents than user reporters, but significantly lower ratings for positive critical incidents.		
Impact on Satisfaction Ratings	**Significant** No significant differences between remote and laboratory-based user reporters.		
Error Severity Ratings	**Significant** Expert 2 gave significantly higher ratings than user reporters did.		
Time to Report a Critical Incident	**Not Significant**		
Number of Times Help was Accessed	**Not Significant**		
Post-test Questionnaire Results	**Significant** • Laboratory-based users gave significantly more favorable ratings on various usability parameters than did remote users. • Laboratory-based user reported gave a significantly higher rating to the statement “I was motivated to report positive critical incidents” than did remote users.		

CHAPTER 4. FUTURE RESEARCH

The results of this research lend support to the effectiveness of applying critical incidents to remote usability evaluation. The extent of this support is limited due to the variability found between the two usability experts, which exacerbated comparisons between the types of critical incident data collected by usability expert versus user reporters. Other limitations arose from the use of the Usability Problem Inspector and the large number of unique critical incident reports generated. Research is needed to find ways in which to address these limitations. The results of this study also provided useful data, in the form of comparisons between user reporting groups and in the qualitative evaluation of on-line training and reporting tools. Additional research is required to further investigate the implications of these findings and to support of the continued development of remote usability evaluation and the critical incident technique. A detailed discussion of issues and areas recommended for investigation is presented below.

4.1 PROTOCOL FOR IDENTIFYING USABILITY EXPERTS

A limitation of this research was inter-expert variability, the implications of which are multi-fold. First and foremost, it suggests that caution must be taken when generalizing the performance of one usability expert across a population of usability experts. For this reason, it is recommended that additional usability experts be recruited to review the videotapes and report the critical incidents observed. This data can be combined with those gathered in this study to increase the overall size of the expert sample and hopefully capture a more representative sample of the usability expert population.

A second implication is that screening mechanisms related to educational backgrounds and experience and knowledge in the application of usability evaluation methods do not provide adequate means by which to guarantee similar performance. Research is needed to determine more accurate predictors of expert performance. Alternatively, pre-tests can be developed to ensure consistent performance prior to

the evaluation. Tests may be designed whereby potential expert participants are asked to review videotape containing a fixed amount of known critical incidents and the number of incidents correctly identified and reported measured. The content of the reports could also be analyzed to evaluate consistency among experts and the quality of information reported. In this case, quality could be assessed indirectly by measuring the ease and accuracy with which a person unfamiliar with the critical incident can reconstruct the incident based on the information given in the report.

A related implication is that the screening mechanisms and recruiting process used in this study did not accurately identify participants who are usability experts *of the critical incident technique*. This expertise was implied in the use of the term "usability expert" in the context of applying critical incidents to usability evaluation. The variability found between the two experts may imply that a certain level of experience and education in the area of usability evaluation methods may guarantee a particular set of general skills, knowledge, and abilities (KSA's). However, there may be additional KSA's that are needed to expertly apply the critical incident technique. These must be identified, quantified, and either made mandatory or incorporated into the usability expert critical incident training program.

Finally, research is required to determine the extent to which inter-expert usability is related to prior knowledge and skill sets versus to individual differences related to the manipulation of the usability evaluation tools. For instance, the use of a web-based report form that requires textual input may allow for better articulation of critical incidents by one expert than by another. Differences in reporting behavior across alternative reporting modes (ex. web forms, free form text, verbal descriptions) should be investigated and some means of predicting which modes are more suitable to one expert versus the next identified.

4.2 EXPANSION OF THE USABILITY PROBLEM INSPECTOR

The Usability Problem Inspector (Hartson et al., 1999) was adopted in this study to supplement a bottom-up classification of the critical incident reports. The benefits of this approach were demonstrated in its ability to systematically classify critical incidents into high-level groupings. Frequency counts allocated to each grouping were amenable to statistical analysis, allowing differences across treatment groups and interface types to be investigated with respect to where in the interaction cycle critical incident occurred. Data generated using the bottom-up classification process was not amenable to statistical analysis, preventing comparisons across treatment groups and interface types from being evaluated.

