REMOTE USABILITY EVALUATION AT A GLANCE*

José C. Castillo, H. Rex Hartson, Deborah Hix
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0106
josec@hci.ise.vt.edu, {hartson, hix}@vt.edu

ABSTRACT

Much traditional user interface evaluation is conducted in usability laboratories, where a small number of selected users is directly observed by trained evaluators. However, as the network itself and the remote work setting have become intrinsic parts of usage patterns, evaluators often have limited access to representative users for usability evaluation in the laboratory and the users' work context is difficult or impossible to reproduce in a laboratory setting. These barriers to usability evaluation led to extending the concept of usability evaluation beyond the laboratory, typically using the network itself as a bridge to take interface evaluation to a broad range of users in their natural work settings.

* This paper is based on the third chapter of the thesis work by Castillo (1997).

DEFINITION OF REMOTE USABILITY EVALUATION

Remote evaluation is defined as usability evaluation where evaluators are separated in space and/or time from users (Hartson, Castillo, Kelso, Kamler, and Neale, 1996). For consistency of terminology throughout this report, the term *remote*, used in the context of remote usability evaluation, is relative to the developers and refers to <u>users</u> not at the location of <u>developers</u>. Similarly, the term *local* refers to location of the developers. Sometimes developers hire outside contractors to do some usability evaluation in a usability laboratory at the contractor's site. Neither term (local or remote) per the above definitions, applies very well to these third-party consultants, but they (as surrogate developers) could have remote users.

In traditional laboratory-based usability evaluation, users are observed directly by evaluators. However, remote and distributed location of users precludes the opportunity for direct observation in usability evaluation. Therefore, with remote evaluation (Figure 1), the network serves as a bridge between users and evaluators, taking interface evaluation to a broad range of networked users (e.g., representative population of users) in their natural work settings.

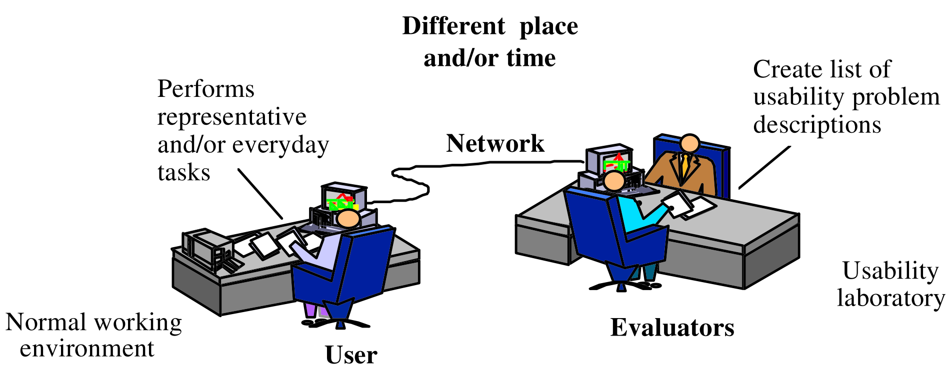

Figure 1. Scenario of a remote usability evaluation session

TYPES OF REMOTE EVALUATION METHODS

Identifying and distinguishing various approaches to remote evaluation is useful to understand the range of possibilities and to avoid comparison of unlike methods. No claim is made for the completeness of this list of types.

- Commercial usability services
  - Third-party laboratory evaluation
  - Third-party usability inspection
- Remote questionnaire or survey
- Collaborative remote evaluation
- Video-conferencing-supported evaluation
- Instrumented or automated data collection for remote evaluation
- User-reported critical incident method

Some remote evaluation situations call for a portable usability evaluation unit, by means of which the laboratory is taken to users in their normal work environment. Portable units often contain special usability evaluation equipment, including a laptop computer, video, and audio equipment. While this is an area of growing interest, it is outside the scope of this work.

COMMERCIAL USABILITY SERVICES

A number of commercial usability evaluation services are now available to software developers. Developers use the network to communicate design documents, software samples, and/or prototypes to remote contractual evaluators, but the network is not used to connect to remote users. The evaluation is local to the contractual evaluators and remote from the developers, with results being returned via the network.

Contractual evaluation services can be a good bargain for developer groups who have only occasional need for usability evaluation and cannot afford their own facilities. However, the quality of these services can vary, the methods can be ad hoc (e.g., intuitive inspection without use of guidelines), and the process is not always suitable for the specific needs of a development group. Two possible variations of local evaluation at remote sites are described below.

Third-party laboratory evaluation

Formal laboratory-based usability evaluation offered by consulting groups provides usability evaluation with representative users and tasks. Results include quantitative performance measures, user opinions and satisfaction ratings, recommendations for application improvement, and sometimes even a copy of evaluation session videotapes for review by the development team.

