Technical Report

Trusting Remote Users… Can They Identify Problems Without Involving Usability Experts?

José C. Castillo
eonBusiness Corporation
7430 East Caley Ave. Suite 200
Centennial, Colorado 80111 USA
jose.castillo@eonbusiness.com

H. Rex Hartson
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061

Technical Report TR-07-06
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061

Trusting Remote Users… Can They Identify Problems Without Involving Usability Experts?

José C. Castillo
eonBusiness Corporation
7430 East Caley Ave. Suite 200
Centennial, Colorado 80111 USA
jose.castillo@eonbusiness.com

H. Rex Hartson
Department of Computer Science
Virginia Tech
Blacksburg, VA 24061

Abstract
Based on our belief that critical incident data, observed during usage and associated closely with specific task performance are the most useful kind of formative evaluation data for finding and fixing usability problems, we developed a Remote Usability Evaluation Method (RUEM) that involves real users self-reporting critical incidents encountered in real tasks performed in their normal working environments without the intervention of evaluators. In our exploratory study we observed that users were able to identify, report, and rate the severity level of their own critical incidents with only brief training.

Keywords
Remote usability evaluation, user-reported critical incident method, critical incident, user-initiated, usability data, contextualized, incident report, post-deployment, contemporaneous reporting

Our Study in a Nutshell
There are several very different approaches to remote usability evaluation—getting field usability data at home—each with its own advantages, drawbacks, benefits, and costs [Castillo & Hartson, 2006] We focused on a user-reported critical incident method because it can identify real usability problems encountered in real work settings.

Our study was not a traditional summative study with statistically-significant results, but an exploratory study with the goal of revealing insight and understanding. The good news from the study is an indication that users with no background in usability engineering or human-computer interaction, and with the barest minimum of training in critical incident identification, can identify, report, and rate the severity level of their own critical incidents. The success of the user-reported critical incident method depends on this important indication. In particular, users were able to identify most of the medium- and high-severity critical incidents, and most of the problems not reported were of low severity.

Users took almost twice as long as we expected (an average of 5 ½ minutes compared to our expected 3 minutes) to make a critical incident report, but few users reported that reporting was intrusive or interfered with work tasks. All of our users reported a satisfaction in being able to report problems to evaluators and designers.

In the second phase we looked at whether evaluators can use remotely reported critical incident data to produce usability problems descriptions. Although data in this phase were too sparse to allow solid conclusions, evaluator subjects did seem generally able to produce usability problem descriptions from the field reports of critical incidents, and most evaluator subjects reported that this task was not difficult.

Although we still believe in the value of video clips showing screen action related to each critical incident, the study did not support this belief. This was likely due to technology limitations, including poor resolution and difficulty in matching video clips with the text of critical incident reports. We expect that newer technology for capturing and annotating screen action (e.g., Camtasia® and Morae® [2006a, 2006b]) will help materialize expected benefits of screen action clips.

Perhaps the most important lesson learned in the study was in relation to the timing of critical incident reports. We had expected our users to report critical incidents as they occurred, while still performing the associated task. Although we even directed users to send a report immediately after encountering a critical incident, there was considerable variation in the timing of critical incident reports. We learned that, in most cases, users preferred to wait until after the task was completed. In general, the more severe the problem, the longer is the delay in reporting.

We now believe that what we call contemporaneous reporting (reporting soon after task performance) has advantages over concurrent reporting (reporting during task performance). The details are almost as fresh in the user's memory, and immediately after task performance is the point in time when the user has the most information about the problem. Additionally, not overlapping reporting with task performance eliminates interference with the task itself.

However, this delay in reporting by our remote users had a significant impact on our RUEM design. We had originally assumed that, in a deployed version of our RUEM, the user would trigger video clip capture by clicking the *Report Incident* button to initiate reporting activity. We could capture, via a delay loop in the screen capture application, the most recent (e.g., the

last two minutes) screen action. Based on our assumption of immediate reporting, we expected this clip would show the screen context for the critical incident being reported. However, the typical delay observed in critical incident reporting meant that video clips rarely were relevant to the critical incident. The unpredictability of timing led us to a redesign, giving the user complete control over the timing of video capture. We hope that users will use this feature to record "re-enactments" of critical incidents, giving evaluators exactly what they need in the clips. Future studies will be required to confirm the simplicity and/or effectiveness of this new design.

Background

Importance of Critical Incident Data

Despite numerous variations in procedures for gathering and analyzing critical incidents [Shattuck and Woods, 1994], researchers and practitioners generally agree on the definition of a critical incident. A critical incident is an event observed within task performance that is a significant indicator of a factor determining success or failure of the task [Andersson & Nilsson, 1964]. In the context of formative usability evaluation, a critical incident is an occurrence during user task performance that indicates something significant about usability, usually a problem to be fixed or a feature that should be considered for redesign.