In light of these benefits, UPI merits future research. Specifically, it is recommended these tools, which were developed primarily for text-based user interfaces, be expanded to address other interface modalities. This expansion is necessary since interaction with a particular interface type likely generates usability issues specific to it as well as those that are general to all types. For example, the temporal nature of speech output and its concomitant demands on user memory may result in problems in assessment that are not relevant to a text-based interface. Better accommodation of positive critical incidents is also recommended as a future research activity. In both cases, data generated in this study can aid in the expansion process. Usability descriptions pertinent to the voice interface can be analyzed to determine new problem categories within each Interaction Activity grouping or sub-grouping or when feasible, the generality of existing problem categories can be increased through their modification to encompass interface-related issues. A similar analysis of usability success descriptions can be carried out. Data from other studies can be consulted or new studies undertaken to generate additional insight regarding interface-specific usability issues. The end result of these expansion efforts will be a set of generic tools that can be applied with equal success across multiple interface types for the purpose of classifying problems or successes.

4.3 USABILITY SENSITIVITY EVALUATION

The analyses conducted in this research were based on the effects and interactions of treatment condition, day, and interface type on the number of critical incidents reported and on the severity ratings they were assigned. Only a high-level investigation was conducted that addressed *what* critical incident was reported. The Usability Problem Inspector tool was used to perform this investigation by classifying critical incidents according to the Interaction Activity in which they occurred. This analysis was facilitated through the use of the Usability Problem Inspector and indicated that distribution of critical incidents across Interaction Activity varied according to treatment condition (i.e. usability evaluation method). This finding suggests that different usability evaluation methods identified a different set of critical incidents.

It is recommended that additional research be conducted to measure the sensitivity of each set of critical incidents to the overall usability of the VEMS interface. A recommended approach would be to sort each set according to some pre-determined criteria and implement the top ten into three separate redesigns of the VEMS interfaces (one per usability evaluation method). A usability evaluation of the redesigned VEMS interfaces could then be conducted using RECITE and analyses conducted to identify the redesign that led to the greatest reduction in problems and greatest increase in successes. The usability evaluation method associated with that particular redesign could then be concluded as being the most effective at identifying key usability issues. This measure of effectiveness would likely require a significant amount of time and resources to implement. However, it would also provide a more direct and compelling means by which to compare RECITE against other usability evaluation methods.

4.4 ON-LINE TRAINING TOOL REDESIGN

Future work activities are required to follow up on training questionnaire responses that identified the need for modifications of the training tool. For instance, many complaints were voiced regarding the

navigational structure of the training tool, citing the use of too many windows and links. Different web architectures can be investigated to determine an optimal design. Several participants indicated the need for additional examples. Narrated video clips could be developed to provide a more realistic depiction of critical incidents and reporting procedures and embedded within the on-line tool. A comparison with the existing design could then be conducted to determine whether this multi-media approach increases transfer of training.

4.5 CRITICAL INCIDENT REPORT FORM REDESIGN

Additional future work activities are required to follow-up on post-test questionnaire responses that identified the need for modifications of the critical incident report form to better support the reporting task. For example, several complaints were made that the detail demanded by the report form was inappropriate for some lower-severity incidents and detracted from task performance. It is recommended that a bi-level reporting system be investigated as a possible solution to this complaint. With this system, the reporter would be given the option of reporting a quick comment in the event that a more detailed critical incident report is unnecessary (i.e. because the feature or event did not cause extreme impact on task performance) or not feasible (i.e. reporter lacks the necessary information or does not have the time). In the latter case, it may be useful to provide a means by which the reporter could return at a later point to add to the description or expand the quick comment into a complete critical incident report.

