A PRIMER ON USABILITY TESTING FOR DEVELOPERS OF TRAVELER INFORMATION SYSTEMS

PEGGY CROSBY, JAN SPYRIDAKIS, JUDY RAMY, AND MARK HASELKORN
Department of Technical Communication, University of Washington, Seattle, WA 98195, U.S.A.

and

WOODROW BARFIELD
Industrial Engineering Program, University of Washington, Seattle, WA 98195, U.S.A.

Abstract—This paper reviews the requirements for conducting successful usability tests, particularly on traveler information and traffic management systems; offers a general approach for conducting such tests; and presents a case study of usability tests on an interactive traveler information system.

INTRODUCTION

Many developers of hardware, software, and documentation have found usability testing to be an extremely useful tool for producing products that their intended users will find both easy and appealing enough to actually use. Given that one of the major determinants of a product's success is whether users perceive it as easy to use, usability testing seeks to ensure that a product is usable before it goes to the marketplace, and that users can interact with software and hardware tools without excessive cognitive demand. The purpose of usability tests is to evaluate products, not users—by having users evaluate ideas, prototypes, or products. Through an iterative process, users' evaluations help guide the design of the final product. Gould and Lewis (1985) point out that "user testing will happen anyway"; if it does not happen in the developer’s environment, it will happen in the user’s environment.

Advanced Traveler Information Systems can greatly benefit from usability testing. Because the aim of these systems is to affect travelers' behaviors, it makes sense to involve potential users (travelers) while the systems are still under development. When products are tested in the users' environments only after development, hard-to-use products may be rejected by their users.

During usability testing, developers attempt to answer questions about a product—ranging from ease of use, to the "look and feel" of an interface, to objective measures of how long it takes someone to use a feature, to more subjective issues of whether users like the product or think they might use it. Joseph Dumas (1989) points out that, regardless of the method used, all usability tests share three common factors: the use of human subjects similar to actual end-users, the performance of typical tasks, and the collection of data from subjects' interactions with the product. Whereas everyone agrees that usability testing should have something to do with users, not everyone agrees on what terms users' opinions/needs should be brought into the design process. Until recently, many viewed usability testing only as a final check, as something tacked on at the end of the development cycle to assess whether the system works as intended. If the system does not work as expected at this point, findings are difficult to incorporate without major hardware or software changes.

The prevailing thought within the usability testing community today is that usability testing should be an iterative process incorporated throughout all phases of the design cycle (Gould and Lewis, 1985; Schneiderman, 1987; Poltrock, 1989; Dieli, 1989). With this approach, findings on how the system should work can be incorporated more easily and inexpensively into the product's design. For example, user surveys and questionnaires can be distributed to identify what type of need exists in the user community. Usability testing can help determine what type of system might be most effective (e.g. radio, computer, VMS). Testing can also help determine the best user interface and interaction methods with the system, and can range from broad to specific solutions.
This paper presents the necessary requirements for successful usability tests, a general approach to usability testing, and finally a case study of a usability test on an Advanced Traveler Information System (ATIS), which exemplifies eight stages of a typical laboratory-based usability test.

**Requirements for Successful Usability Test**

For any usability test to be successful, several organizational criteria must be met. First, the organization must establish an interdisciplinary "usability team" representing various components of the organization. All team members, from researchers to developers to usability specialists, must make a concerted effort to work cooperatively. Second, members of the development team must have confidence in the usability results and recommendations, and be willing to incorporate them into the design process. The positive potential of usability testing cannot be realized if developers do not make a commitment to the process and follow through with that commitment. Third, guidelines must be established for conducting meetings with the various players, testing product features and reporting results, and setting realistic time frames for usability tests and design changes. Fourth, developers must be willing to allocate the needed resources (financial and personnel) and to provide enough time for adequate and intensive testing.

**Cooperation within the Organization**

Whether the organization forms its own usability testing team or hires outside consultants, one major condition for the success of usability testing (and therefore the success of the product or user interface) is that the organization work together cooperatively. Usability testing should be a team effort. Gould and Lewis (1985) note that "superior quality can be attained only when the entire user interface, including software, manuals, etc., can be designed by a single group, in a way that reflects users' needs, and then evaluated and tuned as an integrated whole." (p. 307). They suggest that those interested in conducting usability tests follow four common guidelines: (a) understand end-users and their tasks; (b) include end-users on the design team; (c) test the design early in the design cycle using the performance of end-users as a basis for measurement and change; and (d) iterate the process. Cooperation among team members is also important in terms of setting schedules, because testing early and iterating the design process take time to implement. It is important that usability testing not be mentioned from the development cycle when other areas of development run over schedule.

All sectors of an organization can benefit from usability testing, but software developers especially can reap the rewards. A major benefit is the opportunity to come in contact with actual users, either by watching how they use the product or a prototype, or by accompanying a usability specialist to a field site and watching how users perform relevant tasks in their own work environments. Users often have numerous ideas or needs that are quite easy for programmers to incorporate.

**Commitment to Usability Testing**

A concerted effort by all involved will not be enough for usability testing to succeed if the organization is not fully committed to usability testing and to incorporating the usability results into the product. When developers conduct usability tests only because such tests may sound "fluffy" in promotional literature or professional presentations, usability results and recommendations are often ignored or biased. Such attitudes are counterproductive, as usability testing done under these conditions still costs time and money, yet usually does not result in an improved product. Deciding on "standard operating procedures" that involve all developers who can be affected by the results of usability tests helps in achieving cooperation within the organization, as well as a commitment to usability testing.