Critical incident information, in our experience, is arguably the single most important kind of data associated with task performance in the context of formative usability evaluation. These data are perishable and must be captured immediately as they arise during usage. This is a major reason why lab-based usability testing is effective, because it captures exactly that kind of detailed usage data, in the form of particular critical incident data, verbal protocol, and usability problem descriptions.

What Makes an Incident "Critical"?

It's all about the impact on the users' experience with the software, especially negative impact. Although it's nice to observe positive critical incidents, incidents that illustrate positive impact on users, negative critical incidents are the ones that help us find and fix usability problems in interaction designs. Obviously, with software used to control a nuclear reactor, for example, human error while using the interface can have an enormous impact, potentially creating an industrial disaster (e.g., radiation leak, explosion). However, everyday users also think of problems encountered as critical if they affect their job performance or their usage experience in a negative way (e.g., lost data, no way to recover, system crashes).

In sum, we think of a usage incident as critical (important for finding and fixing problems) if it has a negative impact on usability – i.e., on task performance, user effectiveness or efficiency, user errors, safety, ease of learning, retainability, user satisfaction, or even usefulness of functionality.

Can Users Report Their Own Critical Incidents?

It is reasonable to expect that users might be aware of errors and problems as they occur. Dzida et al. [1993] found that, when users watch a videotape of their own task performance, they can verbally identify when they experienced an error. In our own experience we have found that many computer users are aware of errors and problems they encounter and, therefore, we believe that users performing their own everyday tasks

are in a good position to recognize critical incidents and usability problems caused by design flaws in the user interface. This possibility was the working hypothesis of the exploratory study.

Who Should Identify Critical Incidents?
In the original work by Fitts and Jones [1947], the user (aircraft pilot) reported the critical incidents. Flanagan [1954] used trained observers to collect critical incident information while observing users performing tasks. Returning to the original approach, del Galdo, Williges, Williges, and Wixon [1986] involved users in identifying critical incidents. Then, with the emergence of lab-based usability testing, the critical incident identification role reverted to expert observers (usability specialists).

Dzida, Wiethoff, and Arnold [1993] and Koenemman-Belliveau, Carroll, Rosson, and Singley [1994] make the case for maximum flexibility, allowing that identifying critical incidents during task performance can be an individual process by either the user or an evaluator or a mutual process between the user and an evaluator. Our user-reported critical incident method is similar to that of del Galdo et al. in that users do the reporting, but differs in other ways.

User Reported Critical Incident Method

When using our RUEM, users are located in their own working environment and acquire modest Web-based training to identify critical incidents occurring in the normal course of on-the-job task performance. Whenever users encounter usage difficulty, they click on a *Report Incident* button from their Web browser, which activates an instrumentation routine that:

- opens a textual form in a separate window, for users to enter a structured report on the details of the specific critical incident encountered, and
- causes the user's computer to store a screen sequence video clip showing screen activity immediately prior to clicking the button for the purpose of capturing the critical incident and the context of events leading up to it. (This describes the design for video capture as originally intended for regular use. In order to capture more complete data in our exploratory study, screen activity was captured continuously via a scan converter and videotape.)

The resulting package of usability data—the critical incident report and the screen sequence clip taken together—is called a contextualized critical incident report, sent asynchronously via the network to evaluators to be analyzed into usability problem descriptions that designers use to drive redesign solutions to improve the interaction design.
Because of the vital importance of critical incident data and the opportunity for users to capture it, we developed this method [Castillo, 1997] for capturing critical incident data and satisfying the following situational criteria:

- data are captured from day-to-day tasks as performed by real users,
- users are located in normal working environments,
- users self-report their own critical incidents,
- reporting is done within a short time after the problem occurs (i.e., contemporaneous to the usage session),
- no direct interaction is needed between user and evaluator during an evaluation session,
- data capture is cost effective, and

- data are high quality (high value for identifying and fixing usability problems) and relatively easy to translate into usability problem descriptions.

Exploratory Evaluation Study

We performed an exploratory evaluation study [Castillo 1997] of the user-reported critical incident method for remote usability evaluation. We describe the study as "exploratory" because, while we obtained quantitative data and computed simple descriptive statistics such as mean and standard deviation, we did not apply inferential statistical tests for significance formally to prove or refute an experimental hypothesis. Rather, it was a qualitative study to gain insight and understanding about the strengths and weaknesses of the method under practical operating conditions. Thus, all results reported are to be taken as empirically derived "indications" but not statistically supported claims.

Objective of the Study

Our objective was to seek understanding in the context of these primary research questions:

1. Can users report their own critical incidents and, if so, how well can they do it?
2. Can evaluators use remotely reported critical incident data to produce usability problem descriptions and, if so, how well can they do it?
3. What are the variables and values that make the method work best?

