53 Usability Testing

Joseph S. Dumas and Jean E. Fox

CONTENTS

53.1 INTRODUCTION

At the turn of the millennium, a survey of usability professionals showed that they rated usability testing as the most influential method for having a strategic impact on organizations (Rosenbaum, Rohn, and Humburg 2000). At that time, testing frequently was recommended as a key to stimulating product development organizations to integrate user-centered design into the development process. It had strong face validity: it appeared to evaluate usability fairly, and tests always produced a list of usability problems to be addressed.

Testing's face validity and its value as a tool to influence developers delayed the profession's examination of the details of the method, its reliability, and more importantly its forms of validity. In the past decade, books and research studies have looked at both the strengths and limitations of the large variety of practices that are now part of the umbrella term "usability testing." In this chapter, we discuss those new materials. There are two themes that appear throughout: (1) the widespread use of Agile and other streamlined development practices has increased the pressure to test faster and cheaper and to strip testing of some of its essentials and (2) the lack of consensus about the criteria for what constitutes a valid usability test has made it vulnerable to attacks on what were assumed to be its basic foundations.

In a previous edition of this handbook, we focused primarily on the basic concepts of testing practice that were established over the period of its emergence and growth (Dumas and Fox 2007). In this edition, we focus on the body of research and opinion that has emerged during the past decade.

Downloaded by [RMIT University] at 23:36 01 March 2015

53.2 TYPES OF TESTS

The fact that the term "usability testing" refers to a wide variety of methods becomes apparent when one tries to categorize them. There are at least five dimensions to describe a particular test:

1. Purpose of the test—explore the usability of early design concepts, diagnose usability problems, fix usability problems, validate usability, measure baseline usability, or compare usability of products
2. Scope of the product tested—the whole product, part of it, and/or selected task flows
3. Location of sessions—local or remote
4. Presence of a test moderator—moderated or unmoderated
5. The level of functionality of the product—paper prototype, static screens, interactive prototype, or live code

We are not aware of any empirical data about the frequency of test types. We believe that the most common test, at this time, is a moderated diagnostic test on a subset of a product conducted locally in the middle of development. While the stated desire of many in the usability profession is to test earlier, it is not clear that early tests are most common, though testing has moved from the late stage method it was 20 years ago.

Alternative protocols are gaining in popularity. As we see in this chapter, remote and unmoderated online tests are more common. The rapid iterative test and evaluation (RITE) method is an example of a local, moderated test of a whole or part of a product, conducted early in development, with the purpose of fixing rather than finding problems.

In addition to this classification of types of usability tests, there are other terms that are used in the literature and in practice to describe tests: qualitative, quantitative, formal, and informal. What these terms denote is not always clear. They add to the ambiguity about what a usability test is.

53.3 TRADITIONAL DIAGNOSTIC USABILITY TEST

Over the past 20 years, the basic characteristics of a moderated, diagnostic usability test have been established:

- The focus is on usability. The traditional usability test is intended to uncover usability issues both positive and negative.
- The participants are end users or potential end users. Most usability professionals would agree that to have a valid diagnostic usability test, the participants must be part of the target market for the product. The key to finding people who are potential candidates for the test is a user profile (Branaghan 1997) or a persona (Pruitt and Adlin 2005). A user profile captures two types of characteristics: (1) those that the users share and (2) those that might make a difference among users. The test team must also determine how many participants per user group to include in the test. Five to eight users has become a common sample size.
- The participants perform tasks with a product or prototype, usually while thinking aloud. One of the essential requirements of every usability test is that the test participants attempt tasks that users of the product will perform. When a product of even modest complexity is tested, however, there are more tasks than there is time available to test them, so it is necessary to sample tasks. While not often recognized as a limitation of testing, the sample of tasks is a limitation to the scope of a test. Those components of a design that are not touched by the tasks the participants perform are not evaluated. Almost without exception, testers present the tasks that participants do in the form of a task scenario. For example, consider the following:

> You have just bought a new combination telephone and answering machine. The box is on the table. Take the

product out of the box and set it up so that you can make and receive calls.

- Before the test session starts, the administrator instructs the participant how the test will proceed and informs the participant that the test probes the usability of the product, not the participant's skills or experience. In most diagnostic usability tests, participants are asked to think aloud.
- The data are recorded and analyzed. In a usability test, there will be both quantitative and qualitative data. Quantitative data include measures of efficiency (e.g., task times), effectiveness (success rates), and satisfaction (ease-of-use ratings). Qualitative data include participant comments and tester observations. The data can be collected and recorded in a variety of ways. In the early days of usability testing, the test administrators recorded all data by hand with stopwatches and clipboards. Over the years, numerous tools have become available to automatically record video and data. Many of these tools also conduct basic data analysis, such as calculating average task times and success rates. Much of the data analysis involves building a case for a usability problem by combining several measures—a process that has been called "triangulation." In addition, problems are usually categorized by their severity.
- The results of the test are communicated to appropriate audiences. Test reporting began with lengthy written reports and highlight tapes, but reporting has become less formal.