**Seeing Guidelines**

Planning, preparing, and running a series of usability tests can quite often be very time-consuming. Because the usability testing process typically includes many parallel stages and tasks, it is easy to overlook or bypass necessary steps. For example, because recruiting subjects and preparing test materials often proceed simultaneously, slip-ups can easily occur. Further,
because results of the evaluations are recorded "live"—while subjects are being tested—one can easily forget to record data in a standardized fashion.

Guidelines need to be established for the timing, the participants in the product company, and the content of meetings among various development departments that will be affected by usability testing. The initial meeting should require that representatives from all development departments meet and decide who, what, where, when, and how to test a product (specifics of this are discussed in the next section). In an ideal setting where usability testing is an integral part of the development cycle of a product, the only issue for ongoing meetings might be to set up testing checkpoints. In organizations less familiar with usability testing, other issues need to be addressed, such as what personnel should be involved (e.g., which group of managers to report to at what stages of the testing process), and time frames for meetings and product development (e.g., the timing of tests in relation to development deadlines).

Guidelines for testing or a library of previous test materials and results can save time in preparing future tests. Guidelines or "templates" for reporting results are also important, because representatives from different divisions within the development group may be using the results. If results are presented in a consistent format, people will know where to look in the report for key information. If usability testing is part of all product development cycles, many results can be integrated into the design of future products, as well as current products.

Realistic time frames for testing and integrating test results into the product are critical; in fact, this area can affect test validity more than any other. Many developers are under a deadline to release the product ahead of the competition or, in the case of many transportation projects, user's are in need of renewal or are running out. In both cases, results need to be published quickly.

Unfortunately, when development schedules slip, usability testing frequently slips, is omitted, or is cut back in scope. Cutting back in scope can radically affect the outcome of the tests; consider the impact on external validity (generalizability) of testing only two subjects instead of the originally scheduled fifteen—one cannot be sure that those two subjects really represent the target audience. What if the two participants evaluate the system at opposite ends of the spectrum—who is right? Further, if only two features of the system are tested instead of the eight that were scheduled, one cannot be sure that these two features are the most important in determining the usability of the product. It is for such reasons that contingency plans must be constructed.

Although in many cases some data on a few subjects or features may be better than none, one must be prepared for reduced usability testing time frames. If there is not enough time to test a sufficient number of subjects or features, then perhaps other methods would obtain more useful usability results: for example, expert evaluations, (evaluations of the product by a usability specialist); focus groups (oral discussions of the product by a facilitator and a group of users); or surveys.

Necessary resources

One final requirement for successful usability testing is having and using the proper resources. Two types of resources are needed to ensure the effectiveness of usability testing: (a) a subject database and (b) appropriate equipment and personnel.

Subject database. A large database of potential subjects or evaluators for usability tests is important because the actual resource of evaluators who represent actual end-users can be very time-consuming. A good subject database should include a record of current phone numbers and addresses of potential evaluators, their level of expertise on a variety of products, the number of tests each evaluator has already done for the organization, and any other information that might help the usability group pre-screen evaluators for a usability test (e.g., job description, physical and verbal skills). It is important that the database be kept up-to-date and that new evaluators continually be added.

There is some controversy in the usability testing community about how often to use the same person as an evaluator. Some will use the same evaluator in three to five tests and keep track of previous testing experience in their subject database. Others believe that an individual should be used only once because the experience of evaluating may bias responses in future evaluations. They are afraid of evaluators becoming "professional" subjects or in some way
feeling that they work for the developers and owe them their loyalty, thus creating a halo effect. Although the number of times to use a particular subject is mostly a matter of preference, one must consider the possible implication of this decision on the validity of the test results.

Ideally, subjects should be randomly selected from the group of people meeting the subject selection criteria. Organizations will sometimes bypass this step because of the amount of time required, and use in-house people who "come close" to meeting the subject requirements or who happen to be available to participate in the usability test. Such practices are suspect, as too many factors are uncontrolled. Not only will a cross-section of the relevant population not be represented, but also in-house subjects may belong to the same corporate culture as the tester, be involved in the development of the product, and thus be familiar with company jargon, or find it very difficult to remain objective about the product.

Equipment and personnel. All forms of usability testing require at least a minimum of testing equipment and personnel, whether that equipment be pencil and paper test materials, specialized hardware to run the software to be tested, or audio or video data recording equipment. Necessary personnel might be only the administrator of the test, or several administrators, data keepers, and hardware and software support individuals. Different methods of usability testing will require different equipment and personnel. The important point is that the equipment and personnel requirements be met in a timely manner. All stages of a usability test take time: from audience analysis, to task analysis, to preparing materials, to meeting with developers, to recruiting and training evaluators, to analyzing and reporting test results. Although in an iterative model these stages overlap, testing and implementation of results take longer when fewer people are involved.

A GENERAL APPROACH TO UsABILITY TESTING

Beyond meeting the requirements discussed in the previous section, to perform successful usability tests, one must answer questions about how, why, what, who, and where to test. The following describes a general approach to conducting usability tests for organizations interested in including such tests in the product development cycle. Given this base, readers will better understand the case study of a specific usability test of an ATIS that follows. (For a comprehensive bibliography introducing usability testing, see Ramsey, 1989; for an in-depth discussion of usability issues, see Crosby, 1991.)