Phase 1 of Study: Critical Incident Gathering

Our first research question was: Can users report their own critical incidents and, if so, how well can they do it? We divided this question into the following factors:

- user subject ability to identify and report critical incidents during task performance,
- user subject activity sequencing and timing in reporting critical incidents,
- level of user subject time and effort required to report critical incidents,
- user subject ability to rate severity of critical incidents,
- user subject ability to identify critical incidents at various levels of severity,
- user subject attitudes towards remotely reporting critical incidents, and
- user subject perceptions with respect to interference with user tasks.

PARTICIPANTS

We administered a background questionnaire to non-computer-science majors and selected user subjects based on their having a minimum knowledge of Web browsing and information retrieval. A total of twenty-four students (6 female and 18 male, 22 undergraduate and 2 graduate) participated as volunteer user subjects, from a variety of academic disciplines.

LOCATION OF EQUIPMENT

The best location for users in a study of a remote evaluation method is their own work place. However, the study itself (not the user-reported critical incident method) required a scan converter and videotape deck to make a complete continuous recording of the computer screen during task performance. Since it was not feasible to lend this equipment to each user subject, we provided the next best thing for the user subject: a closed and quiet room isolated from other

people, including us, the experimenters. (For the user-reported critical incident method in the field, digital screen capture software and disk storage will suffice as equipment for any user in any location.)

The experimenter was located in a room adjacent (without two-way glass) to that of the user subjects, who could neither see nor hear the experimenter. An intercom system was installed in both rooms, to be used only as a safety net in case user subjects experienced any hardware or software problems that prevented them from continuing with the tasks. We, as experimenters, did not have any interaction with user subjects during task performance. A scan converter, lapel microphone, and Hi-8 videotape deck recorded video from the computer screen and audio from the user subject.

Experimental Application

The application evaluated by user subjects (as we evaluated the RUEM) in this study is the Internet Movie Database (IMDb) at http://www.imdb.com, which provides free access to extensive movie information. Advertising and sponsorship finance the IMDb site, which contains mechanisms for simple and advanced searching of information about more than 100,000 movies and over 1,500,000 filmography entries. Although no experience was required with the IMDb, three participants had used this application previously.

Critical incident reporting tool

The critical incident reporting tool used in the study was a Web application that allowed user subjects to send structured reports about critical incidents they identified during their experimental session. A "control" window contained the *Report Incident* button and "floated" on the desktop, running independently from the window where the IMDb was displayed. User subjects used the mouse to arrange the IMDb window and the control window so that they could see some of each window on the screen. (In the future the reporting tool could also be implemented by adding a button to the Web browser, saving users some window manipulation.)

Clicking the *Report Incident* button opened a "report" window with questions about the critical incident. This report window was independent from the IMDb window, allowing user subjects to click back and forth between the windows to work on both the task and the critical incident report.

Procedure

We applied minimalist instruction principles [Carroll, 1990] to design critical incident training, a video presentation and a practice session, presented individually to each user subject. To investigate the role of user training for identifying and reporting critical incidents effectively, we randomly assigned user subjects to two separate groups of twelve people in each group. Group 1 watched a training videotape with information about identifying critical incidents, but Group 2 did not, receiving only the training and the practice session.

The practice session, taken by all 24 user subjects, gave hands-on experience in reporting critical incidents using the Web application. The experimenter selected a representative task with the Internet Movie Database, such as finding the biography of actor Denzel Washington, and provided a five-minute overview in how to identify and report critical incidents encountered

while performing this task. Then, in a twenty-minute session, user subjects performed a few more representative tasks and practiced identifying and reporting critical incidents with our reporting tool. The training videotape given to Group 1 before the practice session provided additional examples of critical incident identification in several other applications.

During the study, each user subject performed the same six search tasks using the Internet Movie Database. We created these tasks as representative of what a typical user might do with the movie database (e.g., finding the titles of the four most recent movies directed by Steven Spielberg). User subjects wrote their retrieved responses to these queries on a participant answer sheet, so that correctness of each outcome could be judged unambiguously.

Data Collection

During these tasks, users employed the report window to describe each critical incident that they believed they encountered. The critical incident reporting tool gathered the reports users sent and stored them in the experimenter's computer. Following the evaluation session, each user subject completed a questionnaire about the experience as a remote user.

Data Analysis

For data analysis we reviewed the 24 one-hour videotapes twice, tagging and coding critical incident data involving user subjects. We then identified and counted the critical incidents that user subjects identified, the critical incidents the experimenters identified, and the critical incidents both identified. We also reviewed each critical incident and compared user-assigned severity rankings with the experimenter-assigned rankings. Last, we analyzed all questionnaires to identify user likes and dislikes about the method.

PHASE 2 of Study: Transformation of Critical Incident Data into Usability Problems Descriptions

The second research question of our study was: Can evaluators produce usability problem descriptions from user-reported critical incident data and, if so, how well can they do it? This question was divided into the following areas of investigation:

- ability of evaluator subjects to analyze critical incident data,
- role of textual reports in data analysis,
- role of video in data analysis,
- time and effort required to analyze critical incident data, and
- level of agreement with user subject critical incident severity ratings.

