CONTEMPORANEOUS VERSUS RETROSPECTIVE USER-REPORTED CRITICAL INCIDENTS IN USABILITY EVALUATION

Miranda G. Capra
Grado Department of Industrial and Systems Engineering
Virginia Polytechnic Institute and State University
Blacksburg, VA

The *user-reported critical incident technique* involves end-users directly in qualitative data collection during formative usability evaluations. An *augmented retrospective* variation was developed where participants reported incidents while watching a recording of their usability session, rather than reporting incidents *contemporaneous* to their occurrence during task execution. Retrospective reporting enables controlled comparisons of user-reported and expert-reported methods, since session recordings can be shown to multiple reviewers. It also allows for the collection of incidents without disrupting traditional usability measures, such as time to complete task. A within-subject study with 24 participants found retrospective reporting to be similarly effective to contemporaneous reporting. The study is described and guidelines are provided for the use of both the contemporaneous and augmented retrospective techniques.

INTRODUCTION

Critical incident reporting is an effective way of collecting qualitative information during formative usability testing. Having the end-user, rather than a trained observer, identify and describe the critical incidents increases designers' understanding of user perceptions of the interface being evaluated by increasing communication between users and designers (del Galdo, Williges, Williges, & Wixon, 1986). In the user-reported critical incident technique, users take time out from using an interface to describe interactions that greatly increase or impair their performance (critical incidents). Hartson and Castillo (1998) found the technique to be well suited to remote usability studies where users explore an interface in their own environments with no experimenter involvement, particularly when the incident reports were packaged with a brief screen-capture movie file illustrating the incident. The technique is valuable for participants using deployed software for daily tasks and submitting feedback that developers incorporate into future product revisions. Describing critical incidents can help focus users' feedback, and incident reports can be easily collected over the internet through structured web forms.

The benefit of collecting user-reported critical incidents in laboratory studies is less clear. The technique, as used by Hartson and Castillo (1998), has users reporting incidents when they happen or soon after, which they call reporting *contemporaneously* to task performance. Users frequently re-work tasks and explore the interface to help them complete their reports, which intereferes collection of traditional objective usability measures, such as time to complete task. They also tend to lead to additional feature discovery as users reexamine the interface to understand the incident for reporting, and this additional experience with the interface can affect execution of subsequent tasks.

Furthermore, it is difficult to quantify the benefit of collecting user descriptions in addition to expert observations of participant behavior, since both types of information are qualitative. Two previous studies have compared user-reported and expert-reported critical incidents. Hartson and Castillo (1998) found that users report most of the high-severity incidents identified by expert reviewers, although fewer of the lower-severity incidents. However the expert reviewers were aware of all incidents reported by the user participants, which may have biased their reviews. Thompson (1999) avoided this biasing by using different, non-reporting participants to create the tapes for expert review, but found that individual differences in expert review were too high to draw formal conclusions from the results.

To address these issues with formal comparisons we developed a variation of the user-reported critical incident technique where users complete a traditional usability session, uninterrupted, and then report incidents while watching a tape of their usability session. We call this technique *augmented retrospective* reporting, since users' retrospective recall is aided by watching and listening to the session tape. In previous studies session recordings have been used successfully to cue participant think-aloud comments for retrospective verbal protocol studies (Bowers & Snyder, 1990; Page & Rahimi, 1995). The session tapes created for augmented retrospective reporting are "clean", meaning that they do not show the users identifying and reporting incidents. The tapes can then be shown to expert observers for controlled comparison studies in which the same participants are used for both user-reported critical incidents and expert review without biasing the expert reviewers. The augmented retrospective technique has the added benefit of not interfering with the usability session, eliminating the issues of interruption and exploration. Thus the same participants can be used for both the collection of user-reported critical incidents and traditional task-performance measures, balancing the subjective user reports with objective laboratory measures.

The study described in this paper is a preliminary study to determine whether the augmented technique is feasible, whether it generates similar information to the contemporaneous technique, and to refine the technique itself. The comparison of contemporaneous and augmented retrospective reporting is just one factor in a larger study of user-reported critical incidents. Details of the study and results are presented in Capra (2001).