53.4 UPDATING USABILITY TESTING BASICS

While many usability tests still are consistent with the traditional basics, the variations on what are still called "usability tests" have grown. In this section, we discuss how testing evolved.

53.4.1 From Usability to User Experience

Beginning about the year 2000, there was a concern that the "traditional" view of usability was limiting. These efforts have led the professional to ask whether task effectiveness, efficiency, and satisfaction are only part of the story. For example, Quesenbery (2004, 2005) broadened the ISO definition by adding e*ngaging:* "how pleasant, satisfying, or interesting an interface is to use" (Quesenbery 2004, p. 5). Others have advocated looking beyond traditional views of the scope of the profession to consider "user experience," shaped not only by usability, but by aesthetic, emotional, social, and business factors (Jordan 2002; Teague and Whitney 2002; Karat 2003; Hancock, Pepe, and Murphy 2005). Many industry groups have changed their name from "usability" to "user experience" groups.

This broadened view of what it takes for a product to be successful has had two important implications for usability practice. First, traditional usability measures are being adapted to assess the broader notions of user experience. Second, new methods are being used to supplement the more traditional ones (e.g., Karat 2003; Pagulayan et al. 2003; Murphy, Stanney, and Hancock 2003). Usability practitioners are supplementing traditional measures with value-based metrics and methods drawn from the marketing, anthropology, and psychology disciplines. Questions such as "Is it fun?," "Is it motivating?," and "Does it provide enough variety (as opposed to consistency)?" are a few examples of what usability practitioners are asking today in addition to "Is it usable?"

As a result of these changes, usability testers are including more subjective measures into tests, and testing is often paired with marketing methods such as online surveys to broaden the scope of the evaluation beyond usability issues.

53.4.2 Are Five Still Enough?

Part of the popularity of usability testing has come from its ability to find usability problems with only a few participants. Anyone who watches multiple test sessions with the same set of tasks perceives that the same issues begin to repeat, and that somewhere in the five-to-eight test participant range, with the same user population, it begins to seem unproductive to test more participants. So it was with great joy that testers greeted the research studies by Virzi (1990, 1992), showing that 80% of the total number of usability problems that will be uncovered by as many as 20 participants will be found by as few as five. Virzi also found that those five participants will uncover most of the problems judged by experts to be severe. This finding has been confirmed several times (Faulkner 2003; Law and Vanderheiden 2000). Practitioners conducting diagnostic, moderated tests continue to select small numbers of participants, confident that they are finding most of the problems that they could find.

Those findings lead to a popular rule of thumb for diagnostic tests that "five is enough." But the interpretation of the rule is not as simple as it appears. Among others, the rule has been attributed to Nielsen (2000). But Nielsen placed the rule into an iterative testing context in which he proposed that three iterative tests of the same product each with five participants are better than one test with 15 participants.

In Section 53.12, we discuss some recent studies showing that a single usability test only finds a small fraction of the total number that multiple independent tests will find. How do we reconcile that finding with the studies of sample size? All the studies that have looked at sample size and the number of problems found have done so with a single test by one test team. Apparently, there is a limitation in how many problems a single test team can find. At this point in time, we do not know why test teams have this limitation.

Furthermore, all the sample size studies, except Lewis (1994), tested very simple applications. As Redish (2007) points out, we know very little about the optimal usability testing process with complex systems. So the rule of thumb

Downloaded by [RMIT University] at 23:36 01 March 2015

would be more accurate if it said that five participants will uncover about 80% of the problems that one team can find with a small application. That fact also means that adding more participants may not find more problems as long as the test team does not change.

There also have been a few challenges to the generality of the "five is enough" rule of thumb, most notably by Lewis (1994, 2001) and Turner, Lewis, and Nielsen (2006). Their challenge makes the reasonable case that tests differ in the probability of problem detection. A moderately complicated product being tested for the first time might indeed yield many of its problems with five to eight participants. Those authors have looked at problem detection over a large sample of tests and found that the average probability of detection is about 30%.

But what about a product that is being retested after most of its problems have been fixed? One might expect that it might take more participants because it is harder to detect the problems. It also may take more participants if the user population is very heterogeneous, such as with elderly and disabled users (Grossnickle 2004; ITTATC 2004; Swierenga and Guy 2003). Turner, Lewis, and Nielsen (2006) created and verified a formula for determining how many participants are needed in a variety of testing situations.

