Reprint of paper from *UPA 97 Proceedings*,
(Monterey, CA), published by Usability
Professionals Association, Inc., 1997.

Multiple-User Testing: When One Person Can't See Everything

Laurie Kantner, Deborah Hinderer, and Connie Leas
Tec-Ed, Inc.

This session presents two case studies about designing usability tests where constraints prevented observation of all participants by one person.

1. The usability team needed to run twelve test sessions concurrently, with participants in different rooms.
2. The usability team needed to run a live teleconference, with participants in different rooms.

The goal of the presentation is to explain the methods we used, describe the pros and cons of our approaches, and elicit the opinions and experiences of audience members.

Case Study 1: 36 Sessions in One Day

A large insurance organization wanted to conduct exploratory usability testing of Win95 prototype versions of four DOS programs its field agents use to elicit information from customers and then generate proposal and contractual documents. The development team wanted early feedback about the Win95 implementation before proceeding with coding.

Goals of the Usability Testing

The development team had three goals:

- Test the software with a select group of agent representatives who had participated in requirements definition.
- Conduct the tests during a home office visit by these representatives, who reside throughout the country.
- Gain immediate buy-in from the representatives of strategies for solving problems found by usability testing.

Methodology Selected to Meet the Goals

Within a two-day home office visit by the selected participants, we could not run one-at-a-time test sessions of all four software programs. Therefore, we decided to run concurrent sessions for each software program. Combining two of the software programs into one test, we ran a total of three tests, each in 12 concurrent sessions. This approach met the goals of the organization by scheduling the tests to coordinate with the home office visit.

Problems Posed by the Methodology

The people available to administer the test would be software developers and business analysts who create the products, as well as members of the organization's software testing group. The usability team recognized the following methodology problems:

- Most of the people available to do test administration were untrained in usability principles and practices.
- Multiple administrators mean the testing experience would vary for the different participants.
- People involved in creating the software being tested would be administering or observing sessions and would therefore be at risk to introduce bias.

The UPA annual conference (July 1996) session entitled "The More the Merrier? Methods for Multiple-User Testing" gave timely information about methodological experiments other organizations were conducting. Those presentations confirmed our general approach and alerted us to potential pitfalls.

Solutions Developed to Solve the Methodological Problems

To reduce the impact of the methodological problems, we:

- Created session scripts that minimized administrator interaction with participants.
- Assigned a separate observer/note-taker for each session.
- Provided one day of usability training for the 24 administrators and note-takers.
- Used questionnaires and observer note-taking forms to collect data that would indicate where test administrators might be introducing bias.
- Spot-checked actual sessions to give reinforcing feedback.

We recognized that these efforts would reduce the negative effects we were trying to avoid, not eliminate them. The organization decided this tradeoff was acceptable, compared to using fewer participants, asking the user representatives to extend their visit, or visiting the representatives at their offices.

Minimizing Administrator Interaction with Participants through Script Design. In these exploratory usability tests, we wanted to collect as much qualitative data as possible to gain insight into the users' mental models and preferences. We wanted the administrators to provide neutral prompting to encourage participants to think aloud. With different individuals administering, we were concerned about the consistency of prompting. With developers administering, we were concerned about prompting that would "lead the witness."

We developed scripts that reduced the effects of multiple administrators. These scripts:

- Gave the administrator the language to use for moving the participant from activity to activity.
- Contained task handouts describing the activities to perform, which the administrator gave to the participant.
- Provided the neutral prompt language for specific circumstances.

The scripts were a joint effort, where the client subject-matter specialists defined the typical tasks on which the testing was based and the usability specialists crafted the blueprint that guided the sessions. Although the script included some space for the administrator to take notes, note-taking was the job of a second person observing the session, who used another version of the same materials.

Assigning Separate Observers/Note-takers and Designing Note-Taking Forms. Because we were concerned about the effect of people administering usability sessions for software they created, it was agreed that each usability session would have two people in attendance: the administrator and a separate observer/note-taker. In fact, the note-taker was more likely to be the software developer, while the administrator was either the business analyst or someone from software testing. To encourage quantifiable note-taking and minimize subjective interpretations, the note-taking forms provided columns for degrees of success, into which the note-taker placed checkmarks; the forms also provided spaces in which the note-taker could add information explaining participant problems. The note-taker did not interact with the participant.