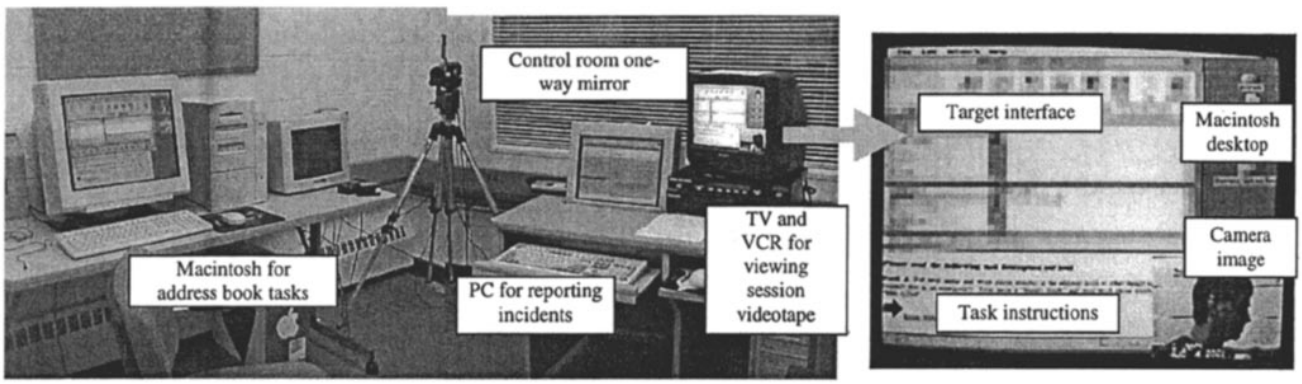

Figure 1 Participant room (composite photo) and session recording (annotated, target interface blurred)

METHOD

Design

A within-subject experimental design was used to compare two reporting techniques: contemporaneous and augmented retrospective.

Participants

Twenty-four individuals with no usability experience participated (12 male, 12 female), ages 19 to 35 (M = 22.4, SD = 3.5). One participant had 1-3 years of computer experience, and the rest had four or more. Two were taking an undergraduate introductory course in human factors.

Materials

Tasks involved both simple address book functions, such as adding addresses and appointments, and complex functions, such as grouping, import and export. The address book software had known interface problems, including non-standard menu names and groupings, inconsistent terminology, unclear icons and unlabeled interface objects. The software was chosen because it had been used in previous studies (Andre, 2000; McCreary, 2001) and because it had known usability problems.

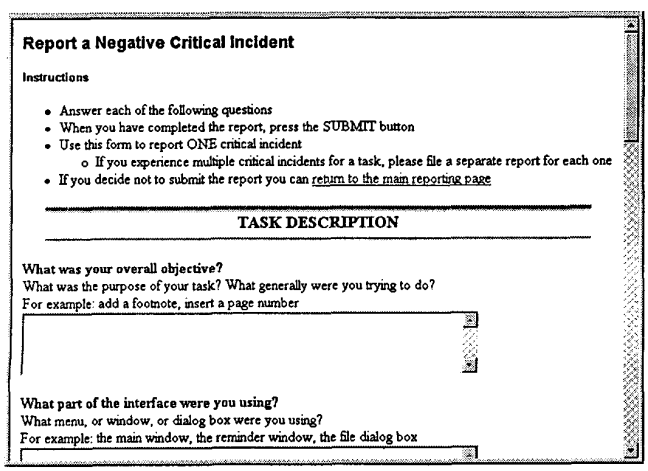

Figure 2 Critical incident report form

The sessions were recorded to S-VHS videotape using a scan converter, and a camera image showing the participant's face was overlaid on the corner of the video (Figure 1). Tasks were performed on a Macintosh computer with the screen set to 640x480 pixels to maximize text readability on the session recordings. All room noises were recorded to the tape's audio track, including clicks, typing, and think-aloud comments.

Participants reported critical incidents on separate computer running Windows NT and the Microsoft Internet Explorer web browser. The incident reporting forms (Figure 2) are web pages based on Thompson and Williges (2000). The forms ask participants to indicate whether the incident is positive or negative, and then to describe their overall objective, part of the interface being used, how the task was being carried out, what happened, how performance was affected, how this made them feel, how they recovered from the incident, and the severity of the incident.

A post-session questionnaire was used at the end of each usability session to assess participants' confidence in their ability to report incidents, and engagement in positive and negative incident reporting (6 questions each, 18 total). For example, one of the statements to assess confidence was "I understand how to report critical incidents," and one to assess engagement in positive incident reporting was "my positive incident reports were thorough and complete." All questions were developed by the experimenter and ratings used a 6-point Likert-type (Likert, 1932) scale.

Procedure

All participants attended three sessions: an introductory session where they learned about critical incidents (positive and negative) and practiced reporting, and two usability sessions: one using contemporaneous reporting and one using augmented retrospective. Half the participants used retrospective reporting first and half second. During each usability session, participants began with a brief review of critical incidents and practiced reporting a negative incident, and then performed six tasks. At the end of both usability sessions participants completed the post-session questionnaire. At the end of the experiment participants filled out a post-test questionnaire asking which technique they preferred and why, and how strong was their preference (very strong, strong, moderate or mild).