Disruption to task flow may also be minimized by integrating the report form directly into the interface being evaluated. This approach may streamline and expedite the reporting process since the participant would no longer be required to access the appropriate web site and then switch between browser windows. Moreover, it may increase a participant's motivation to report an incident, since the reporting mechanism is at all times visible and more easily accessed. Integration would be conducive to the collection of a more rich source of data pertaining to where user was located at the time of the incident,

what feature was being used, what actions were performed, thereby eliminating need for user to provide this information. Since the integrated approach requires that the reporting tool take the same form as the interface being evaluated, it is not valid in the case where comparisons across interface types are to be made.

The design of the critical incident report form has implications both with respect to reporting performance and to the ability to classify incidents into appropriate usability problems or successes. In the current critical incident report form design, task and critical incident descriptions (the information used to classify an incident) were primarily open-ended items. This format, while useful for accommodating various contingencies, increases inter-subject response variability. A future effort could be invested into investigating the use of close-ended questions for task and critical incident descriptions. Guiding this investigation would be the need for questions to be general enough to support the variety of critical incidents that can be encountered and to provide the information necessary to make quick and repeatable classifications of the incidents. It would be worth considering the addition of select lists or checkboxes denoting all possible interface states (ex. all web pages in a web site) or better, only those that are applicable to the current situation (ex. all links accessible from the active web page).

Another possible extension may be a critical incident report form for real-time usability evaluation by usability experts. Currently, the report form is best suited to the scenario in which a usability expert is making observations from videotape that can be stopped and rewound as needed. However, situations may arise in which user interactions cannot be videotaped for future review (i.e. due to cost or time constraints) and must be observed and evaluated directly. Real-time evaluation imposes certain design constraints, including the need to eliminate as many descriptive fields as possible and maximize the number of close-ended questions.

4.6 MOTIVATION AS A KEY TO CRITICAL INCIDENT REPORTING

Occasional differences in the data generated by remote versus laboratory-based users were revealed. For example, it was shown that laboratory-based users report a significantly higher number of critical incident reports than do remote users. Post-test questionnaire results indicate that a key factor in reporting critical incidents may be motivation. Remote users gave significantly lower (i.e. indicated less agreement) to the statement "I am motivated to report positive critical incidents" than did laboratory-based users, although both gave equivalent results to a similar statement addressing negative critical incidents. A laboratory environment may help foster intrinsic motivation on account of being under observation by the experimenter. Users may also tend to be more praiseworthy of the interface in order to appease the experimenter. The potential affect of instructor presence on user behavior is one of the commonly cited drawbacks to conducting usability evaluations, particularly those in the field.

This relationship between motivation and reporting strategies suggests that one means by which to further increase the feasibility of applying critical incidents to remote usability evaluation is to find ways in which to sufficiently motivate participants. One solution may be to make the process of reporting critical incidents a collaborative effort among a group of users interacting with a common interface. This may involve giving group members access to each others' critical incident reports (i.e. via a communal web site) and providing the opportunity to add to these incident or give their own ratings of its impact on task performance and satisfaction. Studies investigating the motivational benefits of collaborative reporting would be required to justify its implementation, although the increase data quality and richness that would likely occur may provide an equally compelling reason to adopt this approach.

4.7 EXPANSION OF THE ROLE OF THE USER

The induction of groupings from the basic critical incident data requires insight, experience, and judgment and is hence a subjective rather than objective process (Flanagan, 1954). Different people may

systematize incidents in different ways (Andersson and Nilsson, 1954). In this study, the experimenter carried out the classification. This approach was the most feasible due to time and resource constraints, but may have been biased due to the experimenter's expert knowledge of the system and exposure to the participants' interactions with the system.