Table 1 summarizes a general approach to usability testing, and addresses issues surrounding the questions of how to test, when to test, what parts of a product to test, who to test, and where to conduct the tests.

In the sections that follow, we examine each of these questions individually, attempt to point out where controversies lie, and identify steps that usability specialists can take to make their tests more effective. Our suggestions are somewhat idealized. In the "real world" there are almost always tradeoffs and constraints involving money, time, or resources, as becomes apparent in the ATIS case study.

How should the testing be conducted and what methods should be used?

Ideally, usability specialists should use a combination of methods that include measures for obtaining both quantitative and qualitative data. The selected testing methods ultimately depend

<table>
<thead>
<tr>
<th>Table 1: A general approach to usability testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>How</td>
</tr>
<tr>
<td>When</td>
</tr>
<tr>
<td>What</td>
</tr>
<tr>
<td>Who</td>
</tr>
<tr>
<td>Where</td>
</tr>
</tbody>
</table>
on the research questions—testing goals can vary from assessing whether users can complete a given task in under ten minutes to ascertaining whether users prefer menus over command lines or one menu design over another.

Various types of tests may map to various stages of product development. Most researchers feel that at the beginning of the product development cycle, testing should be of an exploratory nature when the design, interface, and functionality can all be easily explored and modified. A number of methods can be used at the outset, including expert evaluations where a usability specialist evaluates similar products and makes recommendations for the current product; focus groups where potential users of the system meet with a facilitator and discuss their needs and desires for a new system; and user evaluations where potential users give their opinions on mock-up paper and pencil representations of the design or interface.

Once the design of the system becomes more stable, testing tends to be more evaluative, in prototypes or “finished” parts of the product can be tested with actual users. As a product nears completion, testing starts to take on a more competitive nature. The new product can be compared against existing products that are functionally similar, as well as against standards that have been set (for example, the user being able to complete a task in under five minutes).

The choice of methodology is usually determined by a combination of real-world constraints and the product’s place in its development cycle. Again, depending on the goals of a particular test, the methods might differ. Some typical methods used in usability testing are questionnaires, surveys, interviews, observation (both informal and formal where subjects may be audio and/or video recorded), keystroke recordings, timing of tasks, oral protocols (where users interact with and make comments about the product), expert reviews or evaluations, and field studies.

Possible threats to reliability and validity. In discussing methods for evaluating usability, we must mention the issues of reliability and validity usually associated with empirical testing. Test reliability relates to test consistency, whereas test validity relates to how closely a test measures what it intends to measure. Wenger and Spyridakis (1989) assert that although usability testing methodologies are borrowed from formal empirical approaches, the issues of reliability and validity associated with those methodologies are often ignored. “Reliability and validity demand that testers be clear about their operating assumptions and values before they draw inferences or causal conclusions. If testers can identify the inconsistencies among observers, differences among behaviors of interest, changes in behaviors and measures across time and situations, and disparities among methods, they can be more confident and more credible in recomputing modifications to a product” (p. 267).

Wenger and Spyridakis further point out some pitfalls in five common usability testing methods: protocols, surveys and questionnaires, interviews, expert reviews, and direct observation. Table 2 summarizes their findings. Any of these methods can be further confounded by using more than one test administrator, data collector, or observer, as each individual will bring his or her personal bias.

Possible solutions or threats to reliability and validity. There are several steps that usability specialists can take to try to ensure the reliability and validity of their tests. First and foremost, specialists need to have a clear understanding of the methods they use and the pitfalls inherent in those methods (Wenger and Spyridakis, 1989). Iterative testing can also help overcome some of the pitfalls in usability testing caused by real-world constraints, such as limited time, money, or resources (Dediu, 1989).

Additionally, one should understand and acknowledge the difference in experimental controls between carefully controlled experiments and usability testing. Classic, controlled experiments seek to control as many factors as possible, whereas usability testing strives for “high realism.” However, some amount of control and consistency is necessary to ensure the reliability of usability testing results. For example, if the test administrator needs to give instructions to subjects, these instructions (whether oral or written) should be the same for each subject. If tutorials or other preliminary materials are to be presented to subjects, all subjects should receive the same materials. Further, an interruption-free environment is necessary, and a “Do Not Disturb” sign should be placed on the door. Contrariness should also be planned for occurrences such as software failures.

These are only a few of the consistencies that can be built into a test. The important
condition to strive for is one where high realism is stressed, yet enough controls are in place to allow for reliable and valid results.

When and at what stage of development should a product be tested?

As just discussed, the decision of how to test relates to when tests are conducted. The pervasive trend is to employ iterative testing as early as possible—from initial product concept to post-release. Because of the relative newness of usability testing in many organizations, products will often be in different stages of development. As noted earlier, different goals and testing methods are associated with different parts of the development cycle. Several usability specialists have tried to match testing methods to the life-cycle of a product. Gradin, Ehrlich, and Shriver (1987) note that certain methods such as focus groups, competitive analysis (comparing a new idea with an existing product), and needs-finding interviews (finding users’ needs) should be conducted early on in the life-cycle, while tasks are still being defined. They suggest that a walk-through analysis (stepping through specifications and evaluating against a set of usability criteria) be done as the interface as soon as the first pass of the interface is completed.