During the retrospective sessions, participants first performed all six tasks for the session uninterrupted. Then, after completing the last task, they watched their session recording and reported any critical incidents observed. During the contemporaneous sessions, participants reported incidents either as soon as they happened or when they reached a convenient stopping point. For both sessions, participants were asked to record all incidents associated with one task before moving on to the next. A total of twelve tasks were used for the study, and tasks were always performed in the same order.

Dependent Measures

The dependent measures recorded were: confidence and engagement in positive/negative incident reporting, computed by averaging the six post-session questions for each attribute; overall number of incidents reported per session; number of positive vs. negative incidents, as indicated by the participant; number of high-severity vs. low-severity incidents, as computed by collapsing the four-point scale used by the participants into two groups; and average time to report an incident, per session.

RESULTS

Data Analysis

Participants reported a total of 360 critical incidents (Table 1). Most of the sessions took between one and three hours: 10-20 minutes of review and practice, 15-45 minutes working through tasks, the same amount of time watching the session videotape (retrospective participants only), and 30-60 minutes reporting incidents.

Table 1: Number of Critical Incidents Reported Per User

Type of Incident	Contemporaneous			Retrospective		
	Mean	SD	Min.- Max	Mean	SD	Range
Positive	3.4	1.8	1 – 10	3.5	1.9	1 – 8
Negative	3.8	1.5	1 – 7	4.3	1.7	1 – 9
High-severity	3.5	2.1	0 – 9	3.2	1.7	1 – 7
Low-severity	3.6	2.0	0 – 7	4.7	3.1	0 – 11
Overall	7.1	2.2	3 – 13	7.9	2.6	4 – 14

Table 2: Summary of ANOVA Results

Dependent Measure	Reporting Technique	Order Effects	Session Effects
Confidence	$p = 0.94$	$p = 0.54\approx$	$p = 0.13$
Engagement – Positive	0.56	$0.58\approx$	$0.87\approx$
Engagement – Negative	0.17	$0.57\approx$	0.10
Frequency – Overall	0.27	$0.47\approx$	$0.23\approx$
Frequency – Positive	0.70	$0.87\approx$	$0.55\approx$
Frequency – Negative	0.29	$0.24\approx$	$0.29\approx$
Freq. – High-Severity	0.55	$0.58\approx$	0.18
Freq. – Low-Severity	0.14	$0.33\approx$	0.04*
Avg. Time to Report	0.19	$0.23\approx$	$0.60\approx$

* $p < 0.05$ $\approx p > 0.2$ (equivalent treatment conditions)

Three analyses of variance were performed to test for significant effects due to reporting technique, order effects due to which reporting technique was used first, and whether or not the two usability sessions can be considered equivalent treatments (Table 2). There were no significant effects due to either reporting technique usage or order of exposure, and most dependent measures were equivalent across both usability sessions. There was a significant effect on the number of low-severity incidents reported per session due to day of the usability session; participants reported more low-severity incidents during the first usability session ($M = 3.6$, $SD = 2.2$) than the second ($M = 3.3$, $SD = 1.4$).

Technique Preferences

When asked in the post-test questionnaire which technique they preferred, 15 of the 24 participants chose contemporaneous. A chi-square test indicated that this was not a statistically significant difference ($p=0.22$), but this test did not take into account the strength of these preferences (Table 3). When asked to explain their preference for contemporaneous reporting, 12 participants said it was easier to remember when you reported soon after the incident, three said incidents seemed less important when watching the recording, two commented that some of the text on the recording was blurry, and one commented about the time spent watching the recording. Three participants said that they preferred using a recording because stopping to report was disruptive to task completion, two said that watching the recording multiple times was easier than re-working a task on the computer, and one said incidents sometimes got ignored when focusing on task completion.

Table 3: User Technique Preferences

Strength of Preference	Preference	
	Contemporaneous	Retrospective
Very Strong	4	1
Strong	5	1
Moderate	6	5
Mild	0	2

Observations of Participant Behavior

The following observations were made of participants during the study and from reading their critical incident reports. These observations may help experimenters decide whether the technique is appropriate for a particular design project and understand how to interpret user-supplied critical incident reports. These observations are not specific to either technique, and the incident reports from which they are drawn are included in Capra (2001).

Participants reported information that might have been unavailable to an observer. This includes information such as motivations, feelings, and decisions. Several participants also described how previous software experience affected their understanding of the target interface.

