

# MULTIPLE METHODS AND THE USABILITY OF INTERFACE PROTOTYPES: THE COMPLEMENTARITY OF LABORATORY OBSERVATION AND FOCUS GROUPS

Patricia Sullivan  
Purdue University

Recently, I used a focus group in a usability study of an interface prototype as a balance for a laboratory observation. The clients for this usability study wanted a sense of whether their interface was attractive to a range of users, whether the range of users understood the product, and whether the users could use the interface quickly; they also wanted user feedback on a list of potential features they could include in the next phase of development. Because of very limited resources available for the usability study, and because of the disparate questions the clients had, a focus group for some new users was used to supplement a laboratory observation and interview of other new users. This paper reports on what strengths and weaknesses these methods yielded as complementary approaches to testing the usability of interface prototypes.

The pairing of focus groups with laboratory observation is an instance of one of research designers' long-standing goals, namely, that data be as gathered as richly as possible. We can see designers reach for richness in most usability studies, even in so simple a design as adding an interview to an observation. In adding a method to the design, the study invokes the two aspects of multiple methods -- complementary and converging measures. The interview offers the user's holistic perspective on the use, which both adds richness to the observations and tests interpretations about user actions during observations. It provides a quality of extension and a quality of analysis.

This extension and testing of observation adds depth to research findings. Nowhere is this need more important than during the development of products -- a time when quality information about users needs to be delivered to developers quickly and thoughtfully. Studies conducted during development serve as litmus for developers, and thus steer aspects of the development. But those same studies are conducted under tremendous time pressure, making multiple methods a tempting sacrifice.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.

This discussion aims to support those usability researchers who are struggling to use multiple methods in those pressure cooker studies of prototypes.

## Part 1: On Multiple Methods

### COMPARING COMPLEMENTARY METHODS AND CONVERGING MEASURES

Multiple methods, as I'm using the term, refers to efforts in study design to assemble evidence from more than one source of information. At times discussed under the aegis of complementary methods, or converging measures, or triangulation, the deployment of multiple methods is widely accepted in social science, which points to widespread agreement that an interpretation is stronger (more convincing, more rigorous, or more valid) when it is built out of a variety of types of evidence. But "complementary" and "converging", the two key components of multiple methods, refer to two opposing activities -- the complementary urge one that extends the net of data gathering to include whatever seems appropriate and the converging urge one that sets up multiple measures of phenomena in order to test various observations. "Complementary" is a gregarious and additive move, while "converging" is an analytic one. Still, they are coupled in many discussions because they are the activities of multiple measurement.

Complementary methods, converging measures, and triangulation are embraced by most social science methodologists (Eisner and Peshkin, 1990; Van Maanen, 1988; Yin, 1984; Lauer and Asher, 1988; Doheny-Farina and Odell, 1988; Fowler, 1984; Kirk and Miller, 1986). One reason lies with the featuring of multiple observation. Methodologists tend to distrust every measure, acknowledging all measures are approximations. This leads them to distrust an interpretation that is built on one measurement. If a usability researcher, for example, conducts a study where the users do not read below the first inch on each page, and concludes that all important information needs to be on the first inch, methodologists will question the interpretation. But, if a number of measurements (taken at different times, by different

researchers, using different materials, using different subjects, etc.) agree (or replicate), the methodologists become more satisfied with the interpretation. If a variety of measures show similar results from their differing viewpoints, methodologists are even more satisfied with the evidence for an interpretation. For example, if a field study that looked at documentation found notes and underlinings only within the first inch, or if interviews found people could answer questions that were answered in the first inch more frequently than the bottom inch, etc., the finding of the original lab study would be strengthened. Multiple methods, then, are a sensible research design goal, and they are extremely important in areas where the work done is in large part qualitative or where it depends on the study of few subjects.