Diehl (1989) proposes that usability specialists use “exploratory,” “plan,” and generate hypotheses about usability when products are still at the conceptual stage. One can look at “paper and pencil” plans or do a competitive analysis with a related product or previous version. Later in the development cycle, one can iteratively test documentation samples (tables of contents, indices, chapters/sections) and prototypes, and a little later, early versions of the product and documentation. Finally, when documentation is almost complete and alpha and beta versions of the product are available, one can conduct usability tests to confirm early findings. The final phase occurs when the product is shipped; post-release studies can be done to see how the introduction of the product into the marketplace has changed the audience and tasks. Then, usability test results can be used in planning subsequent versions of the product, as well as new products.

Relying on more traditional notions of the measurement theory, Guillemete (1989) describes the timing of tests as either formative or summative. At the early stages of development the emphasis should be on formative evaluation, the results of which will go into the design of the product. At later phases of development, after milestones of some sort (e.g., draft of documentation, prototype), testing should be summative or confirmatory. During both the formative and summative phases of testing, usability experts use both behavioral and non-behavioral techniques. Behavioral techniques center on observing subjects’ behavior while they interact with the product, as well as obtaining their views through surveys and protocol analysis. Non-behavioral techniques include such metrics as editorial and technical reviews, and readability formulas.
What part or parts of a product and documentation should be tested?

Whereas developers in the past were more concerned with assessing functionality instead of actual usability, usability specialists today focus on the usability and relevance of features for given users. The best, but often unrealistic, scenario is to test an entire product, but unless the product is very simple, this is usually too time-consuming. The next best scenario is to test the most important parts of the product that allow a user to get his or her tasks done. The key to what to test lies in a thorough task analysis: any aspect that will be vital to the user’s ability to get his or her tasks done must be evaluated. Additionally, evaluators may want to focus on areas where there have previously been problems, where developers have conflicting opinions, or where there are risky audiences (e.g. a new, unfamiliar user group). If nothing else, evaluating these areas point to specific areas that may need to be tested more thoroughly.

Another concern is ensuring that the tasks being tested are realistic, that users perform them the way they were hypothesized, and that the chosen tasks are in fact the most important for allowing users to get the job done. Additionally, one must decide who should be involved in identifying relevant tasks for users. For example, several years ago Digital Equipment Corporation (DEC) conducted a usability test of a software product where subjects performed an editing task with a text editor on a six-page document. When the “usability engineers” (DEC’s name for specialists) went into the field to follow up on their usability testing, they discovered that users in the workplace typically edited 30-50 page documents, and thus had a need for navigation and search methods, but this need did not show up and was not assessed in the six-page lab test. Usability engineers also found that in the “real-world” users often used the text editor in conjunction with electronic mail and conferencing, thus creating a need for easy integration of text into other media. This need also was not tested (Whiteside, Bennett, and Holzblatt, 1990). One of the reasons that DEC failed to assess the right criteria was that developers set the criteria without a clear understanding of either the audience or its tasks.

Experiences such as the one described caused DEC to move from laboratory-based usability testing to a field study method called contextual inquiry. Contextual inquiry is a type of in-depth field study where the usability engineer and user are co-researchers; that is, together they agree on what the user’s experiences are. This is in sharp contrast to observational methods where the researcher interprets the user’s responses, actions, etc.

To solve similar problems, many companies conduct focus groups and in-person interviews to find out whether they have defined realistic scenarios and tasks, and to verify their task analysis. Others take their tasks into companies that have representatives from the target audience and ask those representatives whether the tasks are in fact those that users find most important.

Who should evaluate a product/product idea/prototype?

The following issues arise in selecting subjects to evaluate a product:

1. distinguishing between users and customers;
2. choosing and verifying the correct audience;
3. determining who should set usability criteria;
4. finding audience representatives;
5. deciding who should evaluate highly confidential products;
6. assessing how many users are “enough”; and
7. dealing with time restraints.

The most important issues concerns ensuring that those who evaluate the product/product idea/prototype are the people who will be using it. In many cases, marketing groups encourage usability specialists to rely on the marketing view of customers, who are often not the actual user. A brief anecdote clarifies how easily one can target the wrong audience. The XYZ company produces a database product. After discovering that a major competitor’s database contains a feature that makes it easy to convert spreadsheet files into database files, XYZ’s marketing department decides to add such a feature, and not only incorporate a file transfer into its product, but also target spreadsheet users (who outnumber database users three to one) who
do not ordinarily use databases. So, for its usability test, XYZ seeks people who are fairly experienced with spreadsheets, but novices at database products.

The first problem XYZ encounters is finding users who fit this description, because most experienced spreadsheet users are somewhat familiar with databases. After finally finding people who fit marketing's desired profile, XYZ has them compare its product with those of two competitors. The tests show that this target audience performed more slowly with XYZ's product than with other products that had fewer similarities to a spreadsheet. However, previous usability tests showed that novice database users (with only marginal spreadsheet experience) preferred XYZ's product. The question then becomes whether XYZ should target all novice database users or only novice database users who are experienced spreadsheet users. The interface will be designed very differently, depending on which audience XYZ targets. XYZ would have benefited from the use of focus groups to verify that it was targeting the appropriate audience.