Third-party usability inspection

Some developers send their designs to remote contractors who perform local evaluation using ad hoc, intuitive interface inspection, drawing on design guidelines, user profiles, and software standards. Without empirical observation of users and a more formal process, results can vary depending on the knowledge and skills of the evaluators. As an example, this kind of evaluation is done by Vertical Research, Inc. (1997) for Microsoft® Windows™-based products.

REMOTE QUESTIONNAIRE OR SURVEY

A software application can be augmented to display a user questionnaire to gather subjective user preference data about the application and its interface. Appearance of the questionnaire, requesting feedback related to usability, is triggered by an event (including task completion) during usage. Responses are batched and sent to the developers. As an example, the User Partnering (UP) Module® from UP Technology® (Abelow, 1993) uses event-driven triggers to “awaken” dialogues that ask users questions about their usage. Approaches based on remote questionnaires have the advantage that they capture remote user reactions while they are fresh, but they are limited to subjective data based on questions pre-written by developers or evaluators. Thus, many of the qualitative data normally acquired in laboratory evaluation (i.e., the data directly useful in identifying specific usability problems) are lost.

COLLABORATIVE REMOTE EVALUATION

The distributed location of users and evaluators yields the opportunity for collaborative usability evaluation via the network (Hammontree, Weiler, and Nayak, 1994). Evaluators (at a usability laboratory) and remote users (at their natural work setting) are connected through the Internet and/or a dial-up telephone line, and use commercially available software (e.g., Microsoft® NetMeeting™, Farallon® Timbuktu Pro™) to assist them during remote evaluation of the user interface, supporting both synchronous and asynchronous collaboration. Typical tools available in such applications include:

- real-time application sharing (e.g., the capability to open a word processor at the user's computer and share it with the evaluator, even when the evaluator does not have that application installed, such that the user and evaluator take turns to modify a document),
- audio conferencing support for user and evaluators to talk during the evaluation session (e.g., audio could be sent via an Internet telephone tool or regular telephone),
- shared whiteboard with integrated drawing tool for viewing and editing graphic images in real time (e.g., pointing out specific areas by using a remote pointer or highlighting tool, or taking a "snapshot" of a window and then pasting the graphic on whiteboard surface), and/or
- file transfer capabilities (e.g., the capability of a user to transfer a document including snapshots of the user interface being evaluated to the evaluators' computer).

TeamWave® Workplace™, an example of a collaborative tool that could be used for remote evaluation, includes a shared space or virtual room for a work group. As in the case of real-life team meeting rooms, a virtual room may contain several objects (e.g., documents) relevant to the team's work. The user and evaluator can be present in the room at the same time, so they are working together synchronously in real-time, or they may be working in the room at different times, so the user can leave things in that room for the evaluator to analyze asynchronously. This type of asynchronous collaboration gives the evaluator the flexibility to do the evaluation all at once or at a time that is convenient.

VIDEO-CONFERENCING-SUPPORTED EVALUATION

When the user and evaluator are in adjacent rooms, local usability evaluation requires a video/audio cable (or wireless communication) between the rooms. Users located more remotely can be connected to evaluators using the network and video conferencing software, as an extension of this video/audio cable (Hartson et al., 1996). This kind of remote evaluation, using video teleconferencing over the network as a mechanism to transport video data in real time, perhaps comes the closest to the effect of local evaluation. Currently, the primary obstacle to this approach is the limited bandwidth of the network, occasioning communication delays and low video frame rates.

INSTRUMENTED OR AUTOMATED DATA COLLECTION FOR REMOTE EVALUATION

An application and its interface can be instrumented with embedded metering code to collect and return a journal or log of data occurring as a natural result of usage in users' normal working environments. Data captured by such applications (e.g., WinWhatWhere® WinWhatWhere™) represent various user actions made during task performance, are often very detailed, and include logging of:

- program usage (e.g., which programs users utilize most frequently),
- project time (e.g., time spent working on specific application),

- Internet usage (e.g., monitor use of on-line time),
- comments to the system,
- keystrokes and mouse movements, and
- any other activity, producing custom-built reports.

The logs or journals of data are later analyzed using pattern recognition techniques (Siochi and Ehrich, 1991) to deduce where the usability problems have occurred. This approach has the advantage of not interfering at all with work activities of the user and can provide automated usability evaluation for certain kinds of usability problems. However, for formative evaluation it can be difficult to infer some types of usability problems effectively. This method has been used successfully for summative evaluation, marketing trials, and beta evaluation.

USER-REPORTED CRITICAL INCIDENT METHOD

This method applies selective data collection triggered directly by users while performing tasks in their normal work context (Hartson et al., 1996). Users are trained to identify critical incidents and report specific information about these events (e.g., description and severity of problem). The reports are transmitted to developers along with context information about the user task (i.e., what user was trying to do when problem occurred) and the system itself (e.g., name and/or location of screen where problem occurred), as well as clips containing screen sequence actions. Evaluators use these data, approximating the qualitative data normally taken in the usability laboratory, to produce usability problem descriptions.

