Evaluating REAL Users, using REAL Software, performing REAL Tasks, in REAL Contexts

Ilona Posner, Ronald Baecker, and Alex Mitchell*

Collaborative Multimedia Research Group — DGP, University of Toronto, 10 Kings College Rd. #4306, Toronto ON M5S 1A1 CANADA * ISS - Institute of Systems Science, National University of Singapore, Heng Mui Keng Terrace, Kent Ridge, Singapore 119597.

Our group carries out research on collaborative multimedia. We design, build, and test prototype software to aid people in working together on tasks such as writing, making movies, using the Internet, and managing information. We continually face the question: How do we study *real users* working with *real software* to perform *real tasks* in *real work contexts* over *real time frames?*

Evaluation methodologies from human-computer interaction and human factors (reviewed in Chapter 2 of Baecker et al., 1995) provide only modest assistance with this question. *Usability testing* (Nielsen, 1994) tends to be carried on in a laboratory on relatively prescribed tasks of limited duration. *Usability inspection* (Nielsen and Mack, 1994) makes use of the judgments of experts who typically examine the interface for relatively brief periods of time out of a real work context. *Contextual inquiry* (Holtzblatt and Jones, 1993) stresses real users in a real work context, but tends to focus on employing insights about work process into the design process.

None of these methodologies address our needs. Ideally, in order to gather the most information about real system usage, we would be:

- omnipresent
- remembering and able to reconstruct everything we see and hear including precise details about user actions and system responses
- so unobtrusive that we had no effect on the phenomenon we are trying to observe.

This is impossible; the question is how best to approximate this at a reasonable cost and with minimal interference to the work going on.

This paper first reviews some case studies in which we have tackled this problem over the past few years. We provide brief descriptions of the study details, data collected, analyses performed, and the problems encountered. The paper concludes with a summary of recommendations derived from our studies.

1. CASE STUDIES

Collaborative Writing Study — Prejudice Project: We organized an after school program for grade six students who worked in groups to produce a magazine about prejudice (Mitchell et al., 1995). Two groups of 4 students worked together at four networked computers during twelve one-hour weekly sessions. Students worked on-line using collaborative writing software; they also worked on-paper in small groups. We collected large volumes of data including: 100 hours of video (4 video cameras - 2 on people 2 on screens, audio recording enhanced using 4 microphones and an audio mixer), electronic records of documents, marked up paper documents, questionnaires, individual interviews, two teachers' blind evaluations of the final documents, and teachers' evaluations of the students performance in class. We performed qualitative data analysis by having 2 judges annotate 30 hours of video. Some of the pragmatic problems encountered during this project included hardware-software incompatibility, unexpected network conflicts and delays, organization of digital records across different machines, and unsynchronized computer clocks complicating digital record keeping.

Multimedia Authoring Study — MAD Camp: We ran a multimedia summer camp for grade 7 students working together to create motion pictures (Posner et al, 1997). The camp was run for two one-week sessions, lasting 5.5 hours per day over five days. Twelve campers attended each session working in groups of 3 campers and one counselor. Data collected included: four questionnaires administered throughout the camp sessions, paper diaries of group activities, audio journals recording group activities, video records of the moviemaking process (32 hours of videotape), paper artifacts, digital records (1.2 Gigabytes per group for 8 groups), group discussions, expert ratings of movie quality (technical, creative, and overall categories). Quantitative analysis of the data focused on software preference, movie structures, process information, counselor effects, and movie quality ratings. Qualitative analysis involved viewing of 12 hours of specially selected videotape for examination of the movie process, counselors' instruction and feedback effects, and determinants of success. Problems encountered on this project included hardware and software reliability, variability of counseling techniques, and counselor biases towards technology.

Information Visualization Study — TimeStore 2.0: A study of the usage of a time-based email management and visualization system (Silver, 1996). TimeStore II was used periodically during a four week evaluation period, by Eudora users. Data collected included: program instrumentation or logging allowing recording and playback of user interactions, "thinking aloud" audio recordings documenting the context of each usage, and follow up interviews with users. Qualitative data analysis was conducted by playing back user interactions captured by the logging data and the "think aloud" recordings. One problem encountered was user selection — all users worked in the HCI group of one computer company, had prior knowledge about this system, and had preconceived ideas about interactive prototypes in general; these expert interaction designers were unable to focus on usefulness of the system concept and instead primarily focused on the system's usability problems. Another problem concerned synchronization of audio recordings of user interactions and their corresponding digital records; without common time stamps on these, the synchronization becomes significantly more complex as the amount of data increases.

Information Management Study — TimeStore 3.0: A study of the usage of a redesigned time-based email management and visualization system (Yiu, 1997). Users used TimeStore 3.0 daily in their regular environments during a three-week evaluation period. Data collected include think-aloud sessions with screen and audio capture by the users' computers using Microsoft Camcorder (Microsoft, 1997), and weekly interviews with users. Qualitative data analysis was performed by viewing real-time play back of movie files containing the users' on-screen interaction and think-aloud audio. One problem encountered in this methodology was the lack of consistency in the think-aloud protocols; this was most evident from minimal commentary during routine operations. The limited functionality of the software, which for example did not provide a spell-check for email messages, lead to users adopting alternative methods for composing email messages, and consequently reducing their use of the prototype. Finally, since users were in charge of running Camcorder and storing space intensive movies of their interactions, storage space limitations reduced the total data recorded for analysis.

The following table (Table 1) summarizes and compares the evaluation methodologies that were used in the above studies. User experiences were captured using interviews, questionnaires, and journals, while interactions details were recorded using artifact capture, video, logging, and think-aloud recordings. Synchronization of recordings and logs was done mostly manually.