Participants

Four volunteer participants served as evaluator subjects: two graduate students from the Department of Computer Science and two from Industrial and Systems Engineering, all trained in usability methods. Their role was to analyze selected critical incident reports sent by user subjects and convert them into usability problem descriptions.

Equipment and Materials

Since the evaluator subjects did not require any special equipment to analyze critical incident data, they were able to do the analysis at their place of preference such as their home or office. Although these critical incident reports would normally be accessed from the Web, we wanted to be sure that monitor differences and scrolling were not confounding factors, so we presented the

critical incident reports to all evaluator subjects on paper. Two of the evaluator subjects also used a VCR and a video monitor to watch video clips containing visual context for critical incident reports.

Procedure

We reviewed all 74 critical incident reports and randomly selected one good—that is, complete and precise—critical incident report for each of the six tasks for a total of six reports, each report from a different user subject. We next edited the videotape to create a short three-minute video clip for each critical incident, manually determining each clip to be the three-minute interval most useful in identifying the usability problem associated with the critical incident, to simulate the best video clip that could be captured by the critical incident reporting system. The overall result was to produce contextualized critical incident packages for evaluator subjects to review in this phase: one set of six critical incident reports and six related videotape clips.

Data Collection

The two report-only evaluator subjects, each working independently, analyzed the six critical incident reports to create a list of usability problem descriptions. The two clip-and-report evaluator-subjects, again working independently, analyzed the same six critical incident reports plus six corresponding video clips to create a list of usability problem descriptions. In addition, all four evaluator subjects completed a satisfaction questionnaire about their experience as evaluators in a remote usability situation. For data collection, we gathered all the usability problem descriptions and answers to the satisfaction questionnaire.

Data Analysis

We examined the four lists of usability problem descriptions created by the evaluator subjects and analyzed their questionnaires with respect to the following:

- the feasibility (i.e., time and effort) of transforming remotely-reported critical incident data into usability problem descriptions;
- the level of agreement about the severity ratings of critical incidents between user subjects and evaluator subjects;
- the quality of content in critical incident reports; and
- the role of video, text, and audio during analysis of critical incident data.

As a general matter, the selection of a small number of critical incidents involved and the small number of evaluator subjects and individual differences among them strongly colored the indications gleaned from this phase of the study. Therefore, the “results” obtained here are to be considered only points of interest and hypotheses for further studies using larger numbers of evaluator subjects.

The issues indicated by both phases of our study are reported here as a series of expectations and outcomes, but not results with statistical significance.

Evaluation Study Outcomes

This section describes the outcomes of objective issues (what we observed), subjective issues (what users thought), and what we concluded from our exploratory study. In each category, we looked at user-related

issues, evaluator-related issues, and method- and study-related issues.

User-Related Issues: Can Users Identify and Report Critical Incidents?

ABILITY TO REPORT CRITICAL INCIDENTS

We, the experimenters, found 97 critical incidents across all user subjects and all tasks in our review of the tapes (Figure 1).

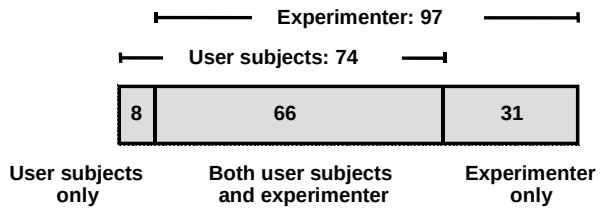

Figure 1. Number of critical incidents identified by user subjects and by the experimenters

Of the critical incidents identified by us, user subjects reported a total 66 critical incidents and missed 31 incidents (mostly of low severity). Interestingly, user subjects reported a total of 74 critical reports (a mean of 3.1 reports per user subject, standard deviation of 1.7). This means that user subjects reported eight critical incidents (all of low severity) that we did not recognize from reviews of the tapes. We did not, however, consider these reports as gratuitous, sent to please us, and concluded that these critical incidents were known in the minds of the user subjects but were not evident in the videotapes. Because we could not confirm them, however, these 8 reports were not considered during data analysis.

This breakdown of critical incident reports, here and in subsequent sections, is conservative because it counts every critical incident experienced by every user, including cases where more than one user encountered the same critical incident. For real data gathering, if two or more users encounter the same critical incident, their reports would be combined by evaluators into one. This latter approach of merging reports over users is a more fair comparison to lab-based usability testing, for example, where the usability problems from all user subjects are combined in the final report of performance.

TYPE OF INCIDENTS REPORTED

We expected user subjects to identify the majority of critical incidents occurring at each severity level. User subjects mostly met expectations by reporting 21 out of 28 (75%) of the critical incidents that we identified as high severity (Figure 2), 19 out of 24 (79%) medium severity critical incidents, and 15 out of 45 (33%) low severity ones, as ranked by us. Thus, user subjects identified 40 out of 52 (77%) of the important (medium and high severity) critical incidents. We found that 26 of the 31 critical incidents not reported by user subjects were of low severity.