Investigating the feasibility of having participants participate in the classification of their own critical incidents would be a worthwhile effort. Participation may involve the user conducting the classification on their own or with the assistance of the experimenter – studies should investigate the feasibility of both approaches. The UPI may serve as suitable classification tool since it is a more structured, and hence more easily explained. The level of specificity to which users can accurately classify their incidents must be investigated. The value of having users, rather than a usability expert, classify the incidents is substantiated by the difficulties experienced by the author in classifying the critical incident reports generated in this study. For instance, it was found that UPI classifications were sensitive to the way in which critical incident reports were worded, and hence to the ability of the reporter to articulate the exact nature of the critical incident. Having the user participate in classification would eliminate this middle interpretation step and its inherent ambiguities. A similar argument to that for having users report critical incidents can also be applied: the user knows the critical incident best, both in terms of its description but also with respect to where in the interaction cycle it occurred and to what aspect of the interface it can be attributed.

A major issue to be addressed in this research is the best way in which to structure the classification protocol. Classification could be done at the time of reporting or at some other point in time (ex. at the end of the usability evaluation). Both approaches should be investigated. If the former approach is shown to be feasible, then a method of implementing the classification into the reporting process must be developed. In its current form, the UPI is a stand-alone, web-based tool and thus easily integrated within

the Usability Evaluation web site. A potential drawback of this approach is that reporting and classifying must be done on separate interfaces and most likely asynchronously, thereby adding time to the process and causing unacceptable disruption to the task flow.

Alternatively, the UPI could be embedded directly into the critical incident report form. The benefit of this approach is that it integrates reporting and classifying into a single activity. Since many of the decisions involved in these activities are similar (ex. determine the effect or impact of the incident), integration may be easily accomplished and help to eliminate redundancies. However, integration would likely require a redesign and simplification of the UPI structure, which may affect its usability. Implications of this approach with respect to its effect on the time required to fill out a report form and on the additional training required also would merit investigation.

4.8 COMPARISONS ACROSS OTHER TECHNIQUES AND INTERFACE TYPES

In this study, the user-reported critical incident method was compared with traditional laboratory-based usability evaluation. Future empirical studies can be similarly conducted to make comparisons with other usability evaluation methods, and particularly with other remote evaluation techniques, such as remote control evaluation and instrumented remote evaluation. Measures such as cost-effectiveness, level of equipment and development effort required, and the quality of data gathered (determined according to the validity of usability problems and successes identified) could be evaluated.

Other empirical studies can be conducted to assess the feasibility of applying the user-reported critical incident technique to the evaluation of other interface modalities such as haptic interfaces and virtual environments. The use of a standard experimental protocol is recommended to facilitate comparisons across studies. Results of these studies can be used to refine the user-reported critical incident technique, with the goal of developing a feasible, yet all-purpose, usability evaluation method.

CHAPTER 5. CONCLUSIONS

RECITE (the REmote Critical Incident TEchnique) is an adaptation of the user-reported critical incident technique developed by Castillo (1997). This technique requires that trained users, working in their normal work environment, identify and report critical incidents that occur during interaction with a target interface. Critical incident reports are submitted using an on-line reporting tool to the experimenter, who is responsible for their compilation into a list of specific usability issues that can be used to guide a redesign of the target interface. Qualitative support for applying the critical incident technique to remote evaluation has been reported (Hartson, H.R., Castillo, J.C., Kelso, J., and Neale, W.C., 1996; Castillo, 1997).

The overall purpose of this study was to quantitatively evaluate the effectiveness of applying RECITE to the short-term (five-day) remote evaluation of web and voice interfaces. Effectiveness was measured as the extent to which critical incident data generated using RECITE was comparable in frequency and severity to those generated by traditional, laboratory-based applications of the critical incident. Specifically, two laboratory-based applications were investigated, differentiated only by the participant assigned the role of reporting critical incidents (i.e. user or usability expert).