Identifying Bias Effects. A column entitled "Facilitator Directs" was provided as a "check and balance" measure, where the note-taker could indicate when the test administrator gave leading hints to guide a participant through an activity. As a separate check-and-balance, participant task questionnaires included a question for each task asking whether the participant had received assistance from the administrator.

Conducting One-Day Usability Training. The final component of preparing for the usability tests was to train the 24 designated administrators and note-takers on the basic principles of usability testing, observe practice sessions, and give administrators and note-takers feedback on their first attempts at performing these new activities. The training day began with a 1-hour lecture describing usability principles and procedures and introducing everyone to the test design and materials.

Following the lecture, teams of two trainees practiced administering and observing partial sessions in 1-hour blocks, during which 30 to 40 minutes were spent in actual practice and 20 to 30 minutes were spent sharing feedback and Q&A with the usability specialists and peers. The usability specialists identified gradual improvement during the day. We also collected suggestions for how to make the materials easier to work with for the final sessions.

Test-Day Spot-Checking. On the testing day, one usability specialist returned to the company to observe sessions by each team. Sometimes the spot-checking was of an entire session, but most of the time it was for a half session. The specialist would give feedback privately; she simply needed to advise the administrator and note-taker to sit closer to the participant so they could see what the participant was doing (participants were using laptop computers). The spot-checking confirmed that the lessons taught had been learned and that the materials were serving well.

The usability specialist also participated in one of the three post-testing analysis sessions that occurred at the end of the testing day. In these sessions, the administrators and note-takers together discussed their observations, referring to note-taking forms and to the questionnaire data, which was tabulated at the same time. From these discussions, the testing teams formulated lists of findings and recommendations.

Feedback about the Process

Using the findings and recommendations developed at the post-testing analysis sessions, the testing teams presented “read-outs” of the results to the test participants to gain their concurrence about what happened and to discuss alternatives for correcting the problems uncovered during testing. Interestingly, the first read-out session encountered the most negative reactions from participants, and perhaps served as a model for how *not* to conduct the next two read-outs, because the presenter focused on the positive results and glossed over the negative results. Participants disagreed quite vigorously with the conclusions. It turned out that participants believed the software would handle their unusual procedures adequately but at the price of making their most typical procedures extremely time-consuming and awkward. Subsequent presenters were sensitized to the possibility of similar participant reactions, and their read-outs were more positively received.

The company concluded that the process worked quite well, and that they collected valuable data for improving their software programs. The usability specialists believe we achieved our goals of minimizing the effects of multiple administrators and developer bias on the test results. We also educated a large group of developers on the process and benefits of usability research.

What would we do differently in the future? Perhaps we would push harder for the software developers to observe the sessions for the programs they did not develop. The software developers believed they needed more knowledge of the other programs to be good observers. We disagree, but we did not campaign early enough to develop people’s expectations of how it would work. This approach would have further removed bias and it would have generated some fruitful cross-pollination (and perhaps across-program consistency) among the development teams.

Case Study 2: A Teleconference of Seven

A teleconference system company wanted to test the usability of the voice interface of a new telephone conferencing product “question and answer” feature. With this feature, a teleconference “meeting moderator” controls who gets to speak during phone meetings and when.

The test design consisted of two slightly different telephone meetings, moderated by a usability specialist. In each test session, seven participants attended the two meetings and played distinct roles in the meetings. The study included two of these two-meeting sessions.

To maintain verisimilitude, we placed each participant in a separate room, and we did not interrupt the meetings to ask them questions. Because the study budget did not permit us to assign a separate observer to each participant, one observer circulated among the rooms to observe participants during sessions.

Methodology

Test scenario. To create a realistic situation that would engage a diverse group of people—from students and homemakers to business people and retirees—we chose meeting topics we felt most people would have interest in, opinions on, or questions about. Each session consisted of a town meeting, in which the topic was a proposal for the development of a mobile home park; and a corporate meeting, in which the topic was a new health care benefits program.

Test tasks. To encourage active dialog at the meetings, participants were given suggested roles and "feeder" statements that could spark their imagination or help them voice their own views more freely. Participants joined the meetings by phone, listened to the meeting moderator introduce the topics, put themselves in line to speak at the meeting, and asked questions or made comments when they were called on by the moderator. Participants were also asked to explore other meeting options available to them, and were assigned one "special" task each, such as verifying their places in line to speak at the meeting.