The user-reported critical incident method for remote evaluation, which is the subject of the work reported here, has potential for cost-effectiveness, since the user gathers the data and evaluators look (at least in theory) only at data that relate to usability problems.

CLASSIFICATION OF REMOTE EVALUATION METHODS

Remote usability evaluation encompasses several methods for evaluating user interfaces at a distant location. Since it is useful to understand how each method works under different situations, this section presents a classification that distinguishes among the characteristics of each remote evaluation method. The classifications made here are not absolute, but are relative and representative, and are based on insight gained during this study, offered to promote an intuitive comparison of the methods. The following attributes of remote usability evaluation methods are considered:

- type of users involved,
- time of evaluation,
- user location during evaluation,
- person or role who identifies critical incidents and/or problems during task performance,
- type of tasks (user’s own tasks or tasks predefined by evaluator),
- level of interaction between user and evaluator,
- type of data gathered,
- type of equipment used for collecting data (e.g., videotape),
- cost to collect data,

- cost to analyze data and create usability problem descriptions, and
- quality or usefulness of collected data.

Laboratory-based usability evaluation is included in each classification as a benchmark method for comparison with the remote evaluation methods.

TYPES OF USERS INVOLVED

The diagram in Figure 2 characterizes typical users involved in each type of remote evaluation described in the previous section. In the user-reported critical incident method, remote questionnaire, and instrumented or automated data collection, only real users participate in remote evaluation, as seen in the left of Figure 2. Collaborative evaluation, video-conferencing-supported evaluation, and third-party laboratory evaluation (center of the figure) can use both real and/or representative users during remote evaluation. Finally, third-party usability inspection is the only method that does not involve users during evaluation (evaluator performs intuitive or heuristic inspection of the user interface).

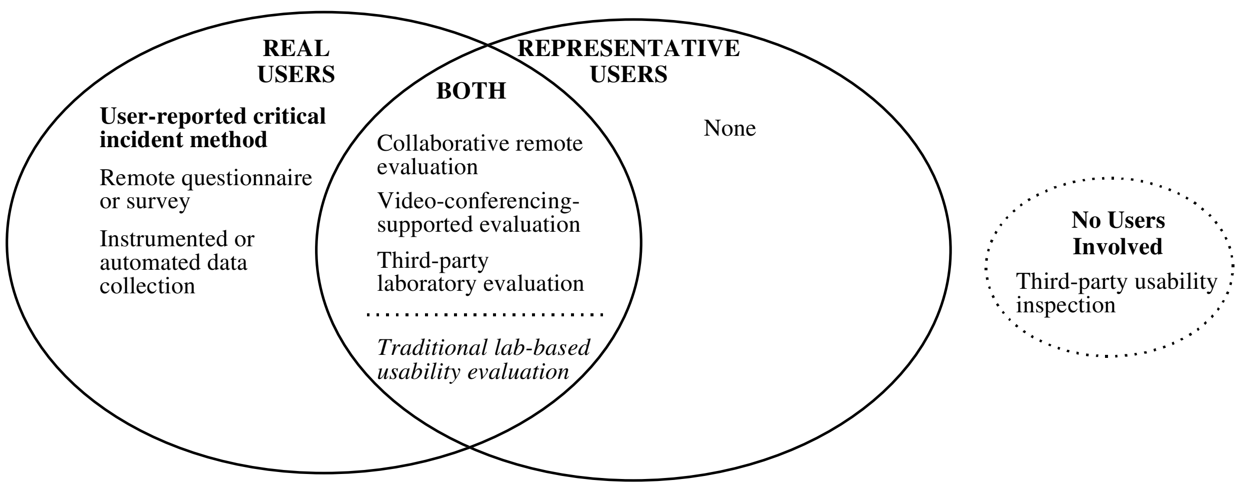

Figure 2. Characterization of types of users typically involved in remote evaluation

TIME AND USER LOCATION

Johansen (1988) distinguished several different situations of group work along the dimensions of time and place. Adapting Johansen's work, Figure 3 presents different situations that can occur during task performance with remote evaluation along the dimensions of time and user location (place).