Due to high inter-expert variability, measures of effectiveness with respect to laboratory-based usability expert reporters were limited to those in which similar comparisons to user reporters could be drawn for both experts. There were three such three measures: reporting performance, critical incident type, and ratings of impact on task performance. User participants did not differ significantly from experts in terms of the time to report a critical incident and in the number of times help was accessed during this reporting process (an indirect measure of difficulty in reporting). However, usability experts were found to report fewer than expected critical incidents related to the assessment stage of interaction activity and

gave significantly different ratings for impact on task performance (lower ratings for successes and higher ratings for problems). A revision and re-evaluation of the interface is needed to objectively assess which ratings better reflect the true impact of the critical incidents on task performance. In conclusion, there is limited support for the effectiveness of training users to report critical incidents versus recruiting a usability expert.

Investigation of critical incidents reported by remote users versus laboratory-based users failed to reveal differences in all but one measure. Specifically, laboratory-based users tended to report more positive critical incidents for the voice interface than remote users. Environmental conditions and motivation to report critical incidents may have contributed to these differences and carry implications with respect to training and the choice of interface to be evaluated using RECITE. Another compelling finding was that user reporting groups identified a non-significantly different number of negative critical incidents. This suggests that remote users are equally as capable as users located in a laboratory environment to identify problems. Since usability problems are typically of most value to designers interested in improving their product, this finding has significant implications. Namely, it suggests that having users participate while in their normal work environment can minimize cost and resources without jeopardizing data quality.

Results of this research indicate that critical incident data changes over time. Specifically, the number of problems reported decreases. A corresponding result was not found to apply to positive critical incident reports. This finding suggests the value of collecting data over time rather than in a one-time evaluation, as is typical approach. A longitudinal study may help distinguish between aspects of the interface that cause recurrent problems versus those that cause initial difficulties for naïve users and are subsequently resolved. This distinction may help designers better allocate resources and ideas on ways in which to address the interface features in question.

Efforts to demonstrating the effectiveness of remote usability evaluation are well timed. In today's competitive markets, two trends are emerging. First, interface usability evaluation is becoming increasingly more integral to the success of a product. This suggests the importance of developing techniques that facilitate the usability evaluation. Complicating this development process is a second emerging trend: the expanding distribution and internationalization of customer bases (Hammontree, Weiler, and Nayak, 1994). As a result, access to representative users is becoming more difficult and the ability to capture usage patterns pertinent to actual work environments more critical.

Based on the findings of this study, a set of criteria can be proposed for usability evaluators wishing to apply critical incidents to remote evaluation. This set of criteria adds to those established by Castillo (1997):

1. Real users of the interface should be used.
2. Critical incident training of users should precede evaluation efforts.
3. Users should be located in their own working environment.
4. Users should identify and report their own critical incidents via an on-line critical incident report tool designed to support efficient reporting.
5. Data should be captured in day-to-day task situations over a given amount of time to allow changes in performance to be identified.

On-line reporting and training tools are feasible means by which to address the above criteria. Continued research will provide impetus for refinement of these tools and help build a full suite of tools suitable to applying critical incidents to remote evaluation.

Remote evaluation using the RECITE method provides a cost-effective means by which to address situations in which the users of interest are distributed and remote and cannot be brought into a

laboratory environment. It also permits usage patterns pertinent to actual work environments to be captured. In this research, RECITE was applied to the evaluation of a product that had already been released for public use. The interface was fairly robust and fully operational. These characteristics are considered pre-requisites for an effectively implementation of RECITE. Specifically, the interface being evaluated must be such that an experimenter is not required to assist in its use, as in a walkthrough, since RECITE only supports asynchronous interactions between user and experimenter. Furthermore, the interface must be such that it can be used in a normal working environment and can meet the user's normal work task requirements. In other words, it must not only be high fidelity, but also reliable and robust to ensure that critical incident reports reflect only those issues applicable to the design and architecture of the interface, rather than those related to software bugs or other programming issues. Finally, due to the subjective nature of critical incident reports, RECITE results are best applied in the revision of an interface versus a summative comparison with other prototypes or competitor products. Therefore, it is recommended that RECITE be implemented as a formative usability evaluation method and implemented towards the end of the product life cycle whereupon a high fidelity prototype (or beta release) can be given to a set of real users and used to support their day-to-day tasks and functions.

