Reprint of paper published
in IPCC 2007 Proceedings.
Seattle, WA: Institute of Electrical
and Electronics Engineers, Inc.,
Professional Communication Society (2007)

Field Usability Testing: Method, Not Compromise

Stephanie Rosenbaum
Tec-Ed, Inc.
stephanie@teced.com

Laurie Kantner
Tec-Ed, Inc.
laurie@teced.com

Abstract

This paper discusses a user research method the authors have refined over several years: field usability testing. Field usability testing combines techniques from traditional laboratory usability testing and condensed contextual inquiry, itself an adaptation of traditional contextual inquiry methods. The authors describe two approaches or models of field usability testing: ethnographic and structured. Three case histories illustrate the method, giving examples of the ethnographic model and the structured model. Keywords: user research, field research, usability testing, ethnography

Introduction

As the practice of observing our target audiences—user research—gains maturity, its teachers and practitioners have formalized and documented its methodology. A wide variety of literature is available about usability testing 2,8), contextual inquiry (4,7), and ethnographic interviews (12). Information-gathering techniques employed within these methods, such as think-aloud protocol (1) and card-sorting (3), have also been addressed in our professional discourse.

Usability testing has its roots in a controlled ("laboratory") environment, where uncertainties of measurement are minimized because all participants use the same computer equipment (or other product) and perform tasks with the same set of data. Thus it's valuable not only for problem identification, but also for competitive evaluations and collecting quantitative data about a product's usability.

In contrast, field research methods (10) such as contextual inquiry and ethnographic interviews involve observing people in their everyday situations—homes, workplaces, and public places—to learn their normal or natural behavior. We develop an in-depth understanding of users by watching and interviewing them performing their real activities in their natural environments.

This paper discusses an adapted method that our consulting firm has refined and used for several years: *field usability testing*. Field usability testing (9) combines some of the techniques of traditional laboratory usability testing and some techniques from *condensed contextual inquiry* (5, 11), itself an adaptation of traditional contextual inquiry methods.

Summary of the "Parent" Methods

The following sections briefly review the two methods from which field usability testing derives: laboratory testing and condensed contextual inquiry.

Laboratory Usability Testing

In usability testing, people whose characteristics (or "profiles") match those of the target audience perform a series of typical tasks. Each participant, usually working one at a time, performs the same tasks under controlled conditions, facilitated by a test administrator. The research team selects the tasks to be performed, based on criteria such as frequency of use or how critical the tasks are to overall successful use of the product.

The basic premise of usability testing is that we can gain unique insights about users' needs and preferences by observing their behavior as they perform typical activities. Laboratory usability testing emulates the expected real-world context of use in a controlled environment. The two most important parts of this emulation are realistic scenarios (situations of use combined with user tasks) and representative users (test participants with the same characteristics as the target audience).

Depending on the stage in the product development process, laboratory usability testing can be exploratory or performance-based. In exploratory testing, the goal is usually problem identification to inform product design or redesign. In performance testing, researchers want usability metrics for benchmarking, and thus collect quantitative data—for example, the number, type, and severity of errors users make.

The design of a laboratory usability test summarizes the issues of concern, the tasks to be observed, the questions to ask, and the criteria for screening the people who participate. The researcher creates and follows a

detailed session script so that all participants receive the same instructions and error remediation, while performing the agreed-upon tasks using the same data.

Laboratory usability testing also has a strong psychological benefit for observers and helps build credibility for user-centered design within an organization. The experience of watching people having problems in a test session is more convincing to product designers and developers than simply hearing the recommendations of user experience practitioners.