Finally, the pressure coming from organizations using an Agile development approach is to test with even fewer than five participants (see Section 53.7). Krug (2010) suggests monthly tests with three participants each. He argues that each test will find more than enough problems to keep the team busy for the next month. Again, Krug is saying that his rule needs to be viewed in an iterative testing context.

53.4.2.1 Sample Sizes with Other Testing Types

Most of the dialog about minimum sample size and all the research have been done in the context of diagnostic test with a moderator. The minimum sample size for comparison and baseline tests is much larger because of the need to measure usability not just to find problems. Minimum sample sizes for those types of tests are similar to those for cognitive science research studies, about 12–15 per group.

One of the strengths of online unmoderated testing is that much larger samples are easier to obtain. These larger samples can make the results of online tests more credible. By adding survey questions in addition to tasks, such tests can gather market research as well as usability data (Albert, Tullis, and Tedesco 2010).

53.4.3 Are "Real" Users Necessary?

The first books on testing procedures stressed that it is necessary to recruit test participants who are part of the target market for the product (Rubin 1994; Dumas and Redish 1993). The rationale was that all the problems that the target market will have would not be uncovered if a different population is tested. This rationale was based on a logical analysis and anecdotal evidence.

The methods for identifying the qualifications of participants were asking marketing experts in the organization, developing a user profile, or more recently, using personas (Pruitt and Adlin 2005). However, these methods result in ranges of qualifications that are difficult to cover with a small sample. For example, if one of the qualifications is knowledge of a software operating system, do you select participants with a little or a lot of experience? The advice is to make sure you have a range, some with a little and some with a lot. This strategy may mean that two subgroups are combined into one. Furthermore, as participants are recruited, compromises in the details of the qualifications are often made. Consequently, the final sample only approximates the profile or persona.

For usability testing, as with other types of research, it is nearly impossible to draw a random sample of the population. You may be limited by issues such as geography (e.g., those close enough to come to your lab), availability (e.g., who can participate during business hours), or willingness (e.g., who wants to participate). To some extent, every usability test sample is at least partly a sample of convenience. The challenge for testers is to determine which characteristics might affect the participants' experiences with a product.

One of the consistent results of tests is that they always yield lists of problems, often long ones. It has seldom been necessary to question whether a different sample would have yielded a different list. But the pressure from Agile development and from startups to get websites to market faster has led to a practice called "hallway" testing (Spolsky 2000), in which "you grab the next person that passes by in the hallway and force them to try to use the code you just wrote. If you do this to five people, you will learn 95% of what there is to learn about usability problems in your code." Krug (2010, p. 42) makes a similar point, "But there are many things you can learn by watching almost anyone use it (a website)."

Until a research study shows that a sample of "real" users, that is people who are part of the target market, yields the highest quality list of problems, some practitioners will continue to see value in recruiting a more convenient sample. As long as such samples uncover usability problems, it will be difficult to argue that the sample invalidates the test.

53.4.4 Does Task Selection Matter?

An essential component of any usability test is that participants attempt tasks. The measures taken during and after tasks provide the empirical data on which the product design is evaluated.

The selection of tasks is a function of the purpose and scope of the test. Testers must also consider the order of the tasks. In some cases, the tasks must be completed in a particular order, such as when a later task relies on the results of an earlier task or when there is a natural task order. In other tests, the order of tasks is randomized or varied in some way to balance any order or start up effects.

Task selection has been identified as one source of the lack of agreement in independent tests. Molich et al. (1998)

Downloaded by [RMIT University] at 23:36 01 March 2015

concluded that differences in usability test results across four teams were at least partially explained by fact that the teams use different tasks. However, Molich and Dumas (2008) found that even when teams used almost the same tasks, the problems they listed did not appear to have any more agreement than for teams with quite different task sets. This may have occurred because the task statement is only the starting point for the task. Participants can go down very different paths from the same starting point, thereby exposing different flaws.

In addition to the types of tasks to include, it is also important to consider the number of tasks. Lindgaard and Chattratichart (2007), using the same data as Molich and Dumas, found that the number of tasks used by teams was significantly correlated with the number of problems found, while the number of test participants recruited was not. Interestingly, they also found that the number of participants was not significantly correlated with either measure or the number of problems found. In this case, the number of tasks had greater influence on the number of problems found than on the number of participants.

Most of the advice about task selection and wording has been given in the context of moderated tests. The challenges of creating tasks for unmoderated online tests are quite a bit different (Albert, Tullis, and Tedesco 2010). Task statements for unmoderated tests must be clear and unambiguous because there is no moderator to clarify them. Careful piloting of wording is essential. "Easy to understand" is not the same as "easy to guess," as the participant may guess rather than perform the task. The best tasks are ones whose successful completion is obvious, such as an answer to a question that can be found on a web page. It may be necessary to constrain the participant in the path they use to complete a task to be sure the test is probing the product design appropriately. Finally, sometimes the participant must indicate whether they believe that they completed the task successfully. In such cases, an analysis of their path through the task may be needed to supplement their belief in their success.