Dumas (1989) warns usability specialists to beware of how user characteristics are specified by either designers or marketers. He suggests that if there is an inability to articulate user characteristics without a clear picture of the audience's requisite computer skills, job experience, etc., then the audience is not really being understood. Another sign that designers or marketers are not taking a user orientation is if they list functions the product can perform instead of tasks for users to perform.

Obviously, the testing team should be concerned about identifying the correct audience and the best places to find representatives of that audience. This is one of the main areas where cooperation between different departments becomes important. Because marketing and development may have different views of who the audience is, the usability specialist must obtain specifics about the audience from both departments. For example, if the marketing department tells the usability specialist that a word-processing product should serve "Mom and Pop," the first question he or she needs to ask is about Mom and Pop's computer experience. For our purposes, let's assume the usability specialist is told the target audience should be novices. One way to clarify the picture of computer experience would be to construct a chart that lists tasks for different types of products (Ramsey, 1990). Rather than defining users as novices, the specialist should define them in terms of the tasks they do and the frequency with which they do them. For example, with word-processing software, setting margins may be considered a "beginner" task, whereas defining stylesheets may be considered an "expert" task. The resulting user definitions are value-neutral as opposed to qualitative judgments by users who define themselves as "novices" or "experts." The task/user definition chart can also be used to identify appropriate subjects through a subject profile questionnaire where the chart serves as a guide for matching recruits to the targeted audience.

If the product is to be a re-release, information from customer support and/or field service departments will be invaluable. Not only will they have a good idea of who has been using the previous version, but they will also know the ways in which the product has been used.

Another excellent way to find out about users is to conduct field studies. Gould and Doherty-Farina (1988) point to seven questions that can be answered about an audience in a field setting: (a) who the users are, (b) what users' work environments are like, (c) which users have access to what information, (d) how information is used in those environments, (e) why users behave the ways that they do, (f) how accurate and useful the information is, and (g) how satisfied users are with the information.

A problem can exist if the product, product idea, or prototype is so innovative and confidential that the organization does not want anyone from the outside to see it. The best situation is still to bring representatives of the target audience. Barring this, an expert evaluation should work, ensuring confidentiality and saving both time and money. There are, however, two pitfalls with expert evaluations: the "expert" may be too far removed in personality from the target user and therefore be unable to accurately act as the user, and, as only one person does the evaluation, only one point of view is represented.

This of course brings up the question of how many users is enough. Many people feel that information from any number of users (or user surrogates in the case of expert evaluators) is better than none. However, most companies with in-house usability specialists aim for seven to ten people per test condition, depending on the length and complexity of the test. These subjects
should be as homogeneous as possible. Ultimately, one should test enough end users to be able to generalize about the target audience.

When running tests, the usability specialist should watch for anomalies in test sessions between subjects. As already discussed, any anomalies created by dissimilar experiences (e.g., software crashes, interruptions) can affect validity. Anomalies in subject performance, when all conditions are constant across subjects, may in fact reflect another characteristic of the user group not previously defined. For instance, when running a group of commuters in a traffic study, we found that one member of the group answered most questions differently than other subjects. Upon further probing, we discovered that we had not sufficiently defined the audience characteristics.

The last issue concerning who should do the evaluation is what to do when there is not enough time to obtain seven to ten users for each iteration of a product. As noted above, expert evaluations can save both time and money. Questionnaires and telephone interviews can also save time; but the usability specialist will not actually be able to see or evaluate the user using the product. These methods, however, along with focus groups, are very effective for defining both audience and task.

Given the previous mention of both lab settings and field research, it is now appropriate to examine the issue of where to conduct usability evaluations.

Where should testing of a product take place?

More and more companies and researchers are becoming convinced that in order to perform thorough usability evaluations, they need to use a mix of controlled and field settings. As stated earlier in the section on how to test, there are a number of usability testing methods. Many exploratory types of testing, where the goal is to explore the design, interface, and functional possibilities, are well suited to field testing. A field setting can also be very useful when one seeks to define audience and task characteristics for a product. Lab settings, where conditions can be more controlled, are often used for evaluative types of testing where parts of products are evaluated on such criteria as ease of use, clarity, or functionality.

Gould and Doherty-Farran (1988) suggest that specialists go into the field (a) to obtain task information, (b) to identify variables for late surveys and tests, and (c) to validate results and instruments after conducting usability tests. Gould and Doherty-Farran summarize this view when they say that “Usability is best explored in many ways. A complete usability program combines qualitative and quantitative research. Qualitative field research, quantitative surveys, and lab testing are interdependent and should reinforce one another” (p. 332).

A Usability Test on an Advanced Traveler Information System

Given the general approach to usability testing described in the previous section, it is now useful to examine the stages of a real-life usability test conducted on the interface of “Traffic Reporter,” an Advanced Traveler Information System (ATIS). Although not all usability tests will follow the stages as presented, this case study serves as a model for laboratory-based usability tests.