Participants had incomplete knowledge about the target interface. Participants that failed to locate a feature sometimes reported that it did not exist. Participants that used a non-

optimal task completion strategy tended to report that the task was complicated. Several participants commented that they wished they had been able to go back and change some of their incident reports at the end of the study, when they better understood the target interface.

Incident reports were unfocused and imprecise. Participants' incident reports frequently mentioned several critical incidents and multiple usability issues. Participants sometimes used imprecise terminology to describe interface elements, such as calling a scrollable selection box a "pull-down menu".

Incident reports were positively biased. Participants frequently submitted a positive report when they figured out a difficult task without submitting a negative report describing what made the task difficult. Several participants reported positive incidents because a confirmation dialog kept them from making a mistake, and yet did not report negative incidents about the interface design that caused them to take the action being confirmed.

DISCUSSION

Interpretation of Results

The study measures indicate that users are indeed able to report critical incidents while watching session videotapes. Although participants' preferences for contemporaneous reporting were stronger than those for augmented retrospective, this was not reflected in their confidence and engagement levels, and participants reported as many incidents during the augmented retrospective sessions as they did during the contemporaneous sessions. The author's judgment, based on observations during the study and reading participants' incident reports, is that participants did as good a job describing incidents during the retrospective sessions as they did during the contemporaneous.

The decrease in low-severity incident reporting during the second usability session may be due to participant fatigue or to the nature of the tasks used for the second usability session, which were judged by the experimenter to be more difficult. For example, several participants during the second usability session were unable to import addresses from a text file, and only a few participants figured out how to export a specific subset of addresses to an external file.

An important difference between the two techniques is that the retrospective technique takes longer. Since participants have to watch their entire session tape, retrospective sessions tend to be longer by the length of the tape. Also, the retrospective sessions do not interfere with task performance thus allowing the measurement of objective usability measures, such as time to complete task. On the other hand, the contemporaneous technique does not require video capturing software, although Hartson and Castillo (1998) recommend it. While participants in this study were not allowed to use the target interface while reviewing the session videotape, such usage could be allowed in other studies.

This was a preliminary study. Further studies are needed to evaluate the contents of the incident reports for differences between contemporaneous and augmented retrospective reporting. For example, Bowers and Snyder (1990) found in a study of verbal protocol that retrospective statements made while watching a session recording were fewer than those made concurrent to task execution, but were more useful to designers because they were more explanatory and less descriptive. Similarly, there may be differences between the reports collected using the two critical incident techniques. For example, augmented retrospective reports are written at the end of the session when participants have more experience with the target interface and so may have different perspectives on incidents. Participants are also less focused on task completion and may be more analytical of their own actions. The "clean" session tapes created during the augmented retrospective sessions could also be used in comparison studies. For example, the same tapes could be used for both user reporting of critical incidents and more traditional expert observation, and the user-supplied and expert-supplied information could be compared for differences in coverage of usability problems, quality of problem descriptions, etc. Such information would help practitioners decide when to use the techniques.

Suggested Changes to Both User-Reported Techniques

Based on experiences running this study, several changes are recommended to both the contemporaneous and augmented retrospective techniques used in this study.

Use digital screen capture for session recordings. Even at 640x480 pixels participants complained about being unable to read some of the screen text on the session tape, and most modern software relies on much higher screen resolutions. Software to capture screen activity to a digital movie file, such as AVI or QuickTime, would record much more detail, and would not require use of a scan converter. While room noise should be easy to capture in this movie file, it might be difficult to include a video of the participant's face. However, this may be acceptable if think-aloud comments are recorded.

Encourage think-aloud comments. Knowing what the participant is thinking can be critical to retrospective recall by the participant and understanding by an expert observer, especially if the participant's face is not included on the session recording. A promising strategy is that used by Ebling and John (2000), which relied on having participants practice think-aloud while playing solitaire on the computer and then prompting them during the study to keep talking.

Ask for negative aspects of every incident. It is well known that participants in laboratory studies tend to be positively biased. Thompson and Williges (2000) found that laboratory-based users reported more positive critical incidents than remote users. Participants in the current study seemed reluctant to submit both negative and positive reports about a single incident and so frequently submitted only a positive report. Several participants commented that it was sometimes difficult to tell whether an incident was positive or negative. Participants could be encouraged to report negative information by having a single reporting form that asks for both positive and negative information. Hartson and Castillo (1998) collected only negative incidents, since these are of the most interest for revising post-deployment interfaces.