53.4.5 Incorporating Thinking Aloud into Usability Testing

One of the early differences between a usability test and a research study was that the test participants typically thought aloud in a usability test. While concurrent thinking aloud is normally done as part of a diagnostic usability test, it is really a method of its own. It has been used in psychological research since the turn of the twentieth century, but it is best known as a cognitive psychology method for studying short-term memory (Ericsson and Simon 1993). Retrospective thinking aloud, that is thinking aloud while watching a video recording of task performance, is also used, especially in situations in which concurrent thinking aloud cannot or should not be done.

Concurrent thinking aloud provides usability testing with most of its drama. Without thinking aloud, it is unlikely that usability testing would have become the most influential usability method. It is the think aloud protocol that grabs the attention of first-time visitors to a usability test and gives a test session the appearance of a science-based method.

When usability testing was first being codified, thinking aloud was borrowed from cognitive psychology without much reflection. It was not until shortly after 2000 that usability specialists began to look at it more closely. Independently, Boren and Ramey (2000), Dumas (2001) and, more recently, Nielsen, Clemmensen, and Yssing (2002) went back to look more closely at what Ericsson and Simon (1993) had described and whether testing practitioners were really following that method. Those reviews showed that the descriptions of how to use the think aloud method that had been provided to usability testing practitioners by Dumas and Redish (1999) and Rubin (1994) were in direct contradiction to the instructions used in cognitive psychology research in which participants are discouraged from reporting feelings or expectations or to make any verbal diversions over and above the content of their actions. In usability testing, participants are encouraged to report on their feelings and expectations and on additional relevant issues.

Only a few research studies have been done on the think aloud method in a usability testing context. Krahmer and Ummelen (2004) compared typical usability testing think aloud instructions to the instructions used by Ericsson and Simon and found that the research instructions do not work well in a testing context. Ebling and John (2000) traced each usability problem found in a usability test back to its source in the test measures. They found that over half of the problems identified in their test came from the think aloud protocol alone. Their study supplements an earlier one by Virzi, Source, and Herbert (1993), who showed that fewer problems are identified when the participants do not think aloud. Eger et al. (2007) found that concurrent and retrospective think aloud protocols found approximately the same number of usability problems. However, when they included eye movements in the retrospective cue, they uncovered significantly more usability problems than in the traditional think aloud condition.

Two interesting questions about thinking aloud are "can everyone think aloud while performing another task?" and "should thinking aloud be done in all tests?" There are now a number of studies and demonstrations that suggest that many user populations cannot perform tasks and think aloud at the same time, including the following:

- Teen and preteen children (Als, Jensen, and Skov 2005)
- Low-literacy populations (Birru et al. 2004)
- People for whom English is not their first language (Evers 2004)
- People from some non-English speaking cultures (Evers 2004)

Evers (2004) conducted think aloud tests and post-test interviews with a sample of 130 high school students from England, North America, the Netherlands, and Japan.

Downloaded by [RMIT University] at 23:36 01 March 2015

The moderator was English. The Japanese students had the most difficulty with the think-aloud sessions. They felt uncomfortable speaking out loud about their thoughts and seemed to feel insecure because they could not confer with others to reach a common opinion. The English also needed reassurance before feeling comfortable with thinking out loud.

Concurrent thinking aloud also is to be avoided in tests of voice response system, tests using eye trackers (Bojko 2005), and tests that include complex tasks or complex environments (Redish and Scholtz 2007). van den Haak, de Jong, and Schellens (2003) found that participants performing complex tasks exposed fewer problems using concurrent thinking aloud than with retrospective thinking aloud.

Some authors have proposed alternatives to concurrent thinking aloud. Redish and Scholtz suggest using retrospective thinking aloud for testing complex and open-ended tasks. Als et al. used a technique with children called constructive interaction, in which children work in pairs on tasks. The pairs who used constructive interaction exposed more usability problems than the children who used thinking aloud. Strain, Shaikh, and Boardman (2007) conducted concurrent think aloud tests with blind participants and found the audio from the screen reader interfered with the conversation. Although the method worked when participants were familiar enough with the screen reader to pause and restart the audio easily, the authors suggest considering retrospective think aloud or what they call "Modified Stimulated Retrospective Think-Aloud." With this method, the participant walks through the application after completing the task.