“Traffic Reporter”

“Traffic Reporter” (TR) is a PC-based graphical information system that obtains traffic information from sensors embedded in the freeway, and delivers real-time traveler information to commuters (Haselkorn, Spyridakis, Barfield, Goble, and Garner, 1991). TR’s original interface was developed by University of Washington researchers using information gathered from on-road surveys and in-person interviews of motorists who were commuting in downtown Seattle on Interstate 5 (I-5), the main North-South corridor in and out of the city. The surveys and interviews gathered information on commuters’ driving patterns, route and time flexibility, and the preferred medium for receiving traveler information (Spyridakis, Barfield, Conquest, Haselkorn and Isakson, 1991; Barfield, Haselkorn and Spyridakis, 1991). The survey responses were then analyzed for generalizations about commuters and their preferences.
Four types of commuters were identified:

1. **route changers (RC)**, commuters willing to change route before or after entering the freeway (21%);
2. **route and time changers (RTC)**, commuters willing to change departure time and route (49%);
3. **pre-trip changers (PC)**, commuters willing to change departure time, route, and travel mode before leaving their residence, but unwilling to change route during their trip (16%); and
4. **non-changers (NC)**, commuters unwilling to change departure time, route, or travel mode (23%).

The differences in these four categories of motorists would affect the sampling plan for the usability test, in that the target audience consisted of a heterogeneous group of commuters.

The surveys and interviews also revealed commuters’ preferences for delivery of traveler information:

1. Commuters preferred information delivered at home.
2. Commuters in the on-road surveys least preferred information delivered via computer; however, commuters in the in-person interviews were quite open to information delivery via computer.
3. RCs, RTCs, and PCs needed commute time information and time estimates for alternate routes, and desired feedback to help them verify that they indeed had made the best choice.
4. Commuters liked receiving time information in a numerical format, yet were willing to receive graphical information.
5. Commuters wanted to receive credible traveler information.

While this list reveals only a few of the surveys’ findings, these ideas quickly began to suggest the types of issues to assess in the tests of TR.

The usability tests of TR consisted of eight stages, which are typical of many laboratory-based usability tests:

1. audience analysis,
2. subject recruitment,
3. task analysis,
4. preparation of test materials,
5. pilot test,
6. actual evaluation,
7. data recording, and
8. analyzing and reporting results.

Although these stages often occur recursively and simultaneously, we present them sequentially for reading ease.

**Audience analysis**

In the case of TR, the ultimate end user is the Puget Sound commuter. However, due to cost and hardware constraints at this time, the initial users of the system will be TV and radio media personnel, and Washington State Department of Transportation (WSDOT) traffic managers. Therefore, in conducting the usability tests, we assessed three different user groups: commuters, media personnel, and WSDOT traffic managers. We selected commuters who were most likely to change their behavior based on traveler information: route changers (RC), pre-trip changers (PC), and route and time changers (RTC). Additionally, because TR is a PC-based system and because a majority of commuters in the initial on-road survey ranked computers last as a preferred medium for receiving traveler information while many in the in-person interviews expressed interest in computer delivery, we thought it important to test both those commuters who stated they would and those who stated they would not like to see a computer system developed for delivering traveler information.
Subject recruitment

Subjects were selected from the 1,000 respondents to an on-road survey of southbound commuters who expressed willingness to participate in an in-person interview and who fit the characterizations of the groups mentioned earlier. From these groups, we randomly selected subjects, and then scheduled a two-hour time slot to participate in a usability test at the University of Washington. Media subjects consisted of volunteers from Seattle News media. Traffic managers from the Traffic Systems Management Center (TSMC) in Seattle volunteered to evaluate the TR system.

Task analysis

Based on the survey results that commuters need commute time estimates and information in order to alter their commute choices, we designed tasks that mimicked real traffic scenarios. We had commuters engage in scenarios in which they faced time constraints or heavy traffic; we then asked them to use the system to find time and speed information for that scenario and assess whether they would change their commute behavior based on the information they obtained. We sought responses from commuters to five broad areas of questions: (a) potential change in commuter behavior, (b) interaction with the system, (c) interface design, (d) potential use of the system, and (e) the desired delivery method of the information provided by the system. We wanted to obtain similar information from media personnel, but with a focus on traffic reporting. From WSDOT traffic managers, we sought subjective evaluations of TR as an addition to other information sources they would continue to use.

A brief review of TR’s functions at the time of the usability tests will help clarify the task design. TR’s main screen displayed a map of I-5 with an overview of traffic speeds indicated by different colors. The displayed freeway grid (from the King-Snohomish County boundary to the north to slightly south of downtown Seattle) was edged by labeled ramps. Users could obtain information from the system about the travel times and average speeds between these ramps and about specific sections of the freeway. Since the purpose of the usability tests was to determine how useful the system would be for commuters, we focused on four functions directly affecting commuters.

Zooming. Users obtain speed information by moving a mouse-controlled cursor, which turns into a magnifying glass, over the freeway map. By clicking the mouse button at a specific point, a section of I-5 magnifies, showing mean speed rates, instead of the broad speed ranges indicated on the main screen.

Locating speed and time estimates. Users obtain speed and time estimates for a specific trip by clicking the mouse on any two ramps. An information box to the right of the freeway map appears and shows estimated speed and travel times between the two ramps. Selecting the best exit ramp. Users determine the best exit for a given trip by double-clicking the right mouse button on an origin ramp; a box to the right of the freeway appears and shows estimated rates of speed and travel times to all possible exit ramps currently displayed. Selecting the best entry ramp. Users select the best entry ramp for a given trip by double-clicking the left mouse button on a destination ramp; a box to the right of the freeway map appears and shows estimated rates of speed and travel time from all possible entry ramps currently displayed.

Both inbound and outbound scenarios were presented to the commuters with various tasks within each scenario. The tasks were designed so that the use of almost any of the functions would allow the evaluators to obtain the “right” answer, but there was one method per task that was most effective.