REFERENCES

Andersson, B-E. and Nilsson, S.G. (1964). *Studies in the Reliability and Validity of the Critical Incident Technique.* Journal of Applied Psychology, 48 (6), 398-403.

Bergman, E. and Johnson E. (1995). *Towards accessible human-computer interaction*. In Advances in Human-computer Interaction, 5. Edited by Jakob Nielsen. Ablex Publishing Corporation: Norwood, New Jersey.

Carroll, J.M., Koenemann-Belliveau, J., Rosson, M.B., and Singley, M.K. (1993). *Critical Incidents and Critical Themes in Empirical Usability Evaluation.* In People and Computers VIII. Proceedings of the HCI'93 Conference, September 1993. Edited by J.L. Alty, D. Diaper., and S. Guest. British Computer Society Conference Series. Cambridge University Press: Cambridge, England. 279-292.

Carroll, J.M., Mack, R.L., Lewis, C.L., Grischkowski, N.L., and Robertson, S.R. (1985) *Exploring a word processor*. Human-Computer Interaction, 1(3), 283-307.

Carroll, J.R., Smith-Kerker, P.L., Ford, J.R., and Mazur-Rimetz, S.A. (1987). *The minimal manual*. Human- Computer Interaction, 3(2), 123-153.

Castillo, J. (1998). *What is remote evaluation?* Remote Evaluation Web Page. <http://hci.ise.vt.edu/~josec/remote_eval/definition.html>. Last updated on November 8, 1998. Last visited on January 31, 1999.

Castillo, J.C. (1997). *The user-reported critical incident method for remote usability evaluation*. Unpublished thesis dissertation. Virginia Polytechnic Institution and State University.

Castillo, J.C., Hartson, H.R., and Hix. D. (1997). *Remote Usability Evaluation at a Glance*. (Technical report from Virginia Tech: TR-97-12).

Charney, D.H. and Reder, L.M. Designing interactive tutorials for computer users. Human- Computer Interaction, 2 (1986), 297-317.

del Galdo, E. M., Williges, R. C., Williges, B. H., and Wixon, D. R. (1987). A critical incident evaluation tool for software documentation. In L. S. Mark, J. S. Warm, and R. L. Huston (Eds.) Ergonomics and Human Factors. New York: Springer-Verlag. 253-258.

Doe, H.L. (1998). *Evaluating the effects of automatic speech recognition word accuracy.* Published thesis dissertation. Virginia Polytechnic Institution and State University.

Fitts, P.M. and Jones, R.E. (1947). *Psychological aspects of instrumented display: Analysis of 270 "Pilot-Error" experiences in reading and interpreting aircraft instruments*. In Selected Papers on human factors in the design and use of control system, edited by H.W. Sinaiko. New York: Dover Publications, Inc. 1-38.

Flanagan, J.C. (1954). *The critical incident technique*. Psychological Bulletin, (July), 51 (4), pp. 326-358.

Gray, W.D. and Salzman, M.C. (1998). *Damaged Merchandise? A review of experiments that compare usability evaluation methods*. Human-computer Interaction, Volume 13. Lawrence-Erlbaum Associates, Inc., 203-261.

Hammontree, M., Weiler, P., and Nayak, N. (1994). *Remote usability testing*. Interactions, (July), 21-25.

Harrison, S.M. (1995). *A comparison of still, animated, or non-illustrated on-line help with written or spoken instructions in a graphical user interface*. Proceedings of CHI'95 Human Factors in Computing Systems, ACM, 82-89.

Hartson, H.R., Castillo, J.C., Kelso, J., and Neale, W.C. (1996). *Remote evaluation: The network as an extension of the usability laboratory.* Proceedings of CHI'96, ACM, 228-235.

Hilbert, D.M. and Redmiles, D.F. (1998) *An approach to large-scale collection of application usage data over the Internet*. Proceedings of the 20th International Conference on Software Engineering (ICSE'98), Forging New Links, Kyoto Japan. Published by IEEE Comp Society, Los Alamitos, CA, USA. 136-145.