Frøkjær and Hornbæk (2005) proposed a technique called "Cooperative Usability Testing" as a way to deal with the difficulties participants sometimes have with concurrent thinking aloud. In their technique, there are two parts to a test session. In the first part of the session, interaction, a test participant performs tasks while thinking aloud in the presence of an evaluator. The session is videotaped and the participant is allowed to ask questions of the evaluator, who takes a more active role than is typical. In the second part of the session, interpretation, the participant and one or more evaluators discuss the video of the interaction session with the goal of clarifying the usability problems. In their study, Frokjaer and Hornbaek report that evaluators and participants liked the cooperative technique and that it uncovered more problems. In addition, participants who just did a traditional think aloud session made negative comments about thinking aloud, including that it was hard to think aloud and perform difficult tasks or read text and that thinking aloud felt like "asocial" monolog. Participants also reported that what they were saying out loud was only a fraction of what they were thinking internally. Similarly, Eger et al. (2007) found that participants rated concurrent think aloud sessions as significantly more unpleasant and unnatural than a retrospective think aloud session. These studies are among the few to record comments about thinking aloud from the participant's point of view. We need more studies that provide data on what the thinking aloud experience is really like for test participants.

Studies of how thinking aloud instructions are actually given and how test moderators prompt participants to think aloud show that moderators are inconsistent (Boren and Ramey 2000; Norgaard and Hornbaek 2006). The think aloud method described in the books on testing techniques and in this section of the chapter is simply not followed in practice.

53.4.6 New Research on Testing Measures

Some recent studies have begun to clarify the relationships among the measures that are taken during tests. Testers have assumed that the measures should correlate. Usability problems often are identified through their impact on multiple measures. For example, a structural problem with the organization of a website might cause task failures, longer task times, errors, the need for assistance, and the participants rating tasks or the product as hard to use.

On the other hand, if the measures were highly correlated, testers would not need so many of them. Frøkjær, Hertzum, and Hornbæk (2000, p. 345) argued that, "Unless domain specific studies suggest otherwise, effectiveness, efficiency, and satisfaction should be considered independent aspects of usability and all be included in usability testing." Supporting that point, Hornbæk and Law (2007) reported weak correlations among efficiency, effectiveness, and satisfaction, with an average Pearson-product moment correlation (r) of about +0.2. The correlations were equally weak among time-on-task, completion rates, error rates, and user satisfaction. But many of the studies they analyzed were not usability tests.

Sauro and Lewis (2009) conducted an analysis of data from 90 summative usability tests conducted in industry settings. The pattern of correlations added some complexity to the discussion of whether measures do or do not correlate. They found that correlations among the performance measures were all significant and in the medium range, around or slightly higher than +0.5. The correlations between the performance measures and post-task ratings were slightly lower. Lowest of all, around +0.2, were correlations between post-test ratings and performance measures. Sauro and Lewis then performed a factor analysis on the correlations, which produced two factors: the first is heavily loaded with the three performance measures while the second is heavily loaded with the subjective measures. They argue that this pattern provides support for a construct of usability with a performance and subjective component and that using multiple measures increases the reliability of testing data.

53.4.6.1 Subjective Measures

Recent studies have begun to clarify several issues about subjective measures in usability testing. One of the issues is the format of rating scales. Tedesco and Tullis (2006) compared five different rating scale formats used for post-task ratings. They found that a simple five level Likert scale from very difficult to very easy was the most reliable. But none of the formats had acceptable reliabilities below sample sizes of

Downloaded by [RMIT University] at 23:36 01 March 2015

8–10 participants. Sauro and Dumas (2009) confirmed those findings and also found that a simple subjective mental effort scale performed as well as the Likert scale.

Tullis and Stetson (2004) conducted a similar study of post-test questionnaires, such as the System Usability Scale (SUS). They found that the 10-question SUS was the most reliable and that none of the questionnaires was reliable with sample sizes below 10 participants.

As discussed earlier, several studies have shown that correlations between post-test questionnaires and other measures are among the weakest. Sauro and Lewis (2009, p. 1617) notes: "It is reasonable to speculate that responses to post-test satisfaction questions elicit reactions to aspects beyond the immediate usability test (past usage, brand perception, customer support)."

For the practitioner, this research means that simple subjective measures are to be preferred, that none of these measures are reliable with the sample sizes typically used in diagnostic testing, and that post-test subjective questionnaires are tapping into factors beyond what happens during the test session.

53.4.6.2 Online Testing Measures

Online tests provide the potential for additional measures. Click stream data can show pages visited, page transitions, and how much time users spend on pages or key areas of pages (Albert, Tullis, and Tedesco 2010). For example, looking at pages visited during failed tasks can provide additional clues about design flaws. The larger sample sizes with online tests also make it possible to break the total population of participants into smaller segments, which is usually impossible with the small samples used in moderated tests.