In usability research, however, there are no clear guidelines for researchers who are trying to assemble a design that uses multiple methods sensibly. Case study research has Yin (1984), ethnography has Denzin (1989), experimental research has such clear guideline that the ground is covered in most methods texts. Usability research of products under development, because it is often qualitative (Campbell, Mack, and Roemer, 1989), finds its work more situated in the needs of particular projects and less boxed into particular methodologies. This coin has two sides: usability researchers can design innovative research methods, but they have few of the traditional assurances afforded to the methods developed carefully inside various fields. A multiple methods approach is helpful here. It gives designers a strategy that makes sense to researchers in other areas at the same time as it allows them to design studies that answer the questions pertinent to the project.

### **STRANGE BEDFELLOWS? LABORATORY OBSERVATION AND FOCUS GROUPS AS PAIRED METHODS**

Most classifications of usability methods assign laboratory observation and focus groups to different types of research questions (Dieli 1986; Sullivan 1989a and b), the former gathering data about users' processes and the later gathering group opinions and ranges of opinion. Lab-based observation, used generally to refer to the family of methods employing a laboratory setting, is probably the most common means of gathering usability data on computer products. Sophisticated labs can gather information about keystrokes, can audio and videotape users, can allow observers to watch unseen and code on the fly, and can even do some computer analysis of tapes. These settings are ideal for gathering process-level information about how various users can go about using products. They are less ideal for gathering actual use information, learning over time information, social influence information, and generalized responses.

The interview is often used to complement laboratory observations of a user's process. In it, the researcher can

ask the users what they found harder and easier, what they remember, or what they preferred, and then check their post hoc responses against their actions in the session. The interview data helps confirm some findings, shed light on some confusing spots, and, yes, confuse some findings. "What they did" can be matched against "what they thought they did" to answer more global questions about understanding and satisfaction. Laboratory observation and interviews work together so naturally that they are almost standard components of a usability study run in a lab.

While the focus group is a group interview, focus groups and laboratory observation are not so obvious partners. For one thing, the focus group gathers a group of people together at one time while the laboratory study watches one or two people at a time. A few studies have linked them in some way (Denning, 1991), but situations where they are advantageous are more specialized.

Focus group research has long been a respected method in marketing research (Lazarsfeld, 1972; Merton, 1956). Its hallmark is its "explicit use of the group interaction to produce data and insights that would be less accessible without the interaction" (12, Morgan). Asking a diverse group to give opinions of real or potential products and services quickly clarifies any disagreements among representatives of target markets for products. The session also provides data that most people can analyze (Zemke and Kramlinger, 1982). A video of the session and an interpretive parsing of the data by opinions or by questions give the marketers a "feel" for the points of conflict. More recently, focus groups have begun to move into social science, communication, and now usability research.

Morgan claims that "What focus groups do best is produce an opportunity to collect data from groups discussing topics of interest to the researcher" (21). This means that they are informal, but somewhat controlled by the questions the researcher poses (there are always tangents and the effect of social discussion that shape it as well). When they deal with questions that people can discuss in public they give good feedback about why people hold particular opinions. They add a depth to interview research that comes out of that group stimulation.

Thus, laboratory observation and focus groups are strange bedfellows in situations where the usability questions focus exclusively on HOW a user acts, or HOW SUCCESSFULLY a user acts, or on WHY a user acts. But, when clients' questions link up the HOW and WHY, the two methods become more comfortable partners.

### **MULTIPLE METHODS, USABILITY, AND PRODUCT DEVELOPMENT**

The development of interfaces can profit from a multiple method approach to usability, particularly in a period where new media are rapidly being integrated into new and

existing interfaces. A complex task is becoming more complex, calling more and more for timely and effective information about usability.

Interface designers have long tried to gain perspective on the complexity of their task by developing guidelines--some developed from psychology (Gardiner & Christie, 1987), some from operating system compatibility, and some from collective wisdom (Gould & Lewis, 1985). They have valued the user, as attempts to integrate user information into development show (e.g., Campbell, et al., 1989; Carroll & Rosson, 1985), but the task is huge.