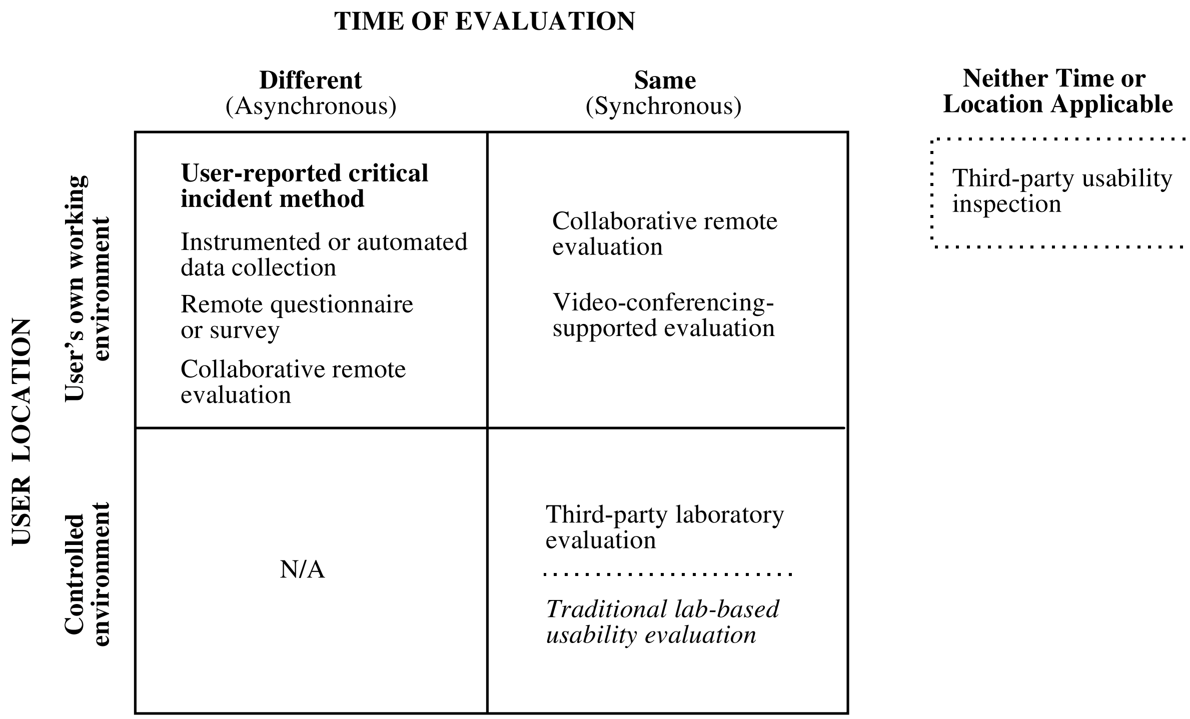

Figure 3. Characterization of user location and time of evaluation

In the user-reported critical incident method, instrumented data collection, and remote questionnaire, remote evaluation occurs within the user's normal work setting. Time of evaluation is asynchronous, meaning that evaluators participate at a different time than do users. User critical incident reports, log files, and questionnaires, respectively, are collected during usage, and evaluators later analyze them to create a list of usability problem descriptions. Third-party laboratory evaluation involves local evaluation with representative users at the third-party contractor/evaluator's usability laboratory (i.e., controlled environment). As in traditional lab-based usability evaluation (same time/controlled environment quadrant), users and evaluators are located in the same place and time.

Because of the generality of the concept, collaborative remote evaluation is the only method than spans across more than one quadrant. Generally, collaborative remote evaluation occurs when user and evaluator are connected in real-time (i.e., synchronously) over the network. The evaluator can also use software (e.g., Microcom® Carbon Copy™) to remotely control the user's computer to observe the user interface. In the asynchronous mode for this method, the user can leave information (e.g., snapshot of one screen of the user interface) in a place known by the evaluator (e.g., documents directory), and the evaluator later connects to the user's computer to retrieve the information.

Video-conferencing-supported evaluation involves local usability evaluation, where the user and evaluator are examining the user interface at the same time from different locations (e.g., adjacent rooms or distant geographical locations). In this method the video and audio cable (or wireless or network communication) operates as an extension of the usability laboratory. Third-party usability inspection is not included in the table on the left of the figure because it does not involve users during evaluation of the user interface.

In the current research, the "different time / user's own working environment" quadrant was of most interest because the "different time" characteristic means the evaluator is not present, making the cost of data gathering very low (relative to traditional lab-based usability evaluation). The "user's working environment" characteristic can yield more realistic qualitative data, since occurs within the user's natural work setting.

PERSON WHO IDENTIFIES CRITICAL INCIDENTS AND PROBLEMS DURING TASK PERFORMANCE

The diagram in Figure 4 presents a characterization of the person who identifies and/or reports critical incidents and/or usability problems during evaluation of the user interface.

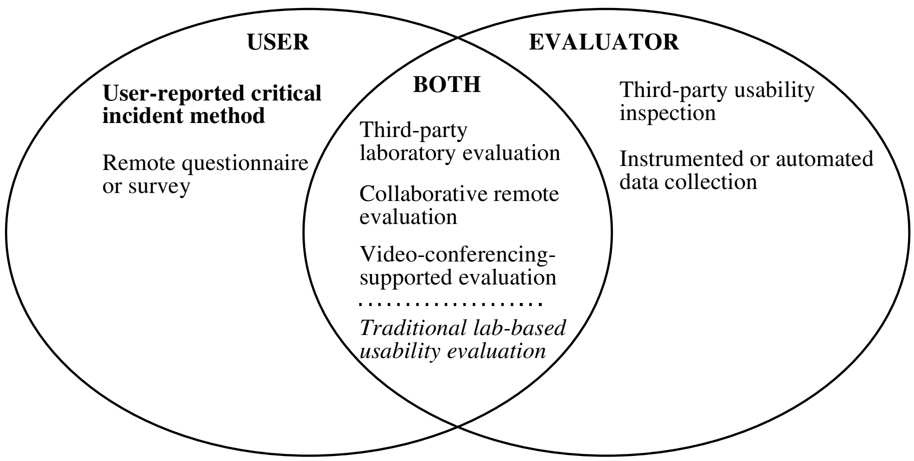

Figure 4. Characterization of person who identifies and/or reports critical incidents and/or usability problems during task performance