Usability test team. The three-person usability test team included:

- A moderator who ran the meetings and observed by listening only
- A stationary observer who also listened and who took detailed notes
- A roving observer who observed participant behavior and who took notes

The stationary observer and moderator were located together in a room separate from the test participants. The room was equipped with the meeting "control panel" (which was operated by a member of the client company's development team) and a speakerphone. The roving observer moved from room to room to observe participant behavior and body language through windows, and to assist participants who couldn't continue without some direction.

Data collection. The stationary observer heard all meeting participants and took detailed notes of their task performance. To collect data about unobservable behavior, we used self-administered data-collection forms (participant questionnaires) and group discussions.

Each participant was given three questionnaires with instructions to complete them at different points during the phone meeting, such as after getting in line to speak. Each questionnaire contained three easy-to-answer questions asking for either an opinion on how easy a task was to perform or a report on what tasks they had completed. In addition, participants were instructed to write—in space provided at the bottom of the questionnaires—their thoughts, feelings, and opinions during the course of the sessions. Each participant was also given a final six-item questionnaire to complete after hanging up the phone.

After each of the two telephone meetings, the roving observer brought participants together with the rest of the usability team for a 20-minute group discussion led by the stationary observer. These discussion groups were structured to collect information about participants' performance, feelings, and opinions.

Data that Resulted from these Collection Methods

Group Discussions. This source yielded data about usability, appropriateness of the tool, and other issues, as stated by participants themselves. The usability team was able to clarify participants' points of view as they were discussed and add "ammunition" to what we had observed both aurally and visually.

Stationary Observer Notes. This source also yielded data about usability, appropriateness of the tool, and other issues, but we consider this source more reliable because it was recorded by an impartial trained observer. However, the observer notes could not capture qualitative data gained from visual signs of participant frustration (or satisfaction) with the tool, which participants may or may not have disclosed in questionnaires or the group discussions.

Self-Administered Questionnaires. This source didn't yield as much quantifiable data as we had hoped. Participants were too focused on the session meetings to jot down their thoughts and feelings as the meeting progressed, or to complete questionnaires at the instructed milestones—they waited until after the meetings ended to complete all questionnaires. Because participants wrote their questionnaire answers "after-the-fact," they didn't accurately describe their behavior or task outcomes. For example, many participants reported they had successfully completed a task, when they actually had not (a fact we could deduce from having heard the sessions and from subsequent group discussions).

Roving Observer Notes. The roving observer activity did not yield as much qualitative data as we had hoped to collect. The distance between participants' rooms was further than ideal—the rover spent a lot of time simply walking from room to room. Also, it was challenging to gauge how much time to spend observing behavior if a participant needing prompting and there were still several participants yet to observe.

Conclusions

Group Discussions vs. Self-Administered Questionnaires. Group discussions are more workable than are self-administered participant questionnaires as a method for augmenting the audible data collected during multi-participant telephone conferencing studies. Motivating participants in different locations to answer questionnaires at specific milestone points during a test is extremely complex to manage. Even though participants eventually answered the questions, we could not be sure their responses accurately reflected their first impressions or initial behaviors. Therefore, we conclude that asking participants to write down their thoughts during a test session is not a viable alternative to "thinking aloud".

However, because participants can give their undivided attention to post-test questionnaires, these devices have the potential to yield satisfactory qualitative and quantitative data—if time can be allotted during the group discussion to clarify any ambiguous or contradictory answers and to expand on participants' responses. Moreover, a single note-taking sheet that includes a few thought-provoking questions might also be more manageable for participants to self-administer.

Stationary vs. Roving Observation. Stationary observation provides the continuity needed to realize the "big picture," whereas roving observation does not gain an accurate overall impression of participant reaction. Spot observation during a test is also challenging for the roving observer to manage, especially if the participant stations are far apart and the roving observer has to weigh "travel" time against observation time. However, our roving observer was able to capture participant behavior data that the stationary observer and moderator could not, and our study benefited from this kind of data.

References

Tahir, Marie Floyd and Kim, Katherine (Lotus); Spine, Tom (Sun) and Stanley, Karen (Digital Equipment Corp.); Wildman, Dan (Bellcore), "The More the Merrier? Methods for Multiple-User Testing", UPA'96 Conference Copper Mountain, CO.

Holleran, Patrick A.: "A methodological note on pitfalls in usability", *Behaviour & Information Technology*, 1991, Vol. 10, No. 5, 345-357.