Nielsen (1990), when commenting on the challenges of designing multimedia interfaces, puts it this way:

"[Actually] modern interaction techniques only increase the need for the designer to pay attention to the usability principles since these techniques increase the degree of freedom in the interface design by an order of magnitude. There are only so many ways to ruin a design with 12 function keys and 24 lines of 80 characters, but a 19 inch bit-mapped display with stereophonic sound can be an abyss of confusion for the user (117)."

Nielsen's order of magnitude example does not mention the variety of uses for these interfaces, or the range of users and their jobs, or the social constraints (such as crowded offices) that may also affect which combinations of media are effective for what tasks. A wealth of office information system research and some of the findings of the researchers who favor a design process that integrates usability suggest that these other factors are important and complex.

The addition of new media creates tensions in the work of the interface designer, as it becomes clearer that an experimental approach to usability alone cannot generate information quickly enough about the multiplicity of factors at work when users encounter multimedia interfaces. Theory about how the media work together for the users is needed to organize the multitude of choices, as is quickly gathered and flexible feedback during development.

## **Part 2: A Case Study Using the Multiple Methods of Laboratory Observation and a Focus Group**

This case presents a somewhat unconventional deployment of a focus group; it is offered as an example that can be used to discuss the feasibility of pairing these methods rather than as an exemplar. The party line in marketing research is that focus groups are used to probe the understanding of people familiar with a product, process, service, activity (Zemke and Kramlinger, 1982). Though some use it more broadly -- to get initial reactions to new

ideas, products, etc. and to conduct qualitative research (Morgan, 1988) -- the obvious way to link the focus group to the laboratory observation is to use it afterward, to let users become oriented to the system then to run a focus group later. This is a particularly good idea in situations where a period of using is important. Users can get started in the lab (where they are observed), go home/work to use the system for a time, and then meet in a focus group to discuss learning, preferences, how it fit into various use environments, and so on (Denning's study addresses this situation). But, with prototypes under development, not all features are operational in the ways that the system will later work. Thus, extended experience with the prototype will not simulate actual use of the ultimate system, nor will users learn that much more about the system. Work with prototypes encourages a more unconventional use of focus groups, one where the users know less about the product when they enter the session.

### **DESIGN OF STUDY**

The Development Team for the product in this study wanted to know more about how intuitive the interface is and about what future features users want in their database under development. A study design problem was: how to get converging evidence about intuitiveness without alerting subjects. After consultation with the client, a two-pronged attack (observation/interview of individual users and a focus group) was adopted to explore the following questions:

- 1) how easy is it for users (those projected as casual users in particular) to search a test database accompanying the interface? (Baseline information)
- 2) what do users think the language and icons used mean?
- 3) how well can users navigate potentially difficult features of the interface?
- 4) how well do users like the interface?
- 5) what features do users want developed in the future?

The study's focal questions asked users to orient themselves in different ways: the "ease-of-use" questions required hands-on, local, experiential knowledge, while the "features of the future" questions required global thinking and vision (See Figure 1 for breakdown of questions). Thus, the study used a two-fold approach:

- Six users (volunteers from the company) were observed using the test database and interviewed about their preferences after gaining some "hands-on" experience, and
- Five different users (also volunteers from the company) attended a focus group where they the interface was demonstrated and they were asked about their opinions on development issues.

Focus	Actions	Lab Observation & Interview	Focus Group
1/ Baseline	check searching skill ask opinion of success ask opinion of ease	√ √ √	√ √ √
2/ Language	ask what words mean ask what icons mean compare menus (A & B)	√ √ √	√ √ √
3/ Working Features	moving around enter info/start search select/display full record select/execute OK CANCEL move forward/back through list SELECT ALL manipulating data FILTER browse list expand a query invoke history opinions	√ √ √ √ √ √ √ √ √ √ √ √ √ √ √	√
4/ Preferences	survey data interview data	√ √	√
5/ Ideal System	opinions about ideal functions ideal services how this system rates strengths weaknesses opinions about future development	√ √ √ √ √ √	√ √ √ √ √ √

**Table 1: Information the Multiple Methods Yield for the Research Questions**

In order to generate converging evidence, overlap was worked into the sessions as possible. For example, the focus group started with a demonstration of the searches that the single session users were asked to perform. Of course, the two sessions could not be isomorphic because the demonstration leader of the focus group handled the keyboard for five people. Further, focus group participants were focused on discussing, the single subjects on acting. Thus, much of the time the two groups were generating complementary data.