Preparation of materials

Since commuters are the ultimate end-users of the TR system, we focused on their needs when developing test materials. For the media, we modified the commuter test materials, but kept the focus of the tasks on the types of information commuters would need. Two types of materials were used for the usability tests involving commuters: introductory materials and test materials.

Introductory materials. The introductory materials consisted of a consent form, subject profile questionnaire, introduction/demonstration, material, and diversionary task. When sub-
jecv arising at the test site, they signed a consent form describing the purposes and benefits of the test, and any risk, stress, or discomfort associated with the test. It also stated that the subject's identity would remain confidential and that subjects did in fact belong to the anticipated commuter and computer groups. The majority of this questionnaire re-asked questions from the original commuter information survey that the commuters completed in September 1988. We were especially concerned with the questions that would elucidate whether the commuters belonged to the RTC or PC groups, as these groups would be most likely to alter their departure time and routes. The questionnaire also asked about commuters' employment and computer use. We believed that answers to these questions could prove useful for long-range planning.

After the subjects filled out the consent and profile forms, they were introduced to the system. The introduction briefly stated the purpose of the usability test and the purpose, scope, and limitations of the system. This introduction was in the form of a script that the test administrator read aloud so that each commuter was exposed to the same information. Next, a scripted demonstration was read aloud and provided information about the main screen, the mouse, the four functions of the system, and the help screen.

Subjects were then given a tutorial that allowed them to watch the cursor change shapes and to become familiar with the mouse. The tutorial described the four system functions: (a) zoom, (b) time and speed information between two ramps, (c) information about specific entrance ramps, and (d) information about specific exit ramps. For each function, the tutorial contained two tasks. In task A, the subject not only interacted with the system to find the answer to a question, but also received detailed instructions on how to answer that question. In task B, the subject again interacted with the system, but this time no instructions were given.

After the tutorial, but before the actual test, subjects were given a diversionary task consisting of setting an alarm clock to counteract the effect of immediate memory on the ensuing experimental tasks. We wanted to test how well subjects had learned the system by doing the tutorial, not how well they could recall the tutorial. A secondary purpose of this task was to give subjects further opportunity to practice the talk-aloud protocol method they would use during the test.

Test materials. A goal of the test was to obtain both qualitative and quantitative information from commuters about a number of questions regarding the system. The test materials consisted of a task list and an open-ended interview. The task list for commuters consisted of two parts. Part one had two commuting scenarios (northbound and southbound) with four tasks per scenario. Each task required subjects to perform one of the four system functions being tested, record their answers, and at times answer qualitative questions on a five-point scale pertaining to the function they had just used. The qualitative questions focused on commuter behavior (changing route or time), as well as certain aspects of the interface (clarity of main screen information, usefulness of various dialog boxes). Part two examined the system overall and consisted of standard questions using a Likert scale about the interface, commuter behavior based on TR's information, TR's information versus other available forms of travel information, additions or changes to the system, the meaning of arrows on the main screen, preference for delivery of TR's information, and use of the system. The open-ended interview also consisted of questions about commuters' thoughts concerning expansions of the system, different color codings or ways to show traffic flow, and other issues related to commuting and the Seattle freeway system in general. All of the scales in part one and two were five-point scales. On all of the scales but one, the numbers 1, 3, and 5 were labeled with subjective adverbs—"not at all," "somewhat," and "very," respectively. In part two, a frequency scale was labeled "not at all," "occasionally," and "frequently."

The testing materials for media personnel also consisted of a task list and an open-ended interview. The task list for media personnel consisted of two scenarios (the morning and afternoon commuting) with four tasks in the first scenario and three tasks in the second scenario. Task questions focused on the types of information that would be transmitted to commuters based on information received from TR. The open-ended interview for the media was the same as the commuters' interview. TSMC traffic managers received the same introduction and demonstration as commuters and media personnel, but performed no tasks. Instead they were asked to interact with the system, comment on it, and discuss its as a group.
Pilot test

Before evaluating commuters, we ran two pilot subjects through the test: one subject who fit the profile but was not from the actual survey group, and a human factors expert. (It is best to select pilot subjects randomly from the target population; however, due to time constraints, we found someone who matched the subject profile but was not from the actual survey population.) After the pilot tests, we revised materials to make the questions clearer and more concise.

Evaluation and data recording

The usability tests of the TR system took between 1½ and 2 hours per subject. Results were obtained by timing subjects on tasks, tape recording subjects’ talk-aloud protocols during task performance, and having subjects record their answers on task sheets. Additionally, both the data keeper and test administrator recorded the number of mouse moves and user of help, and the functions the commuters used to complete each task. When subjects first arrived at the test site, they were taken to the testing area, made comfortable, and then asked to fill out the consent form and subject profile questionnaire. The testing procedure consisted of a preliminary stage and the actual test stage.

Preliminary test stage. In the preliminary test stage, subjects listened to the introduction, watched the system demonstration, and completed the tutorial and the diversionary task. Before the actual introduction, the test administrator reiterated the importance of the subject’s input and stressed that the subject was not being tested, but rather was helping test the system. It is extremely important to emphasize this point several times during the testing process. Evaluators who feel they are being tested may try to memorize through the tasks or answer questions the way they think the test administrator wants (i.e., the halo effect may occur). By stressing that it is the system being tested and by having the test administrator’s voice and language remain noncommittal, more accurate input about the system can usually be obtained.