53.4.7 New Ways of Reporting Test Results

In the early days of user testing, the test team almost always created a formal report and a highlight video tape. Testers needed those deliverables to communicate what they did, what they found, and to justify the testing method itself. Now, it is more common for the results to be communicated more informally, such as by scheduling a meeting soon after the last test session to discuss the results and/or creating a slide presentation for a briefing that may also contain sections of video from the sessions. Collaboration tools such as Wiki workspaces are also used to create "living" documentation to which subsequent design recommendations and user-interface concepts are added (Luef and Cunningham 2001).

53.5 TESTING STEPS OUT OF THE LABORATORY

With remote usability testing, the test administrator and participant are in different locations. Hartson and Castillo and colleagues began exploring remote usability testing as early as 1996 (Hartson et al. 1996; Castillo, Hartson, and Hix 1998). Tools to conduct remote usability testing were becoming available, and they saw the benefits of remote testing. Since then, technologies have improved and become less expensive, making it easier to conduct the tests. As a result, user experience professionals have continued to develop and explore methods of remote usability testing.

There are a number of advantages to remote testing:

- You can reach a worldwide population of participants because you are not limited to the local testing area. This may be especially helpful when there are not many users, and they are geographically dispersed.
- It is easier to get participants to volunteer because they do not have to travel.
- Participants work at their desks in their work environments, which may make them more comfortable and the testing more realistic. This can be especially helpful in recruiting disabled participants, who may find it difficult to travel or who use specific assistive technologies when they use the computer.
- You do not need a usability lab.

In the past, the technology to conduct such sessions was not good enough to allow usability specialists to get the information they need (Dumas 2003). That is no longer true because of several factors:

- The Internet has made it possible for usability specialists and participants to work together without installing special hardware or complex software on both the tester's and the participant's computers.
- There are tools available for instrumenting websites to collect usability measures automatically and to insert questions and ratings as the participants work.
- Collaboration software that works over the Internet makes it possible to share desktops and control the cursor.
- Recording software makes it possible to store good quality video and sound in files that are not large by today's standards, often less than 50M for a 2-hour session.
- PC processors and RAM are fast enough to run both recording software and the application you are testing simultaneously. In addition, participants often have broadband or high-speed transmissions, so they are not limited by slow modem connections.

Remote testing takes two forms: (1) synchronous, in which the moderator and the participant work together, communicating over the phone or through their computer, and (2) asynchronous, in which the participants work on their own without the direct guidance of a moderator. Each has its strengths and weaknesses.

53.5.1 Synchronous Remote Testing

Synchronous remote testing is similar to a traditional usability laboratory test, except that the participant and tester are in different locations. The two methods tend to use similar

Downloaded by [RMIT University] at 23:36 01 March 2015

protocols and similar methods of analysis. As a result, synchronous remote tests generally also involve the small numbers of participants.

With synchronous remote testing, the participant and moderator will use screen sharing software so that the moderator and other team members can observe what the participant is doing.

Typically, the administrator cannot see the participant (a webcam can be used, but usually is not). We do not yet know what the impact of not seeing the participant is, but one laboratory study indicates that usability specialists judge usability problems as less severe when they cannot see the participant's face (Lesaigle and Biers 2000).

Some remote testing configurations may present security problems. For example, participants could obtain screen shots without the knowledge of the moderator. In addition, allowing participants to share applications on computers inside your organization's firewall may be prohibited. Some organizations may be able to address this with a nondisclosure agreement, while others may require a special computer outside their firewalls.

Downloaded by [RMIT University] at 23:36 01 March 2015

53.5.2 Asynchronous Remote Testing

With asynchronous remote testing, participants complete the tasks on their own, and the test team reviews the session results later. Recently, the first book length discussion of this type of testing has appeared (Albert, Tullis, and Tedesco 2010). These are unmoderated tests. Asynchronous remote testing can be conducted by providing the participant with two browsers (one for the product or prototype and one for instructions). The instruction browser includes the tasks to be attempted, buttons to click at the beginning and end of a task, a free-form comment area, and questions or ratings to answer during or after each task. Asynchronous remote tests can also be conducted with tools specifically designed for that type of testing. Whichever arrangement is used, participants must be able to start and complete the entire test session on their own.

The primary advantage of asynchronous over synchronous testing is a larger sample size, because the number of participants is not limited by time requirements of the moderator. In addition, participants can complete the study at their convenience. For example, Tullis et al. (2002) tested 88 participants in a short period of time.

The disadvantage is that you cannot see or interact directly with the participants. However, in the Tullis et al. (2002) study, the participants provided an unexpectedly large volume of feedback in the free form comment field. These comments provided insight into the usability problems with the product.