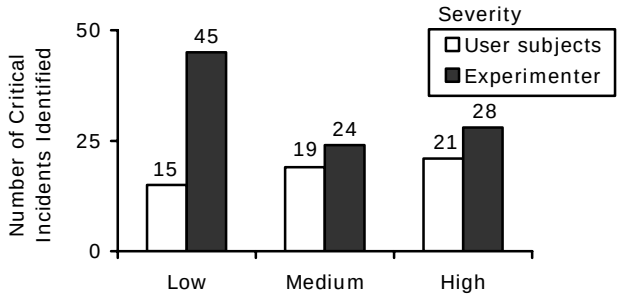

Figure 2. Number of reported critical incidents by severity ranking

We observed that user subjects were generally aware of errors and problems they encountered that had a negative effect on their task performance (which were rated as high- and medium- severity critical incidents) and that most of the critical incidents missed by users, but which might be identified by an expert, were low severity.

Agreement with Severity Rankings

For most cases, we expected user subjects to agree with our severity ratings. Users made severity ratings on a scale of one through five, with one being the lowest severity and five the highest. As an abstraction, we converted the ratings to severity rankings, where the low severity rank corresponds to ratings one and two, medium severity rank corresponds to rating three, and high severity rank corresponds to ratings four and five.

Across all 24 subjects, user subject rankings agreed with ours for 55 out of 66 (83%) of the critical incidents reported by both the user and the evaluator subjects. The others were balanced, with six reports being lower severity than ours, and five higher (Figure 3).

Figure 3. User subject severity ratings compared to our ratings

Time When Users Reported Incidents

We expected the flow of the activities during task performance and reporting to be somewhat structured; in particular, we expected most users subjects would report critical incidents immediately after they occur. The expectation for a structured flow of task and reporting activities—task performance, critical incident identification, followed by reporting—was not met in most cases. Not surprisingly, high severity critical incidents had the most disruptive impact on task performance and flow of activities. Eleven user subjects sent high severity critical incident reports for tasks they never completed. Sometimes when encountering a critical incident, user subjects gave up and continued with the next task without any effort to complete the current task. Sometimes they jumped to the next task but later came back to work on the troublesome task (with or without success).

We observed considerable further variation in user subject behavior with regard to the timing of critical

incident reports. User subjects reported critical incidents during task performance, immediately after the task ended, at a later time, or sometimes did not report the critical incident at all. Although we directed user subjects to send a report immediately upon encountering a critical incident, they sent 52 (70%) of all 74 critical incidents reports (Figure 4) after the task ended.

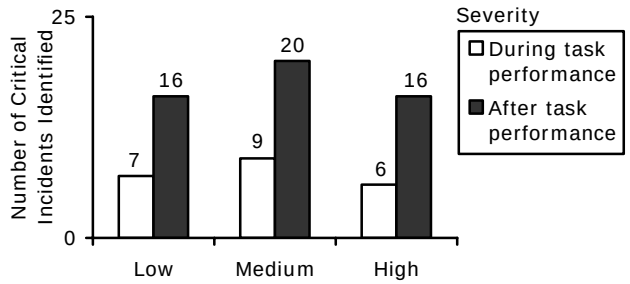

Figure 4. Most critical incident reports were sent after task completion

User subjects may have wanted to delay reporting until they were done with the task to avoid added complexity of concurrent activity and to gain understanding of the problem. In an attempt to learn more about the nature of the timing issues, we watched all 24 videotapes yet again. For each critical incident report, we determined the point in time when it was first evident that a critical incident had occurred and compared that with the time the user reported it.

We also measured the time it took to produce each report. As a rule of thumb, delays in reporting, as well as the time required to produce a report, corresponded roughly with the severity of the critical incident. It seems reasonable that a more severe critical incident requires more information to report and, therefore, results in both a larger delay before reporting and a longer time to make the report. Figure 5 illustrates the mean time interval for all critical incidents between the onset of the critical incident and the point of reporting.

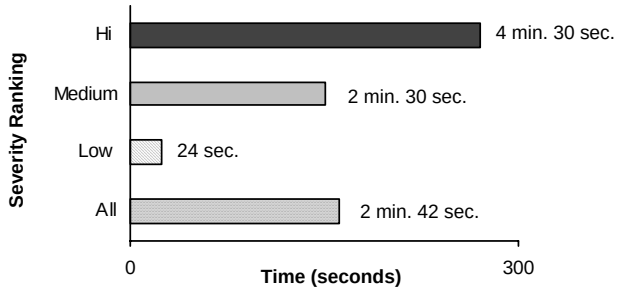

Figure 5. Average delay in reporting after clear onset of critical incidents

Time Needed to Enter Reports

Based on our feasibility case study, we expected user subjects to spend an average of about 3 minutes entering critical incident reports. As seen in Figure 6, user subjects took significantly longer to make critical incident reports, spending an average of 5.4 minutes (standard deviation of 2.3) per report, in a range from 2.08 minutes to 12.25 minutes.