Condensed Contextual Inquiry

Contextual inquiry is a qualitative data-gathering and data-analysis methodology adapted from the fields of psychology, anthropology, and sociology. It consists of observing and talking with people in their workplaces and homes as they do normal activities. Key characteristics of contextual inquiry include:

- Users become partners with the researchers in the inquiry; an ongoing dialog enhances data collection
- The inquiry is based on a set of general concerns to guide observation, not on a list of specific questions to ask
- The result is concrete data based on users' expertise in their own activities

Classic contextual inquiry requires hours of time with each user—up to a full day each. Because our consulting firm wanted to be able to gain the benefits of contextual inquiry even when time is short on commercial projects, we developed the condensed contextual inquiry. It identifies a more constrained set of concerns to investigate, allowing researchers to focus on a few critical issues during sessions with users. The condensed method takes 90 minutes to two hours with each participant, and retains the strengths of contextual inquiry, by exploring:

- People's use of products within the restrictions of their actual work
- When and how companion software and artifacts such as notebooks, yellow stickies, and forms are used to complement the product
- Details about tasks while they occur, to avoid misunderstandings about what users did and why

Contextual inquiry often involves fewer participants than other methods, so we must be especially careful to choose them carefully. Most contextual inquiry projects include at least three participants per profile, to help minimize the effects of individual differences.

The contextual inquiry session protocol includes an outline for the facilitator, the high-level concerns, and sample probing or follow-on questions for each concern. The protocol is high level because the structure of each session is unique, dictated by the user and the user's tasks (the tasks performed are a combination of those the user selects and those the facilitator suggests). To prepare for the sessions, the researchers work closely with the product and hypothesize situations they might observe.

Because contextual inquiries involve conversation as well as observation, they require a high degree of skill from researchers, who must ask appropriate questions without interrupting the participants' workflow or influencing their responses. Our consulting firm conducts condensed contextual inquiries with a team of two researchers, one of whom is the lead facilitator and the other who takes notes and operates recording equipment.

After completing the contextual inquiry sessions, the researchers create a text database of notes (including participant quotes) that reflect their observations. This method assists the initial data analysis and makes it easier to "mine" the data to address follow-up. Summarizing the sessions and compiling the qualitative data are more time-consuming than tabulating data from more structured research.

The Adaptation: Field Usability Testing

Field usability testing adapts the well-known methodology of laboratory testing by conducting the sessions in the participants' own environments, on their own computers (or other equipment). Two motivations led to the development of this method:

- We wanted to learn the kinds of insights we gain from usability testing, but in a natural environment, not an artificial situation
- Some target users are reluctant or unable to leave their normal environments and come to a usability lab

In field usability testing, we design tasks that address the participants' own goals, where task objects include the users' files, bookmarks, or databases. These adaptations give us qualitative data about the target audience that we can't collect in the lab. Especially in home-based research or in small business settings, participants' choices of computer, software, and Internet service noticeably affect their experience and behavior with products and services.

Field usability testing is best suited to exploratory objectives, where we want to learn what problems users encounter as they follow their own work processes. Owing to the variations in computer equipment and user tasks, it is less suitable for performance measurements, especially "time on task" metrics such as comparing which version of a form is faster to complete.

Broadly, we use two approaches or models to field usability testing. The *ethnographic model* more closely resembles contextual inquiry; its goal is to gain insight into how people use a product, even if their behavior varies from its intended use.

In the ethnographic model, we observe participants working with their own data on installed software or released websites. The difference between this model and

condensed contextual inquiry is that we supply the task objectives or high-level tasks for the session, rather than observing people in their actual daily activities. We give each participant the same objectives, although we expect (and usually find) that they make quite different choices in how they carry out the high-level tasks.

The *structured model* more strongly resembles a traditional usability test. The research takes place in the field for one of two reasons:

- The nature of the product and data prevent a realistic emulation in the usability lab
- The participants can't or won't come to the lab

In the structured model, we sometimes observe use of a released product or website, while other times we bring prototypes or storyboards of new designs with us to the field setting. Some study designs include both kinds of activities. For both the ethnographic and structured models, even when we run through the session script with the product or prototype in the lab, the first session in the field also serves as our pilot test and may result in changes to the script.

The key to successful field usability testing methodology is identifying for each study what kinds of collected data can be compared among participants, and which must be used descriptively in creating individual scenarios of use. The creation of data tables or databases for organizing qualitative data (6) is especially valuable for field usability testing.

Case Histories of Field Usability Testing

Three case histories of field usability testing illustrate and explain the differences in methodology needed to make a field usability test successful.