53.5.3 Comparing Laboratory and Remote Testing

There have been just a few studies comparing results of usability tests conducted in a laboratory and remotely, and the results are not always consistent. Relating to performance measures, Tullis et al. (2002) reported no substantial difference between asynchronous testing and laboratory testing in terms of performance measures. However, West and Lehman (2006) found that asynchronous remote participants completed the tasks faster and were more likely to abandon a task than participants in a laboratory, but they showed similar success rates. Further, Thompson, Rozanski, and Haake (2004) also reported that asynchronous remote participants were faster. They also reported that these participants made fewer errors.

Regarding the number of problems identified, both Tullis et al. (asynchronous) and Thompson et al. (synchronous) found no difference with laboratory testing. However, in a study with blind participants, those in the asynchronous remote condition found fewer problems per website than those in the laboratory condition (each participant completed 2 tasks on each of 10 websites) (Petrie et al. 2006).

Clearly, we need to better understand the benefits and challenges of each method. As a result, research on this topic continues and is expanding into new domains such as remote testing with mobile devices.

53.5.4 Testing Mobile Devices

Conducting usability tests with mobile requires that testers be able to see both the screen of the device and the participants' hands. Early efforts used a computer-based emulator or a single camera pointed at a mobile device mounted on the table. These configurations captured the participants' interactions with the devices, but the experience was not realistic. Testers then developed creative solutions to capture the screen and the participants' hands. For example, both Catani (2003) and Schusteritsch, Wei, and LaRosa (2007) attached two small cameras to mobile devices, one to capture the screen (since many mobile devices have no "video out") and one to capture the participants' hands.

Another challenge is that mobile devices are intended to be used "on the go," not in the quiet office setting typically simulated in a usability test. Factors such as weather, signal strength, and background noise can all impact the users' experiences. In addition, mobile device users are often preoccupied by other tasks.

Several studies have evaluated the differences in the results from both laboratory and field usability tests, but there is little consistency in the findings. Kaikkaner et al. (2005) found exactly the same problems in both a laboratory and a field setting. Betiol and Cybis (2005) found more usability problems with a phone mounted to a desk than with a computer-based emulator or with a camera mounted on a mobile device used in the field. Duh, Tan, and Chen (2006) found more critical problems in the field than in the laboratory. On the other hand, Kjeldskov and Stage (2004) found more usability problems in the laboratory than in the field. However, the differences appear to be primarily in problems classified as "cosmetic," not in problems classified as "critical" or "serious." The great variety in the research results suggests that we need to continue to study this issue to better understand the methods of testing mobile devices.

53.6 ROLE OF THE TEST ADMINISTRATOR

Most usability specialists learn the skills of moderating tests through apprenticeship. They watch a few sessions, then moderate a few sessions under supervision. Quite quickly they move into a journeymen status during which they almost never receive feedback on their interaction skills unless they request it. In the first book published on the topic, Dumas and Loring (2008) have provided a systematic rationale for how to moderate a test session. They describe 10 rules for interacting with participants that put the first stake in the ground on the topic. The rules attempt to cover the common situations that moderators encounter rather than unusual incidents.

Dumas and Loring propose that moderators play three separate but overlapping roles:

1. The gracious host, who is responsible for making participants feel welcome from the moment they arrive to the moment they leave and who attends to their physical comfort, ensuring that the session goes smoothly and that they have a positive experience overall
2. The leader, who respects participants but who is clearly in charge of the direction and pacing of the session
3. The neutral observer, who is unbiased and objective and who keeps interactions to a minimum while providing support and encouragement to the participant when needed

Balancing those roles is one of the skills new moderators learn.

53.6.1 Training and Education of Moderators

Some of the early books on testing had chapters describing the skills needed and how to deal with selected situations. In the past 10 years, there have been a few Master's degree programs teaching moderating skills. But most usability professionals still learn on the job from more experienced moderators (Dumas 2007).

The usability profession has not established any educational or training qualification to become a moderator. Dumas and Loring (2008, p. 7) list the following qualifications:

- Understanding the basics of usability testing
- Interacting well with test participants (using our 10 rules)
- Ability to establish and maintain rapport with participants
- Lots of practice

Krug (2010) believes that all that is needed to be a competent moderator is a few hours of training in a workshop. He restricts his view to diagnostic testing. He says that he has never seen a bad moderator. He has challenged his readers to bring him a case in which a moderator has made a product less usable as a result of user testing. He believes that encouraging more moderators to run more tests is a path to making technology work better for its users. Clearly, we need more research on what makes a successful moderator.