The test administrator read the introduction aloud while demonstrating the system to each participant. Subjects were allowed to ask questions afterward, although they were not yet allowed to use the system. Next, subjects performed the tutorial while reading aloud to practice the talk-aloud protocol method. If subjects seemed extremely confused during the tutorial, the administrator tried to get them back on track by asking them questions to help them re-think what they were trying to do. Also, if subjects answered the tutorial correctly but used functions other than the necessary ones, the administrator told them that they had obtained correct information but asked them to try to find the answer again using the specified function. Subjects were told that if at any time during the tutorial or tasks they needed a reminder on any functions, they would have to use the Help screen. Finally subjects completed the diversionary task.

Actual test stage. The actual testing stage consisted of two components: the usability tasks and an open-ended interview. When subjects were ready for the usability tasks, the administrator turned on the tape recorder and asked them to proceed. Using a stopwatch, the data keeper timed the subjects from the moment they finished reading a task until they recorded the correct answer on the task sheet. The test administrator and the data keeper both recorded the number of times subjects accessed the Help screen and the order of the functions used to obtain the answer. In addition, the data keeper noted any special circumstances, such as pertinent comments, outside interruptions, etc. After the subject finished the usability tasks, the administrator conducted the open-ended interview. The administrator, subject, and data keeper discussed various aspects of the system as well as overall traffic situations that commuters encounter.

Some subjects offered solutions to Seattle’s growing traffic problem. When subjects finished, each was thanked and given a silk-screened T-shirt of the TR system as a token of our appreciation for his or her participation.

Analyzing and reporting results

As stated earlier, the questions we sought to answer fell into five categories:

1. changing commute behavior,
2. interacting with the system,
3. interface design,
The answers were measured through scaled information, timed information (including mouse clicks), and ranked information. Additionally, commuters were asked open-ended questions about why they would or would not change a commute behavior, what additional information they might use to help with their commute decisions, and any other comments they might have pertaining to the system or their commute.

The results of the usability tests were reported as part of a final report to the Washington State Department of Transportation (Hasekorn, Barfield et al., 1991). Since the usability section of the report was organized by user category—commuter, media, TSMC—each section contained results and recommendations. In the “Commuter” section, results were broken down at two levels, as can be seen in Table 3 (a reproduction of a portion of the Table of Contents from the report). The top level represents the five general areas we were interested in, and the bottom level addresses specific questions pertaining to those areas.

Although the conduct of usability tests may at first appear to be complex, the stages are successfully performed through diligent team efforts where those involved in product development cooperate and commit themselves to the needs for and benefits of such tests. If those involved in designing usability tests carefully identify and access the correct audience, design and implement appropriate tasks, and analyze and report their results, the tested products should become more user-friendly. Given that many ATIS engineers are trained in empirical research methodology, they must simply apply their knowledge in a new area. However, they must keep in mind the differences between traditional research methods and usability testing. As the goal of usability testing is “high realism,” ATIS developers must allow themselves to relax some of the more rigid controls associated with classic experiments for the gain of ensuring usable systems. Further, because human subjects are the most important part of the usability testing process, to reap the benefits of their views, developers must be willing to learn and accept social science methods for the use of human subjects. Finally, usability tests will only be successful when iterative tests accompany the product throughout the development cycle.

The recent trend in the development of systems to deliver real-time transportation information to commuters and transportation managers, be it through ATIS, dynamic route guid-

CONCLUSION

Although the conduct of usability tests may at first appear to be complex, the stages are successfully performed through diligent team efforts where those involved in product development cooperate and commit themselves to the needs for and benefits of such tests. If those involved in designing usability tests carefully identify and access the correct audience, design and implement appropriate tasks, and analyze and report their results, the tested products should become more user-friendly. Given that many ATIS engineers are trained in empirical research methodology, they must simply apply their knowledge in a new area. However, they must keep in mind the differences between traditional research methods and usability testing. As the goal of usability testing is “high realism,” ATIS developers must allow themselves to relax some of the more rigid controls associated with classic experiments for the gain of ensuring usable systems. Further, because human subjects are the most important part of the usability testing process, to reap the benefits of their views, developers must be willing to learn and accept social science methods for the use of human subjects. Finally, usability tests will only be successful when iterative tests accompany the product throughout the development cycle.

The recent trend in the development of systems to deliver real-time transportation information to commuters and transportation managers, be it through ATIS, dynamic route guid-

Table 3. Commute results section of final report

| Commuter comments on departure time
| Results of departure time questions
| Commuter comments on route
| Results of route questions
| Commuter comments on mode
| Results of mode questions
| Commuter comments on interface design
| Commuter comments on interface detail
| Potential use of the system
| Results of potential use of system questions
| Commuter comments on delivery medium
| Preferred medium for receiving traffic information
| Commuter comments on delivery mode
| Rankings of delivery mediums
| Commuter comments on system control
| Commuter comments on delivery mode and system control
ance, or real-time traffic control, makes the need for usability testing critical to the transportation field. Transportation engineers must recognize the difference between their view of system functionality and capability and the user’s view of system usability. If a majority of users cannot easily and successfully use newly developed systems, functionality will be of little consequence. As more and more tax dollars are spent to help solve traffic congestion around the world, system developers must take responsibility to ensure that their newly developed systems are in fact usable.

REFERENCES