Guidelines for Usage of Both Techniques

The user-reported critical incident technique is effective for gathering qualitative information through usability studies during formative evaluation. The strength of the contemporaneous technique lies in its suitability for remote studies, where participants use an interface on their own time for daily tasks and submit incident reports across a network. The augmented retrospective technique is best suited to a usability laboratory setting, considering the equipment and lengthy sessions involved. Retrospective reporting allows collection of both task performance measures and critical incident descriptions from the same participants, and creates "clean" session tapes that can be used to compare evaluation techniques.

The following tips should be helpful for the practitioner wishing to use either the contemporaneous or augmented retrospective user-reported critical incident technique.

- Expect retrospective sessions to be longer by the same amount of time the participant spends using the target interface.
- Make sure the recording medium has sufficient detail to discern all interface objects and read all text.
- If structured tasks are used, make sure that:
  - Participants have the task text while reviewing the session recording, either on the recording itself or on a separate piece of paper.
  - The recording clearly shows when the participant moves on to a new task.
- Expect to review and interpret user-supplied information in order to:
  - Identify misconceptions about the interface.
  - Infer design problems from incomplete descriptions.
  - Infer negative interactions from positively-phrased descriptions.
  - Extract individual usability problems from lengthy incident descriptions.
- Decide whether or not to have participants review and add comments to all submitted incidents at the end of the session or study; participants may have additional perspective based on later experiences with the target interface.
- Decide whether or not to allow participants to review all previously submitted incidents at any time during the study; participants request this feature, but reviewing previous incidents allows cut-and-paste of descriptions and encourages participants to ignore incidents that have already been already reported.
- Decide whether or not to allow participants to explore the target interface while filing incident reports; it may help participants remember what happened, but can also lead to feature discovery.

ACKNOWLEDGEMENTS

The author is grateful for the advice of Robert C. Williges, H. Rex Hartson and Tonya L. Smith-Jackson. This study builds on research by the Usability Methods Research Laboratory (http://research.cs.vt.edu/usability/) and the Human-Computer Interaction Laboratory (http://hci.ise.vt.edu/) at Virginia Tech.

REFERENCES

Andre, T. S. (2000). *Determining the effectiveness of the usability problem inspector: a theory-based model and tool for finding usability problems.* Unpublished dissertation, Virginia Tech, Blacksburg, VA (available via http://scholar.lib.vt.edu/theses/available/etd-04122000-09440030/).

Bowers, V. A., & Snyder, H. L. (1990). Concurrent versus retrospective verbal protocol for comparing window usability. In *Proceedings of the Human Factors and Ergonomics Society 34th Annual Meeting* (pp. 1270-1274). Santa Monica, CA: Human Factors Society.

Capra, M. G. (2001). *An Exploration of End-User Critical Incident Classification.* Unpublished thesis, Virginia Tech, Blacksburg, VA (available via http://scholar.lib.vt.edu/theses/available/etd-11122001-143830/).

del Galdo, E. M., Williges, R. C., Williges, B. H., & Wixon, D. R. (1986). An Evaluation of Critical Incidents for Software Documentation Design An Evaluation of Critical Incidents for Software Documentation. In *Proceedings of the Human Factors Society 30th Annual Meeting* (pp. 19-23). Santa Monica, CA: Human Factors Society.

Ebling, M. R., & John, B. E. (2000). On the contributions of different empirical data in usability testing. In *Proceedings of Designing Interactive Systems '00: Processes, Practices, Methods, and Techniques* (pp. 289-296). New York, NY: ACM Press.

Hartson, H. R., & Castillo, J. C. (1998). Remote evaluation for post-deployment usability improvement. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '98)* (pp. 22-29). New York: ACM Press.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, No., 140*, 55.

McCreary, F. A. (2001). *InTouch Usability Evaluation* (Technical Report TR-01-11). Blacksburg, VA: Virginia Tech (available via http://eprints.cs.vt.edu:8080/archive/00000535/).

Page, C., & Rahimi, M. (1995). Concurrent and retrospective verbal protocols in usability testing: is there value added in collecting both? In *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting* (pp. 223-227). Santa Monica, CA: Human Factors and Ergonomics Society.

Thompson, J. A. (1999). *Investigating the Effectiveness of Applying the Critical Incident Technique to Remote Usability Evaluation.* Unpublished thesis, Virginia Tech, Blacksburg, VA (available via http://scholar.lib.vt.edu/theses/available/etd-121699-205449/).

Thompson, J. A., & Williges, R. C. (2000). Web-Based Collection of Critical Incidents During Remote Usability Evaluation. *Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting* (pp. 602-605). Santa Monica, CA: Human Factors and Ergonomics Society.