**Data Collection: Laboratory Observation Searches.** In the single sessions users were in a word processor and were asked to start the prototype and perform several searches. The original starting point was inside another application because it is likely that they will use the database in that way.

A sample query was:

You are interested in what marketing research is available on multimedia applications. You are particularly interested in risks involved in marketing these products.

*Language and Icons.* Users examined the icons using a color screen dump and told what they thought was behind each icon. Then they tried to place seven commands and options into two potential menu bars and select the better menu bar. Finally, they looked at all the language in the menu bar, new search box, and main interface. They told what they thought various words and options meant.

*Features.* Users were then asked to explore how features work. They would be asked to perform various tasks -- some navigational in nature, dealing more with handling the resulting data.

*Constraints.* Laboratory conditions, learning, and the number of participants were constraints in this study. A long range goal of the product is to improve use information. The laboratory test used a test database and could determine how users respond to certain features, but could not project the problems a larger database would introduce.

Learning was another constraint. We would have liked to query users about language before they had a chance to use the system, for example. But, a strong measure of intuitiveness is ability to immediately search. Thus, the searches had to go first, as we could not risk the language discussion teaching users how to search. By the time we got to exploring features, then, it is reasonable to assume that some people learned how to answer the questions though their experience earlier in the session. Because the searches were short, the potential of a learning confound is modest. Finally, Six users--three casual and three more advanced-- can point to trends and to problems but cannot conclusively demonstrate. They more easily point out negatives than they confirm positive features.

#### **Data Collection: Focus Group**

The two hour focus group session began with introductions and questions about the participants current database use habits. Then there was a demonstration of the prototype and question-answering, followed by a posing of questions related to the system specifically. More general questions, related to the future development of the product, came late in the session.

*Demonstrating the Process.* Early in the session, the demonstrator searched the test database for the same queries that single users searched for above while the leader narrated what was happening and why. This gave the participants an introduction to how the prototype works that was necessary for them to answer the first group of questions about the system.

This round of questions focused on responses to the database's operation, icons, language, saving work, setting up personal and group profiles, displaying information, and browsing. For example:

- What queries do you think you'd use frequently?
- What would you do if you're getting no hits?
- Do you think you'll use the system more to look for general information or to find a specific item?

- When information is displayed on the screen, what information would you like displayed? How would you like the information to be formatted?

*Projecting the Future.* Before the questions on future developments, how the interface would come up on screens was demonstrated, as were icons that represented further database developments. This was used as a basis for discussing their opinions of future developments and their priorities.

This round of questions explored responses to the future developments, the icons, ideal systems, and training. For example:

- What would you like added to this system in the future?
- How would you make it seem accessible, "next door"?
- What kind of training would you/most people need to use this application?
- What would make the interface more intuitive, easy-to-use, and convenient to use?

*Constraints.* The focus group got its experience of the prototype through demonstration and discussion rather than through hands-on experience. They never got the opportunity to use the keyboard and thus could not give feedback on micro-use issues (e.g., did you have any trouble using NEXT and PREVIOUS?). Because they learned through demonstration, they tended to think the system was easier to use than users who tried to actually use it, in much the same way we watched demonstrations of desktop publishing products and thought they would take half an hour to learn. This led them to discount the amount of documentation and training that would be necessary for the successful introduction of the working system.