Biomedical Engineering Library Product

A publisher of engineering journals wanted to learn how effectively people could use a new biomedical engineering library product to locate reference material, specifically journal articles and papers tagged as biomedical subject matter. Usability testing was especially important because this product had a new interface and was intended to be a template for future products. Our consulting firm conducted field usability testing at customer sites so the users would have reminders of real information they wanted to look up.

The study focused on two categories of users, in both corporate and academic settings:

- Researchers in biomedical engineering disciplines who frequently consult the literature of their field
- Librarians who help biomedical engineering researchers obtain literature necessary to their research

The field usability test explored these questions:

- What are users' first impressions of the product?
- How well do users understand what the product enables them to do? How well do users understand that it provides full text articles, papers, and standards, not simply citations?
- How easily and successfully do users perform typical tasks, such as finding pertinent articles?
- How well do users understand that they are searching the full content set of the collection, not simply keywords or article titles?
- What terms or concepts in the user interface are confusing?
- What do users want to do that the product does not support?

We held a total of ten test sessions of 60 minutes each, in three locations: a pharmaceutical firm, a commercial biochemical research laboratory, and a university conducting biomedical research. The participants were six librarians or information scientists and four researchers, including people with job titles of electronic resource analyst, reference librarian, coordinator for engineering collections, senior scientist, and team lead for genomic department, as well as two Ph.D. students in biomedical engineering.

All participants performed six tasks, interspersed with Likert-scale questionnaires:

1. Explore the product home page
2. Find an article or paper of their own interest
3. Find an article or paper by a specific author
4. Find an article or paper on a specific topic, published before a certain year
5. Browse for an article or paper about a specific topic
6. Search for an article in a specific journal

We collected both quantitative and qualitative data; we measured:

- The number of problems and "wrong turns" participants encountered while performing test tasks.
- The number and level of administrator interventions. We noted four levels of administrator prompting, used if and when participants needed assistance.

None	Participant responded correctly without prompting.
Try again	Participant responded correctly when first told just to try another way.
Generic instructions	Participant responded correctly to a "generic" indication of the right approach.
Specific instructions	Participant responded correctly only when the right choice was named explicitly.

- Participants' opinions about the product, based on answers to the Likert-scale questionnaires.

We also recorded:

- Participant behavior the researchers observed during task performance.
- Comments participants made during the test sessions as they "thought out loud" while performing activities.

Overall, study participants responded positively to the capability offered by the new product and described their experience at the end of the session as "Good." The 36 findings in our results report ranged from "keepers" (findings which promote ease of use) to a few "show-stoppers" that we recommended our client correct before the first release of the product.

This study exemplifies the structured model of field usability testing. Prior to the sessions, we collected background information about the participants' job responsibilities and the biomedical resources participants used at work. The participants were in their offices during this telephone interview and thus could consult their own computers and bookshelves.

During the test sessions, all participants used the same product (the beta-test version that was available to their institution). The protocol we followed and use of the beta-test product could have taken place in the laboratory. However, we wanted participants to have access to their current work and bookmarks to help them recall things they actually wanted to look up. Also, the research institutions might not have agreed to participate in the study if their staffs had to leave the workplace.

Online Banking for Vision-Impaired Users

In another study, our consulting firm evaluated the accessibility of the online banking experience for users with vision impairments, by observing them in their homes as they performed basic tasks on the website of a major bank:

- Viewing account activity (viewing a statement, getting check images)
- Paying bills (setting up a new payee, paying a bill, and viewing bill payment history)
- Transferring funds

We employed field usability testing because it was important to observe these people in the context of their own computing environments, which included assistive tools such as screen magnifiers and screen-reading software (JAWS or Window-Eyes). The test participants were all vision-impaired (some were low-vision, and some were blind), and a few also had motor impairments such as tremor.

We wanted to observe each of these users following their normal procedures with online banking, considering their vision impairments and unique computing environments. However, usability testing of financial products almost always faces the challenge that participants are unwilling to use their own bank accounts or credit cards. Even for this ethnographic-model field usability test, we established a separate bank account for participants to manipulate.