Table 1: Evaluation Methodologies Comparison of the Case Studies

	Collab Writing	**Multimedia**	**TimeStore 2**	**TimeStore 3**
Interviews	post project individual	post project group discuss	frequent? weekly	weekly
Questionnaires	pre, mid, end	pre, mid, end	none	none
Artifact Capture	papers, documents, notes, all video	all documents, selective video	excluding private materials	excluding private materials
User Journals	user experience recorded in daily journals	process info recorded in paper and audio journals	audio records of user feedback	session logs with think-aloud
Logging Method	complete video recording	selective video recording	data capture & audio recording	session logging with think-aloud
Synchronization	manual	manual	manual	automatic

2. RECOMMENDATIONS

Our experience with these and other studies of real software systems leads us to these recommendations which form the basis of a new methodology:

- Build technology with data analysis in mind — Event trackers can be easily incorporated into software during development but are very hard to implement after the development is complete. Off-the-shelf data recorders are helpful but only able to provide large scale details; these can also be disk space intensive and their analysis extremely time consuming.

- Iteratively design the study — Run pilot tests and pretest as much as possible with "similar users" to ensure that study design is sound, user instructions are clear, and user tasks are reasonable. Make adjustments and test them. Inadequate pre-testing can derail the entire study.

- Collect varied and redundant data — Collect maximal information using questionnaires, think-aloud protocols, user interviews, video records of interactions, but be careful not to overload and distract the users. Save product information including time stamped digital records, and all related artifacts such as paper notes, work diagrams, brainstorming lists.

- Save history of software interactions — Automatically log the history of user interactions for verifying software usage patterns and users' problems. Such logs are key to data analysis and can serve to focus retrospective discussions with users about their experiences.

- Use video recording selectively — Video tape provides a very rich record with relatively low cost data collection and storage. One caution is the high cost of in-depth video analysis even with support tools (Fisher and Sanderson, 1996). Selected samples of video can also be very helpful in understanding user interactions, especially if data logs determine sample selection.

- Record surrounding events in context — Try to record the context of the interaction not only machine readable ones so that it makes sense during the analysis which may be far removed from the experimental context. For example, directly recorded screen shots along with synchronized thinking-aloud protocols are extremely helpful for interpreting the user actions in context. This type of holistic data gathering produces a unified data source (a digital movie of user interaction) and greatly reduces the effort required to manage and analyze reams of data.

- Begin analysis immediately — If possible begin analysis while the study is still in progress. While the users are still accessible make sure all user intentions and problems are clearly documented, questionnaires are complete, and conflicting or confusing responses are clarified.

- Use visualization tools for improved analysis — Enhanced visualizations of the data, such as Timelines (Harrison, et al. 1994) and process visualizations (NUD-IST, 1994), can facilitate the management, grasping, and interpretation of data and experimental results.

- Consider the Internet as an evaluation tool — Internet technology adds a new dimension to evaluation. Software users can be far removed from their observers with the internet providing a high speed, high bandwidth link between them. In such cases the users may need to take a more active role in the data collection process by initiating recordings of what they consider to be critical events (Hartson et al., 1996). Collecting data invisibly yet ethically without interfering and changing the users' interactions with the system is an ongoing challenge for evaluators.

Running an original study or experiment with novel software is as much art as it is science. Regardless of how much preparation is done in advance, researchers must be vigilant and creative in observing the users, analyzing on the fly, and adapting their methodology in order to obtain maximal insight from the study.

References

Baecker, R., Grudin, J., Buxton, B. and Greenberg, S. (ed.) (1995). *Readings in Human Computer Interaction: Towards the Year 2000.* Second Edition, Morgan-Kaufmann.

Fisher, C. and Sanderson, P. (1996). Exploratory Sequential Data Analysis: Exploring Continuous Observational Data. *Interactions* 3 (2), March 1996, 25-34.

Harrison, B.L., Owen, R., and Baecker, R.M. (1994). Timelines: An Interactive System for the Collection and Visualization of Temporal Data. *Proceedings of Graphics Interface '94.*

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,* 228-235.

Holtzblatt, K., and Jones, S. (1993). Conducting and Analyzing a Contextual Interview. In *Readings in Human Computer Interaction: Towards the Year 2000,* Baecker, R., Grudin, J., Buxton, B. and Greenberg, S. (eds.), 1995, Second Edition, Morgan-Kaufmann.

Mitchell, A., Posner, I., and Baecker, R. (1995). Learning to Write Together Using Groupware. *Proceedings of CHI'95,* April 1995, 288-295.

Microsoft Corp. (1997). Camcorder. www.microsoft.com/msoffice/office97/camcorder/

Nielsen, J. (ed.) (1994). *Special Issue on Usability Laboratories of Behavior and Information Technology,* Vol 13 (1 & 2), January 1994.

Nielsen, J. and Mack, R.L. (eds)(1994). *Usability Inspection Methods.* John Wiley & Sons Inc.

NUD-IST (1994). Qualitative Solutions & Research Pty Ltd. Vic, Australia, nudist@qsr.latrobe.edu.au.

Posner, I., Baecker, R., and Homer, B. (1997). Children Learning Filmmaking Using Multimedia Tools. *Proceedings of ED-MEDIA/ED-TELECOM 1997,* Calgary, Canada.

Silver, N. (1996). *Time-based Visualization of Electronic Mail,* M.Sc. Thesis, University of Toronto.

Yiu, K. S., (1997). *Time-Based Management and Visualization of Personal Electronic Information*, M.A.Sc. Thesis, University of Toronto.