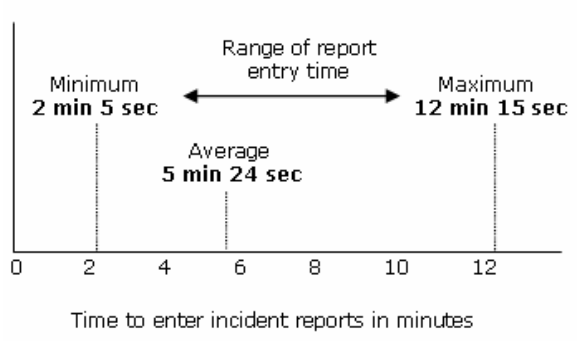

Figure 6. Average time to enter critical incident reports

User subjects who reported critical incidents during task performance spent more time (mean of 6.7 minutes, standard deviation of 2.7) producing critical incident reports than those user subjects who waited until the task ended (mean of 4.8 minutes, standard deviation of 1.6). The higher average time for reports during task performance is possibly due to the added overhead of interleaving two tasks, each with a significant cognitive load.

Time to Report High-Severity vs. Low-Severity Incidents

We expected user subjects to spend more time entering reports for high-severity critical incidents than for low- or medium-severity critical incidents. Contrary to expectations, user subjects spent more time (mean of 6.1 minutes, standard deviation of 2.8) reporting low-severity critical incidents than medium-severity incidents (mean of 4.9 minutes, standard deviation of 1.8) or high-severity critical incidents (mean of 5.3 minutes, standard deviation of 1.9). We were unable to explain this effect. We did notice that many user subjects appeared to take less time to report critical incidents as they gained some experience with the reporting process. During pre-testing, two pilot subjects trained in human-computer interaction and usability methods spent more time in reporting their first critical incident than in reporting subsequent critical incidents. This also happened to 13 of the 24 (or 54%) user subjects who reported more than one critical incident.

Desire to Report Problems Remotely

In a satisfaction questionnaire all 24 user subjects agreed (14 user subject strongly agreed, 8 moderately agreed, and 2 agreed) that, as users, they want to be able to report critical incident information remotely to evaluators.

The following statements made by user subjects tend to confirm this preference anecdotally:

- "I believe that the idea presented in this study is long overdue! Many times you have problems and resort to searching manuals and email addresses to find information on how to fix them. It would be wonderful to let developers know what you have problems with… The real world is full of different people and problems that most certainly are not found in a controlled environment like a lab."
- "It also allows me to feel better knowing that I told someone about the problem encountered, and I don't get as frustrated".

Level of Interference with Task Performance

Usage problem reporting by remote users working on real tasks for real work has the potential to interfere with task performance. However, contrary to our expectations, 19 (or 79%) of the 24 user subjects moderately or strongly agreed that identifying and

reporting critical incidents was not intrusive and did not interfere with their tasks. This counter-intuitive indication was reinforced by the fact that some of the best and most complete critical incident reports (associated with long reporting times) came from user subjects who said that they felt that reporting did not interfere much. Perhaps user subjects felt that their work flow was already interrupted by the critical incident and they might as well report it as long as the task was off track. Perhaps also feelings of interference were offset by the satisfaction of increased understanding of the problem and/or relief of frustration due to being able to report problems. These points need further exploration in future studies.

Ease of Rating Critical Incidents

We expected user subjects to find it easy to rate the severity of critical incidents. Twenty two (or 92%) of the 24 user subjects agreed that it was somewhat easy to determine the severity of critical incidents encountered during the evaluation session. However, we observed that some user subjects still had difficulties in determining critical incident severity, including:

- uncertainty about rating low severity critical incidents properly;
- lenient attitude toward problems ("I'm used to trying things four or five times in different ways to get something done, and if I make it work after a couple of tries, I might forget the details of the initial difficulties");
- unwillingness to read long descriptions for each severity rating option; and
- users' inclination to select the middle point of the scale when uncertain about which option to choose (about 35% of critical incidents were reported by user subjects as medium severity).

Evaluator-Related Issues: Can Evaluators Use Critical Incident Data to Produce Usability Problem Descriptions?

Two report-only evaluator subjects worked independently of each other, each analyzing the same six selected critical incident reports to create a list of usability problem descriptions. The two clip-and-report evaluator-subjects also worked independently, each analyzing the same six critical incident reports plus six corresponding video clips to create a list of usability problem descriptions. All four evaluator subjects completed a satisfaction questionnaire about their experience as evaluators in a remote usability situation. The data collected in this phase are too sparse to allow conclusions stronger than conjecture.