For each task, the field usability test explored the following questions:

- How easily can users begin the task?
- How easily can users find the information they need to make decisions while performing the task?
- What barriers (if any) prevent users from completing the task, or from completing the task as efficiently as they would like?
- Where do users say they would "bail out"? Why?
- What improvements would make the website better support task initiation, performance, and completion for users with vision impairments?

The banking site already featured many accessibility features—for example, cascading style sheets for navigation and text tags for graphic elements. The goal of the study was to learn what problems vision-impaired users had when using the site with their own assistive devices and personal experience using these devices.

During the tasks, we were particularly interested in how the participants navigated around the screen, and how they handled non-text content such as check images. For example, did blind participants try to read the information in the check image with JAWS? Did participants with low vision try to adjust the check image in any way (such as zoom or rotate)?

Blind users employ a screen reader, most often JAWS, while low-vision users use either a screen reader or a screen magnifier such as Zoomtext. These different assistive technologies require different interaction strategies with websites. Thus although we observed a variety of strategies, the collected data took the form of "cases" or scenarios, rather than usability metrics.

The results of the field usability testing informed the bank's continuing design improvements to the website. We observed behavior trends in the participants, such as which features of the assistive technology software they used or didn't use, that will help the bank's design decisions for making the site more accessible to vision-impaired customers.

Iterative Research for Medical Publishing

In an ongoing program of field usability testing for a major medical publisher, researchers visited physicians, residents, medical librarians, and medical students in clinics, homes, and hospital settings. The two studies described next fit both key rationales for field usability testing. We cared about the environments in which participants used print and online medical references, and health care professionals are notoriously difficult to

recruit as research subjects if they must leave their own settings.

The first field usability test collected behavioral and perception data from ten people in the internal medicine field about the ease of use and usefulness of a new release of an online medical information resource. The goals of the field usability test were to learn:

- How easily and successfully participants found information that answered their internal medicine questions
- How easily participants could download information from the medical database to their PDA
- What tools were of value to participants
- How participants perceived the new medical resource in comparison to its previous online version, the printed book, and other online medical information sites for internal medicine physicians

The researchers conducted ten usability test sessions: five in the New York City area and five in Southeastern Michigan. Sessions were between 1.25 and 1.5 hours long. Six of the sessions took place at the participants' offices or at computers in hospitals or clinics. Three sessions took place at participants' homes, and one took place at Tec-Ed's usability lab in Ann Arbor (so in fact the study was 90% a field usability test).

The participants included four practicing clinicians, four medical residents, and two medical students. All used either an earlier version of the medical resource or the hard-copy edition of the information, as well as other online medical information sources; these requirements—plus the difficulty of recruiting high-demand physicians as research subjects—called for field usability testing methodology, with its greater convenience for participants.

Each session began with a discussion about the printed book or previous online edition, then progressed through two to four lookup tasks and a PDA download task. For the final activity, the participant filled out a questionnaire about the experience. The lookup tasks were of the participant's own choosing and thus were unique from session to session. This approach increased participants' interest in the task and its outcome, and exemplifies the ethnographic model of field usability testing.

Tec-Ed videotaped all ten sessions. In the New York area sessions, a representative from the medical publisher attended the sessions with the researcher, while two researchers conducted the Michigan sessions.

The initial field usability test identified many problems; participants were successful in only nine out of 36 total lookups. However, they were loyal users of the printed resource. As one participant said, "[it] will always find a place." To be useful for practicing clinicians, the online resource must help them find answers "between patients," which is not how the study participants perceived using it. These results led to the iterative program of field usability testing our consulting firm is still conducting.

The second field usability test collected behavioral and perception data from a different set of ten people in the internal medicine field about the ease of use and usefulness of a larger medical resource from the same publisher (the larger resource actually included the content from the smaller one). We evaluated the version of the product reflecting improvements recommended from the prior research.