In the user-reported critical incident method, the user both identifies and reports critical incidents during task performance and the evaluator analyzes these reports to create a list of usability problem descriptions. When answering remote questionnaires or surveys, users provide subjective data and may indicate problems that they encounter during task performance. Video-conferencing-supported evaluation, collaborative remote evaluation, and third-party laboratory evaluation involve both the user and the evaluator (trained in human-computer interaction and usability methods) to identify critical incidents and/or usability problems during task performance. For instrumented data collection, the evaluator analyzes log files of usage data and deduces possible usability problems encountered by users. Finally, third-party usability inspection exclusively involves only the evaluator in identifying and reporting critical incidents during task performance.

TYPE OF TASKS AND LEVEL OF INTERACTION DURING REMOTE EVALUATION

The diagram shown in Figure 5 characterizes various situations that can occur along dimensions for type of tasks and level of interaction between users and evaluators during remote evaluation. Unlike Figure 3, none of the remote evaluation methods span quadrants of this table.

LEVEL OF INTERACTION BETWEEN USERS AND EVALUATORS

TYPES OF TASKS	None	Significant
Real or everyday tasks	**User-reported critical incident method** Instrumented or automated data collection Remote questionnaire or survey	none
Representative tasks (predefined by evaluator)	none	Third-party laboratory evaluation Collaborative remote evaluation Video-conferencing-supported evaluation *Traditional lab-based usability evaluation*

Neither type of tasks nor level of interaction applicable

Third-party usability inspection

Figure 5. Characterization of types of tasks performed by users and level of interaction between users and evaluators during usability evaluation

The top left quadrant of this table contains methods (user-reported critical incident method, instrumented data collection, and remote questionnaire or survey) in which remote evaluation occurs while users perform real or everyday tasks, with no interaction with evaluators during task performance and/or evaluation. Similar to laboratory-based usability evaluation, the lower right quadrant includes remote evaluation methods (third-party laboratory evaluation, collaborative remote evaluation, and video-conferencing-supported evaluation) in which users generally perform representative tasks (predefined by evaluator) and significantly interact with evaluators during usability evaluation. Again, third-party usability inspection is located outside the table on the left of the figure because this method does not involve users during evaluation of the user interface.

TYPES OF DATA GATHERED

Table 1 shows the various types of data gathered within each remote evaluation method. With the exception of instrumented data collection, remote questionnaires or survey, and third-party usability inspection, most remote evaluation methods typically gather three or more types of qualitative data. A check in a cell of Table 1 indicates that the corresponding method has the capability to gather the associated type of data, but does not indicate that this kind of data is always used or can be used at low cost. More types of data involved typically imply more equipment required for data gathering. On the other hand, for evaluators, gathering different types of data may be important to ensure that the most useful data are captured (e.g., context information about critical incidents). For example, in the case of video-conferencing-supported evaluation, evaluators can obtain several different kinds of data (e.g., videotape showing the user's face and including audio of user comments, videotape of screen action recorded via scan converter, evaluator notes taken during evaluation session, subjective information obtained from an online satisfaction questionnaire). In contrast, instrumented methods yield only logs of user actions and commands, and the remote questionnaire method yields only subjective opinion of users.

Table 1. Types of data gathered by each remote evaluation method

	Continuous videotape of entire session	Audio with user comments	Evaluator notes from evaluation session	Log files, automatically recorded	Data obtained from collaborative work	Critical incident data (reports)	Critical incident data (video clips)	Subjective data from user questionnaire
Remote questionnaire or survey								√
Instrumented or automated data collection				√				
Third-party usability inspection			√					
User-reported critical incident method		√				√	√	√
Third-party laboratory evaluation	√	√	√				√	√
Video-conferencing-supported evaluation	√	√	√				√	√
Collaborative remote evaluation	√	√	√		√		√	√
Traditional lab-based usability evaluation	√	√	√				√	√

TYPE OF EQUIPMENT USED AND QUANTITY OF DATA GATHERED

Figure 6 illustrates how the equipment required to collect data and the quantity of data collected vary among remote evaluation methods.