Finally, because they watched a screen and listened to a script that directed their attention to various features, they actually noticed more of prototype's interface than the single-users did. This led them to notice more elements of the screen than the laboratory users did, but not in a natural, intuitive, or untutored way. They saw more as we wished them to see. This constraint was a potential weakness for this type of focus group -- one with the correct users but with none having experience with the product at hand.

#### **STRENGTHS AND WEAKNESSES OF THIS DESIGN**

The design had a kind of situational strength that would make me repeat it again: it provided information on the client's key questions in a timely fashion and in a way that allowed a limited number of users to give several types of feedback on the prototype. Because the client had many exploratory

questions to be answered, the focus group of potential users could probe those questions in a shepherded manner. Because the client needed baseline information about the usability of the interface as well, the laboratory observation could point to any clear stumbling blocks for typical users.

In other situations, this design would not remain the same. From the standpoint of converging measures, the design would be "stronger" (design-wise), for example, if it used the same subjects for the laboratory observation as it did for the focus group. The repetition of the subjects, would make the converging evidence from the two methods less noisy, at least theoretically. As it is, the converging measures (see table 1 where both method share a check) are confounded by differing users and by differing types of experience. This can be demonstrated through a general finding. One group thought the interface much easier to use than the other. Which? The focus group. But which had more experience to use in judging its ease-of-use? The laboratory group. They did not have comparable experience with the prototype. From the standpoint of complementary measures, though, the variety in experience may have given richer feedback to the group. In situations where varied feedback is the order of the day, the current design is more desirable.

Several problems arose during the study that deserve some thought.

*It was difficult to coordinate/approximate two radically different experiences of the system into one interpretation of the results.*

*It was more difficult to write out that interpretation quickly.*

Although I tried to acknowledge the differences in the types of data provided by the laboratory observation and the focus group, I found them hard to write about in one report. It was easier to quote numbers of users who could successfully locate some information, or times, or types of process problems than it was to classify the range of responses generated by the focus group. It was more comfortable following a user work through a process than it was to follow a group ebb and swell through a discussion.

As a way to force myself to take both methods seriously, I started with the focus group, generated some hypotheses, and then tested them against the single subjects. That produced some modest coordination. But the focus group was better at visionary discussion than the laboratory users, so cross-checking them against one another did not do either method justice.

As a second way to coordinate them, I returned to the questions of the study and sorted the data into five piles -- those related to each client question. This worked somewhat better, because each method had at least one question to itself, and thus was clearly included in the report. But, it was difficult to build an interpretation that used two distinctive

types of data, and it was more difficult to fit it into a report-like argument. I had difficulty quickly relating appropriate contextual information as I moved from observation to focus group, or vice versa.

Longer qualitative studies, often are months or years in interpretation and writing and have the "luxury" (it's pain really) to synthesize data into narratives that are rich with nuance. To the extent that the interpretation of focus groups resembles those slowly developing narratives of anthropology and sociology, the more difficult they become to write quickly. This is a potential deterrent to their use, as researchers may think that the extra time or effort necessary to yoke focus group findings to other observation is not worthwhile.

A real question for focus groups in the study of the usability of interface prototypes is: can they be interpreted quickly enough? Since I was able to write the analysis in a week, I offer a tentative "yes." But I also felt that I was rushing the interpretation of the data. We need to think hard about the reporting of studies that contain findings from multiple methods that are not usually employed together.

## CONCLUSION

Laboratory observation and focus groups are usefully complementary for usability research that addresses user responses to an interface under development. They provide diverse input on key interface issues of look and feel, accessibility, and intuitiveness. Though not the "ideal" pairing of methods for all situations, as responses to a combination of HOW and WHY questions that explore the possible features of an emerging interface, they can add interesting complementary information.

The need for complementarity is great in research methods that address the usability of computer interfaces. Since the focus in such studies is on speed and richness of feedback on issues important to the interface (such as accessibility, intuitiveness, look and feel, and navigation), multiple methods that complement each other and can be stitched together quickly are desirable to usability researchers. Harnessing multiple methods to this task is not a simple and easy answer (as this extended example shows), but it is one that holds promise for increasing the quality of usability feedback on products under development.