We conducted ten field usability test sessions: six in the New York City area and four in Southeastern Michigan. The methodology was essentially the same as the first study. Each session began with a discussion about the participant's use of the content, then progressed through two to four lookup tasks of the participant's own choosing, followed by a PDA download task. Each participant filled out a questionnaire about the experience, and we videotaped all ten sessions.

The participants included two internal medicine physicians, two medical residents specializing in internal medicine, two medical school librarians, one nurse practitioner in internal medicine, one physician's assistant, and two medical students. Based on our experience from the first project, we tripled the honorarium for the physicians, and this more substantial sum did help soften their resistance to participating (perhaps because it reflected the respect due their profession, rather than the money itself).

In this field usability test, the participants fully succeeded in meeting 14 of the 31 search goals they expressed in their sessions with the medical resource. Participants partially succeeded in meeting an additional 5 of the 31 search goals, and failed to meet 12 of the 31 search goals.

Because this field usability test was the second one our consulting firm conducted on the product family, we were able to make more specific and targeted recommendations for improving the navigation in ways that would yield more successful searches. We noted the improvements in participants' experiences, as well as the benefits participants found in quality, timeliness, and trustworthiness of content. Further field usability testing of this product and related products for this medical publisher continues.

Conclusion: When to Choose Field Usability Testing

The case histories in this paper illustrate how field usability testing differs both from traditional usability testing and from other field research methods. A summary of these differences is in the following tables (where TUT is Traditional Usability Testing, CCI is Condensed Contextual Inquiry, and FUT is Field Usability Testing).

Task context

Traditional UT	Condensed CI	Field UT
All participants use same equipment for tasks; data common to all is either supplied or can be accessed online	Methodology requires participant's own equipment and data	Usually employs participant's own equipment and/or data, not necessarily both

Task scenarios

Traditional UT	Condensed CI	Field UT
All participants perform the same tasks, designed by researchers to be as realistic as practical	Participants perform their own real tasks (within a predefined area of focus), sometimes including ones suggested by the facilitator	Researchers design high-level tasks or task objectives that participants implement in their own ways

Facilitator interaction

Traditional UT	Condensed CI	Field UT
Virtually all interactions with the participants are scripted in advance, including prompting and error remediation	Session protocol includes outline and probing questions, but dialogue with participant requires ad hoc interactions	Depending on the degree of structure of the FUT, the facilitator interactions may be highly scripted or more naturalistic

Data collected

Traditional UT	Condensed CI	Field UT
Problem identification to inform design, quantitative metrics such as number, type, and severity of errors, participant quotes	Descriptive anecdotal stories about participants' work processes, participant quotes	Problem identification to inform design, some measures that can be compared among participants (such as types of errors and number of task failures), participant quotes

Resource requirements

Traditional UT	Condensed CI	Field UT
Because participants come to the same location, and the session protocol rarely changes after pilot testing, schedule and budget are easier to control	Obtaining permissions to visit homes or businesses, plus travel and set-up time, mean that about 9 CCI sessions require the same resources as 12 TUT sessions; schedule is typically 25% longer	FUT can consume more resources than either TUT or CCI, because it requires the scripting granularity of TUT and the logistics of CCI

Skill requirements

Traditional UT	Condensed CI	Field UT
Most usability practitioners begin their careers with TUT, because the detailed scripting enables practice and rehearsal before the sessions to achieve quality results, as well as minimizing facilitator errors during sessions	Requires the most facilitation skill, because researcher must achieve rapport and conversational flow with participants despite interruptions in the participant's environment and without influencing responses	Requires project management and logistics skills; facilitation skill requirements depend on the degree of structure of the FUT (can be as demanding as CCI)

Every user research program should include some field studies and some laboratory testing. How do we decide which method to use, and when to use it?

As described in an earlier article (9), in general, when we need to collect metrics about a product's usability, measure how it compares to the competition, or make a go/no-go decision on a particular feature, we usually conduct traditional usability testing. When the primary goal is to understand users or customers better—to learn what they really do with our products, or to explore which new features to add—we suggest contextual inquiry or other field methods such as ethnographic interviewing.