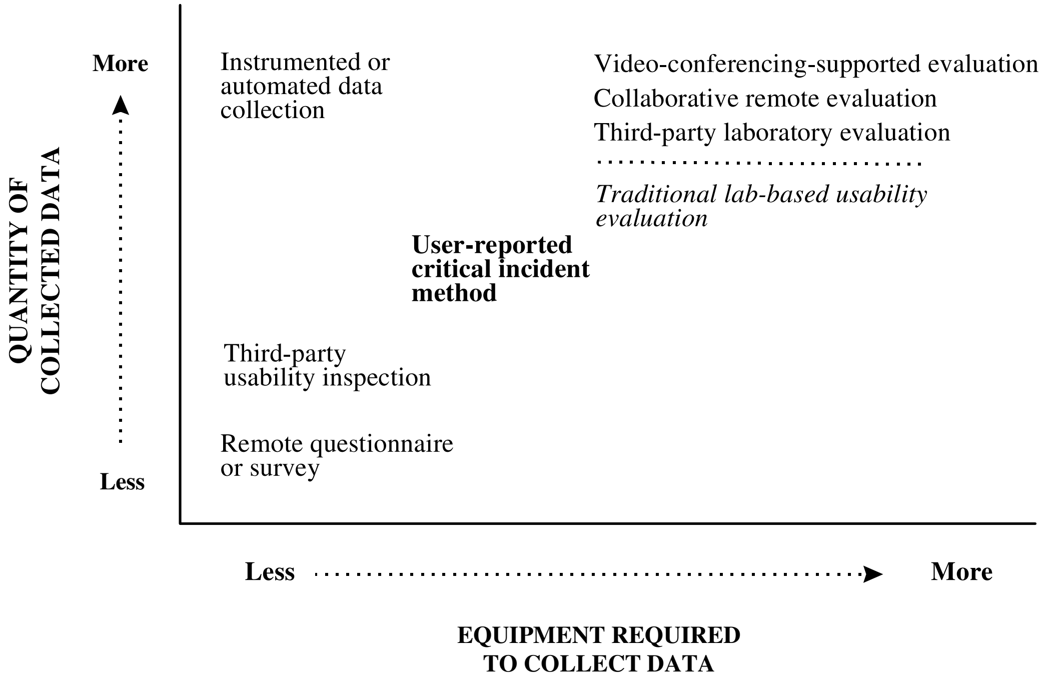

Figure 6. Equipment required to collect data and quantity of data gathered by each remote evaluation method

To illustrate the concept, video-conferencing-supported evaluation, collaborative remote evaluation, third-party lab evaluation, and traditional lab-based evaluation all require a relatively high level of equipment (e.g., video and audio equipment) to collect data and each provides a large quantity of data. Therefore, this group of similar methods appears in the upper right portion of Figure 6. In contrast, instrumented approaches require a low level of equipment (e.g., embedded metering code to collect a log of data including information such as keystrokes and mouse movements) but also yield large quantities of data. Thus, instrumented methods appear in the upper left portion of the figure.

Remote questionnaires and surveys require almost no added equipment, but produce only a small amount of data. This constraint on the amount of data stands for the need to minimize workload for users and interference with task performance. Remote questionnaires typically contain relatively few questions, not extensive surveys. Therefore, this method appears in the lower left of the figure. The user-reported critical incident method requires somewhat less total equipment and produces somewhat less data than traditional lab-based evaluation, so this method appears roughly in the middle of the figure.

COST TO COLLECT AND ANALYZE DATA

Figure 7 illustrates the relationship between cost to collect data and cost to analyze it, for each remote evaluation method. Cost is limited to those incurred by an "in house" development team (mainly referring to evaluators). Developers are only "in house" personnel" and not, for example, a third party retained to perform external evaluation. Costs to the development team, which, however, include fees to third-party evaluators, include:

- equipment and technology,
- time and effort (person-hours),

- external services (third party),
- training (all, including developer and user training for evaluation),
- travel and time of user-subjects, and
- travel and time for developers to visit user sites.