Ability to Analyze Critical Incident Data

Generally, all evaluator subjects were capable of analyzing critical incident data to produce usability problem descriptions. Report-only evaluator subjects reported similar or related usability problem descriptions for five out of the six critical incident reports. Clip-and-report evaluator subjects were similar in their ability to produce usability problem descriptions. However, differences among the styles of problem description across the four evaluator subjects made it difficult to compare them. The evaluator subjects each described the same usability problem in different ways and organized their lists in different ways. For example, one evaluator subject's list described usability problems found for each task while the other list described usability problems found in the user interface as a whole. In a future such study, a

more structured and standardized reporting format probably would help reduce these differences.

USEFULNESS OF VIDEO TO ANALYZE DATA

We expected that clip-and-report evaluator subjects would find the video clips (in addition to the textual reports) to be helpful in creating usability problems descriptions. They did not meet our expectations, somewhat or strongly disagreeing that videotape clips added value to the critical incident reports for creating usability problem descriptions. One evaluator subject commented "it was somewhat difficult to match the two together—that is, the critical incident report and the video clip.

In an apparent contradiction, however, clip-and-report evaluator subjects disagreed with the idea of not using video clips to supplement the critical incident reports for determining usability problem descriptions. One clip-and-report evaluator subject indicated in the questionnaire, "I'm not sure how I would have liked just reading the critical incident reports."

Further, these evaluator subjects mentioned that:

- " The video clips helped me clarify the order of events…"
- "…much of user strategy (e.g., searching menus) would have been lost without the clips."

As a matter of conjecture, report-only evaluator subjects moderately or strongly agreed that they believed video clips (with audio) of screen action, in addition to the critical incident reports, would have helped them create usability problem descriptions. Many of the evaluator subjects' difficulties with the video clips can be attributed to the low resolution of the scan-converted analog video, which made it difficult to discern detail (such as text) on the user's screen. We still believe that a video account of the user's screen action, especially if narrated with audio, would be helpful for understanding critical incidents and usability problems, since many usability problems are about the specifics of user actions on user interface objects and the cognitive activity behind such actions. Using a full resolution digital screen capture program, such as TechSmith® Camtasia Studio® [TechSmith®, 2005a] or Morae® [TechSmith®, 2005b] will solve this perceptual problem in the future. Obviously, this aspect of the problem needs more study.

Lessons Learned

Need to De-Couple Critical Incident Identification From Reporting and Video Capture

One of the most significant observations in the study involved the delay in reporting critical incidents and its effect on the relevance of our intended video clips. In any design, automatic screen sequence capture requires some trigger mechanism to initiate the process to capture video clips to accompany critical incident reports as visual context. We originally intended to capture (via a continuous recording loop) the video clip during the two-minutes leading up to the critical incident, which we hope would show the events leading to the critical incident onset. However, our study shows a highly variable, and often large, delay between the clear onset of a critical incident and the time when user subjects. Tape clips of the two-minutes leading up to the clicking of the *Report Incident* button most often would have missed user actions relevant to the critical incident. Thus, this study reveals the need for a different trigger mechanism for recording critical incident actions, separate from the *Report Incident* button. In our new design that trigger mechanism is

the user-controlled *Record Video Clip* button, as shown at the right area of the window in Figure 7. The label of the button changes to *Stop Video Clip* after users click the *Record Video Clip* button.

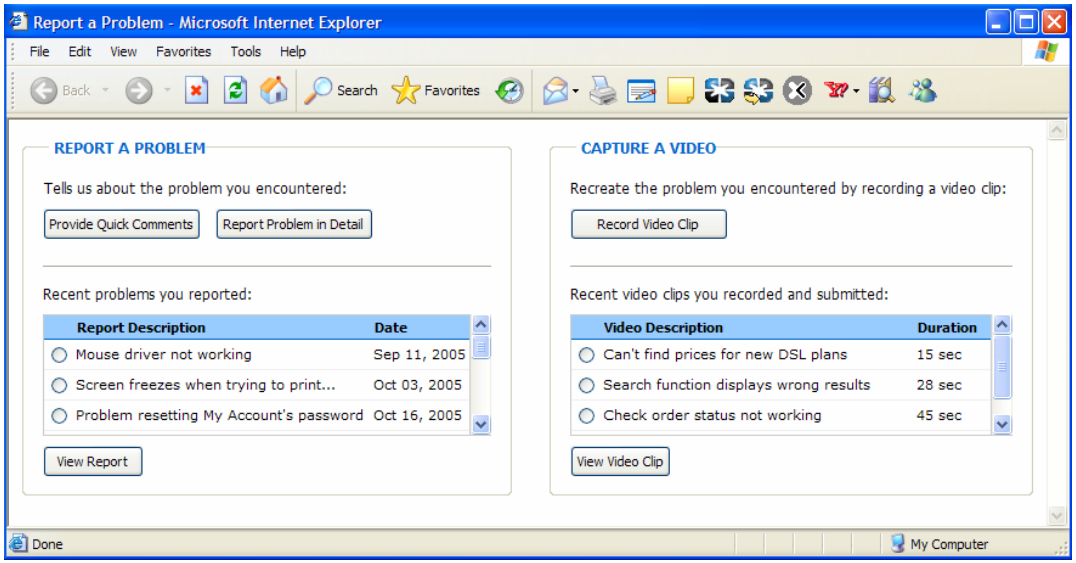

Figure 7. Redesigned critical incident reporting tool

Giving the users control of when to start and stop the recording led to the realization that users could make explicit reenactments by clicking on the Record Video Clip button and performing a narrated demonstration of the critical incident. With users in control, video capture can now be simpler and more effective and the video clips can be focused clearly on the critical incidents.