Kamm, Candace. *User interfaces for voice applications*. In Voice Communication Between Humans and Machines. Edited by D.B. Roe and J.G. Wilpon. National Academy of Sciences, 422-441.

Karis, D., and K.M. Dobroth. (1991) *Automating services with speech recognition over the public switched telephone network: Human factors considerations.* IEEE Journal on Selected Areas of Communications, 9, 574-585.

Koenemann-Belliveau, J., Carroll, J.M., Rosson, M.B., and Singley, M.K. (1994). *Comparative usability evaluation: critical incidents and critical threads*. Proceedings of CHI'94, Boston, Massachusetts, USA, April 24-28, 1994, ACM, 245-251.

Lane, D. (1999). Introduction to Between-Subject ANOVAs (Chapter 12). HyperStat Online. <http://www.ruf.rice.edu/~lane/hyperstat/A47912.html>. Last modified on: November 2, 1999.

Makhoul, J. and Schwartz, R. (1994). *State of the art in continuous speech recognition.* In Voice Communication Between Humans and Machines. Edited by D.B. Roe and J.G. Wilpon. National Academy of Sciences, 165-198.

Nielsen, J. (1998) *Ten Usability Heuristics.* Heuristic Evaluation. Jakob Neilsen's website. <http://www.useit.com>. Last updated on November 30, 1998. Last visited on February 28, 1999.

Richards, J. *Guide to Writing Accessible HTML*. Adaptive Technology Resource Center. University of Toronto. October 29, 1997. <http://www.utoronto.ca/atrc/rd/html/html.html>. March 1998.

Ruben, J. (1994). *Usability Engineering*. San Diego: Academic Press, Inc.

Schmandt, C. (1984). *Speech synthesis gives voiced access to an electronic mail system.* Speech Technology, 2(3), 66-69.

Schmandt, C. (1994). Voice Communication With Computers - Conversational Systems. New York: Van Nostrand Reinhold.

Shattuck, L.G. and Woods, D.D. (1994). *The critical incident technique: 40 years later*. In Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting, pp.1080-1084.

Strathdee, M. *RIM's wireless pager plugs into email*. Article in the Kitchener-Waterloo Record. Wednesday January 20, 1999. Kitchener, Ontario, Canada.

Swallow, J., Hameluck, D., and Carey, T. (1998). User Interface Instrumentation for Usability Analysis: A Case Study. HCI + Learning Laboratory, University of Waterloo (March). This paper was previously published at Cascon '97 (Toronto, Ontario) Nov, 1997.

Walker, M.A., Fromer, J., Di Fabbrizio, G., Mestel, C., and Hindle, D. (1998). *What can I say?: Evaluating a spoken language interface to email.* Proceedings of CHI'98, ACM, 582-589.

Wiedenbeck, S., Zila, P.L., and McConnell, D.S. (1995). *End-user training: an empirical study comparing on-line practice methods*. Proceedings of CHI '95 Human Factors in Computing Systems, ACM, 74-81.

Williges, R.C., Thompson, J., and Andre, T.S. (1999). Usability Evaluations of a Voice Email System. Technical Report HCIL-99-01. Industrial and Systems Engineering, Virginia Tech.

Yankelovich, N., Levow, G-A., Marx, M. (1995). *Designing SpeechActs: Issues in speech user interfaces*. Proceedings of CHI'95, ACM, 369-376.

Hartson, H.R., Andre, T.S., Williges, R.C., and van Rens, L. (1999, August). The user action framework: a theory-based foundation for inspection and classification of usability problems. In H. Bullinger & J. Ziegler (Eds.), Human-Computer Interaction: Ergonomics and User Interfaces, Vol. 1 (Proceedings of HCI Interational'99), (pp. 1058-1062). Mahway, NJ: Lawrence Erlbaum Associates