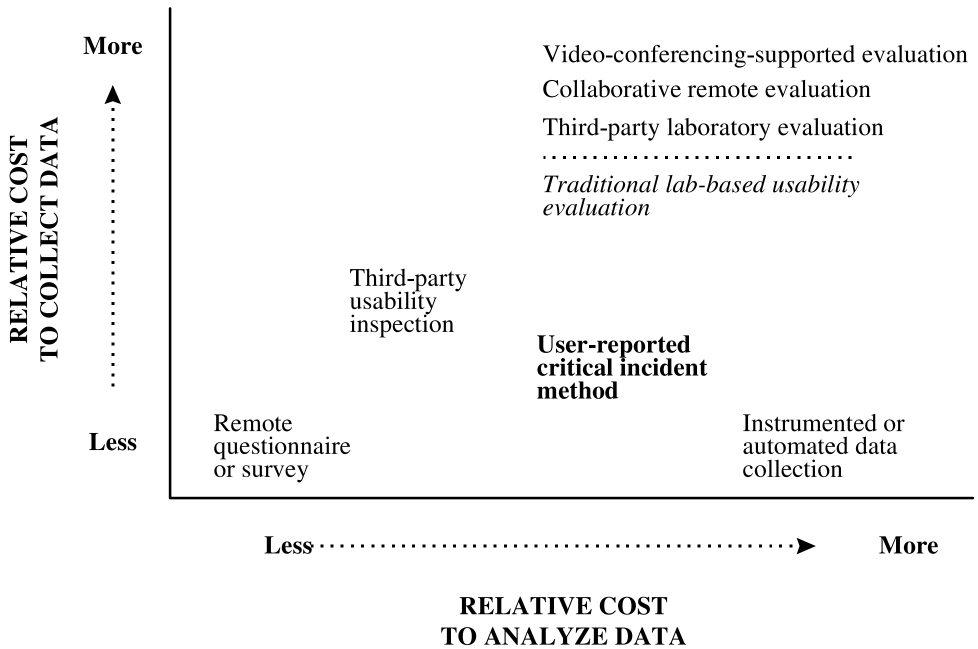

Figure 7. Relative costs to collect and analyze data

Remote questionnaire or survey appears again in the lower left of the figure because the cost to collect data are low and, since there is a small amount of data, the cost to analyze it is also low. The group of methods similar to traditional lab-based evaluation appear again in the upper right because the cost for both collection and analysis of data are high relative to the other methods. Instrumented data collection methods are somewhat an extreme case in that they are probably the lowest in cost to collect data and possibly among the highest to analyze them. In the user-reported critical incident method, the cost for collecting data are a little higher than instrumented approaches but significantly lower than the traditional-lab based method, and is approximately the same as the lab-based methods for analysis, because analysis is very similar. Interestingly, third-party usability inspection falls in the middle, making it a good compromise in terms of cost when resources are limited.

Figure 8 shows a comparison of total data cost (i.e., collection and analysis) for a single project against quality or usefulness of data in identifying usability problems. Remote questionnaire or survey appears again in the lowest left of the figure, and once more, the group related to traditional lab-based evaluation appear in the upper right because they are the most costly but also yield the most useful data. The user-reported critical incident method is somewhat lower on each axis, costing less but producing data that are almost as useful. Third-party usability inspection appears somewhat in the middle of the graph, once again representing a possible compromise when resources are limited.