Need for Short, Quick Reports

Users expressed a need for more than one kind of problem report. For situations where users want to send a complete critical incident report describing the issue in detail, they can click on a Report Problem in Detail button. For other situations were the critical incident is not seen as important enough, users asked for a "quick report" capability, which is accommodated in the new design via the Provide Quick Comments button. Quick reports feature free-form textual

comments rather than use of the structured questionnaire.

Need to Browse and Review Previous Reports
Users indicated a need for support in situations where they have identified a critical incident but are not sure whether they have reported that particular incident or a similar one earlier. This problem is solved in the new design by showing a list of reports previously sent by a user who can later select critical incidents by descriptive names for browsing and editing. Similarly, a list can be provided for users to review previous video clips that they have recorded and submitted.

Future Work

We recognize the following research activities as possibilities for furthering this work by refining the user-reported critical incident method for remote usability evaluation:

- Study to determine the necessity for, and effectiveness of, various kinds of user training to recognize and report critical incidents.
- Study to determine further the importance of the role of video as contextual support for critical incident reports.
- Study to investigate user preferences and effectiveness in verbal (via audio capture) versus textual critical incident reporting.
- Study in a real remote usability evaluation setting to confirm the present indications using separate expert subjects (different than the experimenters) to compare the usability problem lists created by evaluator subjects.
- Study to determine the value of using narrated reenactments of critical incidents by the user as the video clip. This approach is definitely deserving of more consideration. We believe users can easily be trained explicitly to show critical incidents to evaluators via video demonstrations coupled with audio explanations.

Acknowledgements

This paper stems from a Master's thesis [Castillo, 1997] and is an adaptation and extension of a conference paper [Hartson & Castillo, 1998] presented at Advanced Visual Interfaces (AVI '98) in L'Aquila, Italy. We gratefully acknowledge helpful comments by Robert C. Williges and Pawan R. Vora.

References

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

[2] Castillo, J. C. (1997). The User-Reported Critical Incident Method for Remote Usability Evaluation. Unpublished Master's Thesis, Virginia Tech, Blacksburg, VA 24061 U.S.A.

[3] Castillo, J.C. & Hartson, H.R. (2006). Remote Usability Testing Methods a la Carte. Submitted to the

[4] del Galdo, E. M., Williges, R. C., Williges, B. H., & Wixon, D. R. (1986). An Evaluation of Critical Incidents for Software Documentation Design. In Proceedings of Thirtieth Annual Human Factors Society Conference Human Factors Society, Anaheim, CA, 19-23.

[5] Dzida, W., Wiethoff, M., & Arnold, A. G. (1993). ERGOGuide: The Quality Assurance Guide to Ergonomic Software: Joint internal technical report of GMD (Germany) and Delft University of Technology (The Netherlands).

[6] Fitts, P. M., & Jones, R. E. (1947) Psychological Aspects of Instrument Display: Analysis of Factors Contributing to 460 "Pilot Error" Experiences in Operating Aircraft Controls. Reprinted in Selected Papers on Human Factors in the Design and Use of Control Systems (1961). Sinaiko ed. Dover Publications, Inc., New York, 1947, 332-358.

[7] Flanagan, J. C. (1954). The Critical Incident Technique. Psychological Bulletin. 51, 4, 327-358.

[8] Hartson, H. R. & Castillo, J. C. (1998). Remote Evaluation for Post-Deployment Usability Improvement. Proceedings of the Working Conference on Advanced Visual Interfaces (AVI'98), L'Aquila, Italy, 22-29.

[9] Hix, D., & Hartson, H. R. (1993). Developing User Interfaces: Ensuring Usability Through Product & Process. New York: John Wiley & Sons, Inc.

[10] Koenemann-Belliveau, J., Carroll, J.M., Rosson, M.B., & Singley, M.K. (1994). Comparative Usability Evaluation: Critical Incidents and Critical Threads. Proceedings of the CHI Conference on Human Factors in Computing Systems, 245-251.

[11] Shattuck, L. W. & Woods, D. D. (1994). The Critical Incident Technique: 40 Years Later. Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting, 1080-1084. Santa Monica, CA: HFES

[12] TechSmith® Camtasia Studio® (2006a). http://www.techsmith.com/camtasia.asp [Visited: 08/01/2006]

[13] TechSmith® Morae® (2006b). http://www.techsmith.com/morae.asp [Visited: 08/01/2006]