We choose field usability testing partly because of the users' constraints and partly because of our goals. If key audience groups are unable or unwilling to come to the usability laboratory, then we use field methods to collect their data, even when such methods might not otherwise be our first choice.

If we want to collect structured data, and product or system usage is not sufficiently realistic in the laboratory setting, some (but not all) metrics can be collected with field usability testing. If we want both structured data and insight into users' actual context of use, then field usability testing is an excellent choice, and worth its attendant skill and resource demands.

References

[1] Boren, T., & Ramey, J. Thinking aloud: reconciling theory and practice. *IEEE Transactions on Professional Communication, 43*(3), 261-278, 2000.

[2] Dumas, J. & Redish, J. *A practical guide to usability testing* (revised ed.). Fishponds, Bristol, UK: Intellect LTD, 1999.

[3] Gaffney, G. "What is Card Sorting?" *Information & Design*, 2000.

[4] Holtzblatt, K. & Jones, S. (1993). Contextual inquiry: a participatory technique for system design. In D. Schuler & A. Namioka (Eds.), *Participatory Design: Principles and Practices* (pp. 177–210). New Jersey: Lawrence Erlbaum.]

[5] Kantner, L. & Keirnan, T. *Field research in commercial product development*. Proceedings from UPA 2003: Ubiquitous Usability. Scottsdale, AZ, USA, 2003, June.

[6] Kantner, L., Sova, D., & Anschuetz, L. *Organizing qualitative data from lab and field: challenges and methods.* Proceedings from UPA 2005: Bridging Cultures, Montreal, Quebec, Canada, 2005, June.

[7] Raven, M.E. & Flanders, A. Using contextual inquiry to learn about your audience. *ACM SIGDOC Journal of Computer Documentation, 20*(1), 1996.

[8] Rosenbaum, S. *Not just a hammer: when and how to employ multiple methods in usability programs.* Proceedings from the UPA 2000 Conference. Asheville, NC, USA, 2000.

[9] Rosenbaum, S. *Stalking the User: Practical Field Research.* Intercom, Society for Technical Communication, 2003, December.

[10] Wixon, D. & Ramey, J. (1996). *Field methods casebook for software design.* New York, NY, USA: John Wiley & Sons, Inc.

[11] Wixon, D., Ramey, J., Holtzblatt, K., Hackos, J., Rosenbaum, S., Page, C., & Laakso,S. *Usability in Practice: Field Methods Evolution and Revolution.* Extended Abstracts from CHI 2002: Changing the world, changing ourselves, Minneapolis, MN, USA, 2002.

[12] Wood, L. The ethnographic interview in user-centered work/task analysis. In D. Wixon & J. Ramey (Eds.), *Field Methods Casebook for Software Design* (pp. 35-56). New York, NY, USA: John Wiley & Sons, Inc., 1996.

About the Authors

Stephanie Rosenbaum is founder and president of Tec-Ed, Inc., a 15-person consulting firm specializing in usability research and user-centered design. Headquartered in Ann Arbor, Michigan, Tec-Ed maintains offices in California and New York. An IEEE Senior Member and recipient of an IEEE Millennium Medal from PCS, Stephanie is also active in ACM SIGCHI, the Human Factors and Ergonomics Society, and the Usability Professionals' Association. Stephanie recently co-authored a chapter in *Cost-Justifying Usability, An Update for the Internet Age* and contributed an invited chapter on "The Future of Usability Evaluation" to a forthcoming volume on *Maturing Usability* (Springer HCI Series, 2008) by the European COST294-MAUSE usability research community.

Laurie Kantner is Tec-Ed's Vice President of Client Services. Laurie has conducted numerous field usability studies using contextual inquiry and ethnographic research methods and, in the past five years, employing usability testing in the field as well as in the lab. Studies have taken her as far as Regina, Saskatchewan, into inner-city Chicago, to hospitals in New York City, to the sanctum of computer company help desks and university researchers' labs, and to a rural Indiana county courthouse. Laurie is a member of the Usability Professionals' Association (where she served on its Board of Directors and was chair of the 2002 conferences), the Society for Technical Communication, and ACM SIGCHI.