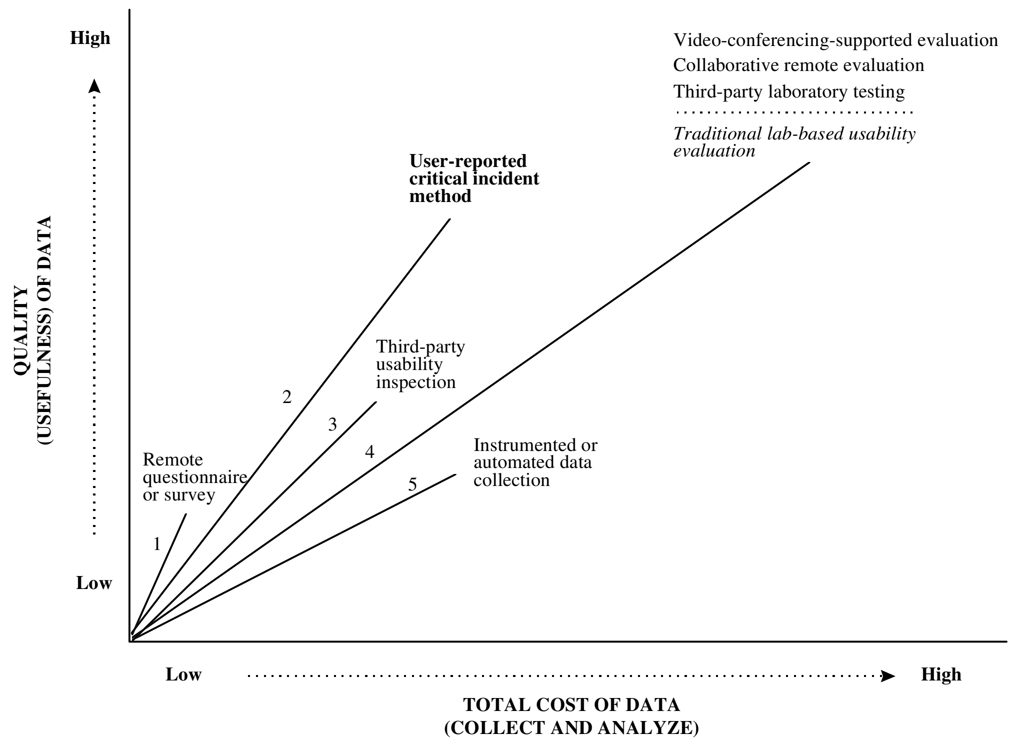

Figure 8. Quality of usability data by remote evaluation method

The interesting aspect of Figure 8* is the slope of the lines from the origin to each method. This slope is a measure of a kind of efficiency, namely the quotient of quality or usefulness per unit cost, of data collected by each method. In the graph, a higher slope suggests higher efficiency. Interestingly, remote questionnaire or survey has the steepest slope in the graph. Thus it has high efficiency. However, the high slope comes from a quotient from low usefulness and low cost. This limited usefulness may not alone be sufficient for the needs of the development team. On one hand, if the budget is extremely limited, some data are better than nothing. But a questionnaire only produces subjective data and, as Elgin (1995) states, "Subjective feedback is generally harder to interpret than objective feedback in a known setting…". The next most efficient method according to the graph is the user-reported critical incident method. It was, in fact, a goal of this research to develop critical incident reporting as an efficient remote evaluation method, one that was less costly than traditional lab-based evaluation but still maintained its usefulness in determining usability problems. Third-party usability inspection comes next, perhaps, again as a middle-of-the-road possibility.

Group 4, the group of remote evaluation methods related to traditional lab-based usability evaluation, appears a bit less of a bargain, despite the high usefulness of data, because of the higher cost. Because costs portrayed here are costs for one project, they include amortization of equipment, lab facilities, and training costs over several projects. Thus, the relative difference between the slopes of method 3 and the

* Of course, this graph is based on intuitive reasoning and represents general characteristics of the methods as qualified by the rationale set forth in this chapter. In specific instances of these methods as used by particular development groups, the results can vary.

group represented by 4, and indeed the methods within group 4, can vary depending on how these fixed costs are amortized. In particular, if they are not well amortized, third-party laboratory evaluation or third-party usability inspection might be more attractive than setting up a traditional lab-based facility. Additionally, both traditional lab-based approaches and third-party usability evaluation can be quite expensive in terms of travel and time for representative users to visit the laboratory. Thus, these two methods of group 4 could have decreased slopes due to this factor.

REFERENCES

Abelow, D. (1993). Automating Feedback on Software Product Use. *CASE Trends* (December), 15-17.

Castillo, J. C. (1997). *The User-Reported Critical Incident Method for Remote Usability Evaluation.* Unpublished thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA.

Elgin, Bruce (1995). Subjective Usability Feedback from the Field over a Network. *SIGCHI Bulletin* (October), 27 (4), 43-44.

Farallon Communications, Inc. (1997). Farallon Home Page. Title: Timbuktu Pro 1.1 for Windows. Internet WWW page at: <http://www.farallon.com/product/tb2/tb2winover.html> (Accessed: 7/1/97).

Hammontree, M., Weiler, P., & Nayak, N. (1994). Remote Usability Testing. *Interactions* (July), 21-25.

Hartson, H.R., Castillo, J.C., Kelso, J., Kamler, J., & Neale, W.C. (1996). Remote Evaluation: The Network as an Extension of the Usability Laboratory. *Proceedings of CHI '96 Human Factors in Computing Systems*, 228-235.

Johansen, R. (1988). *Groupware: Computer Support for Business Systems.* New York: The Free Press.

Microcom, Inc. (1997). Microcom Home. Title: Carbon Copy™. Internet WWW page at: <http://www.microcom.com/main/ProductsTech/remoteaccess/cc.htm> (Accessed: 7/1/97. Updated: 6/25/97).

Microsoft Corporation (1997). Microsoft® NetMeeting™ Complete Internet Conferencing. Title: Overview. Internet WWW page at: <http://www.microsoft.com/netmeeting/> (Accessed: 7/1/97. Updated: 6/10/97).

Siochi, A.C. & Ehrich, R.W. (1991). Computer Analysis of User Interfaces Based on Repetition in Transcripts of User Sessions. *ACM Transactions on Information Systems* (October), 9 (4), 309-335.

TeamWave Software Ltd. (1997). TeamWave Home. Title: TeamWave Workplace Overview. Internet WWW page at: <http://www.teamwave.com/overview.html> (Accessed: 7/1/97. Updated: 6/20/97).

Vertical Research, Inc. (1997). Vertical Research Home Page. Title: Remote Inspections. Internet WWW page at: <http://www.nolan.com/~pnolan/rem_insp.html> (Accessed: 7/1/97. Updated: 6/10/97).

WinWhatWhere Corporation (1997). Welcome To WinWhatWhere. Title: What does W^3 do? Internet WWW page at: <http://www.winwhatwhere.com/bsiwww1.htm> (Accessed: 7/1/97).

TeamWave Software Ltd. (1997). TeamWave Home. Title: TeamWave Workplace Overview. Internet WWW page at: <http://www.teamwave.com/overview.html> (Accessed: 7/1/97. Updated: 6/20/97).

Vertical Research, Inc. (1997). Vertical Research Home Page. Title: Remote Inspections. Internet WWW page at: <http://www.nolan.com/~pnolan/rem_insp.html> (Accessed: 7/1/97. Updated: 6/10/97).

WinWhatWhere Corporation (1997). Welcome To WinWhatWhere. Title: What does W^3 do? Internet WWW page at: <http://www.winwhatwhere.com/bsiwww1.htm> (Accessed: 7/1/97).