## REFERENCES

- Campbell, R. L., Mack, R. L., & Roemer, J. M. (1989). Extending the scope of field research in HCI. *SIGCHI Bulletin*, 20, 4, 30-32.
- Carroll, J. M., & Rosson, M. B. (1985). Usability specification as a tool in iterative development. In *Advances in Human-Computer Interaction*, Vol. 1, H. R. Hartson (ed.), Ablex, Norwood, NJ, pp. 1-28.
- Denning, S. (1991). *The advantages of focus groups in usability research*. Microsoft Corporation, unpublished mss.
- Denzin, N. K. (1989). *Interpretive Interactionism*. Newbury Park, CA: Sage.
- Dieli, M. (1988, Oct.). *Integrating usability evaluation into the computer documentation development cycle*. Paper presented at the ACM SIGDOC Conference, Ann Arbor.
- Dieli, M. (1986). *Designing successful documentation: An investigation of document evaluation methods*. Dissertation, Carnegie-Mellon University.
- Doheny-Farina, S., and Odell, L. (1985). Ethnographic research on writing: Assumptions and methodology (pp. 503-535). L. Odell and D. Goswami (eds.). *Writing in nonacademic settings*. New York: Guilford.
- Eisner, E. W., & Peshkin, A. (1990). *Qualitative inquiry in education: The continuing debate*. New York: Teachers College Press.
- Fetterman, D. M. (1989). *Ethnography: Step-by-step*. Newbury Park, CA: Sage.
- Fielding, N. G., & Fielding, J. L. (1986). *Linking data*. Beverly Hills: Sage.
- Fowler, F. J., Jr. (1984). *Survey research methods*. Beverly Hills: Sage.
- Gardiner, M. M., & Christie, B. (Eds.). (1987). *Applying cognitive psychology to user-interface design*. New York: John Wiley & Sons.
- Goldman, A. E., & McDonald, S. S. (1987). *The group in-depth interview*. Englewood Cliffs, NJ: Prentice-Hall, 1987.
- Gould, J. D., & Lewis, C. (1985). Designing for usability: Key principles and what designers think. *Communications of the ACM*, 28, no. 3 (1985), 300-311.
- Hirschheim, R. A. (1986). Understanding the office: A social-analytic perspective. *ACM Transactions on Office Information Systems*, 4, 4, 331-344.
- Kirk, J., & Miller, M. L. (1986). *Reliability and validity in qualitative research*. Newbury Park, CA: Sage.
- Krueger, R. A. (1988). *Focus groups: A practical guide for applied research*. Newbury Park, CA: Sage.
- Lauer, J. M., & Asher, J. W. (1988). *Composition research: Empirical designs*. New York: Oxford University Press.
- Lazarsfeld, P. F. (1972). *Qualitative analysis: Historical and critical essays*. Boston: Allyn and Bacon.
- Merton, R. K., Fiske, M., & Kendall, P. L. (1956). *The focused interview: A manual of problems and procedures*. Glencoe, IL: The Free Press.
- Morgan, D. L. (1988). *Focus groups as qualitative research*. Newbury Park, CA: Sage.
- Nielsen, J. (1990). Traditional dialogue design applied to modern user interfaces. *Communications of the ACM*, 31, 5, 552-561.
- Sullivan, P. (1989b). Beyond a narrow conception of usability testing. *IEEE Transactions on Professional Communication*, 32, 4, 256-264.
- Sullivan, P. (1989a). Usability in the computer industry: What contribution can longitudinal field studies make?. *IPCC Conference Record*, pp. 12-16.
- Van Maanan, J. (1988). *Tales of the field: On writing ethnography*. Chicago: University of Chicago Press.
- Yin, R. K. (1984). *Case study research: Design and methods*. Beverly Hills, CA: Sage.
- Zemke, R. & Kramlinger, T. (1982). *Figuring things out*. Reading, MA: Addison Wesley.