53.7 FITTING TESTING INTO AN AGILE PROCESS

One of the important forces from outside of the user experience community that has had a major impact on its practices is Agile development (Frishberg 2010). Agile development methodology grew out the frustrations that the software industry has had managing the development process. After more than 25 years of trying, software was still released later than planned, over budget, and filled with bugs. Previous to Agile, the most common approach to development was the "waterfall" method, a sequential software development process in which progress is seen as flowing steadily downwards (like a waterfall) through the phases of conception, initiation, analysis, design, construction, testing, and maintenance. Starting about 2001, Agile development was a reaction against the waterfall model. The term "Agile" refers to a family of processes that share some common characteristics. Product requirements are addressed in a series of 2–4 week cycles by a dedicated team that is co-located. Each cycle ends with tested, working code. While code is documented, paper documents such as specifications are not part of the process.

While Agile development has begun to grow in popularity, it is unclear how traditional user experience methods, especially usability testing, can be integrated into it. Over the past few years, user experience professionals have been changing the way they work to remain players in these fast moving Agile cycles. Some of the important changes have been the following:

- Practicing iterative design and evaluation. The concept of iterative evaluation has been touted for decades, but the traditional waterfall model with traditional testing made iteration expensive and hard to justify (why are we testing again?). Iteration was, perhaps, the least practiced principle of user-centered design. Because iteration is at the foundation of the Agile model, user experience professionals have had to find a way to implement it. One approach is for the user experience team to be on a separate track from the coders, a track that is one cycle ahead (Lu, Rauch, and Miller 2010). While the user experience team is on Cycle 2, the coders are on Cycle 1. The user experience team does its user research and design concepts for Cycle 3 while conducting usability testing on the Cycle 1 user interface. The testing that is done is usually with very small samples and sometimes with internal staff rather than target users. Quantitative measures typically are not taken.

Downloaded by [RMIT University] at 23:36 01 March 2015

- Integration into the development team. The friction between user experience professionals and developers using the waterfall model kept user experience professionals on the outside looking in. Testing was often performed too late to impact design, and developers often viewed testers as people good at finding fault rather than fixing problems. The Agile method requires all members of the team to be co-located and to be engaged full-time. This face-to-face contact seems to create more cooperation and respect than was typical with the waterfall model.

Fitting testing into an Agile model was been facilitated by a new approach to testing (Medlock et al. 2005). Known as the RITE Method, it focuses on fixing designs rather than just finding problems. In outline, the method consists of the following:

- Key decision makers for the product participate in the study with the usability specialists.
- The team selects the tasks to be run and attends all sessions. As with traditional usability testing, users who are part of the target market for the product are recruited and sessions use the think aloud method.
- After each session, the usability specialist identifies problems and their severity. The team then decides whether they have enough data to verify each problem and how to refine the design to address the problem.
- The design team refines the design and tests it with the next participants.
- Problems are identified again, including whether the refinements have mitigated previous problems. If not, new refinements are created.
- The team decides which problems they can fix and which need to be examined in more detail or require resources that are not currently available.
- Additional participants are run until the major problems have been fixed or there are no more resources to continue.

With its emphasis on iteration and an integrated team, the RITE method fits nicely into the requirements of the Agile model (Douglass and Hylton 2010). Both Agile and RITE have the potential to change the way testing is performed and perceived. Those methods put pressure on testers to conduct tests quickly, to focus on fixing problems, and to require that developers be present during sessions.

As this chapter shows, since the early days of testing, there has been an emphasis on a faster process, scaled-down reporting, and getting modifications into the product. The traditional laboratory test with 5–8 participants, taking 2–4 weeks, with a report following some days later fit well into the waterfall model but not into the Agile model. The new approaches have some advantages in that they are more integrated into development and provide for iteration. But they also have the potential to make it convenient to test very small samples and to not use target users. It remains to be seen as testers move farther from the tradition testing basics whether diagnostic testing will remain as the most influential evaluation method. We desperately need some research to evaluate how effective testing is with these new models.

53.8 WEBSITE TESTING TOOLS

53.8.1 Eye Tracking

Eye tracking has slowly become more prevalent in usability testing. The technology has advanced to a point where it is noninvasive and almost unnoticeable to participants. Further, the software available to analyze the data also has improved. Although the equipment is still expensive, the prices are more affordable than in the past. Testers can even rent eye tracking equipment for short-term use at an even lower cost. These factors have led to an increase in the use of eye tracking in usability testing.

Eye trackers indicate where a participant is looking throughout a task or a whole test session. Eye trackers emit a pattern of infrared light (invisible to humans) and track the reflection of these patterns on the participants' eyes with special cameras. Participants no longer need to wear bulky head devices or stabilize their head when using an eye tracker. (Some eye trackers that can be used outside the laboratory are head-mounted but they are not as cumbersome as earlier models.)

Eye trackers can generate huge data files but vendors have developed sophisticated software that has greatly simplified the analysis process. This has been essential to the growing popularity of eye trackers, as they can sample data up to 120 times per second. Testers can now quickly determine the number or length of fixations on any particular area of a stimulus (such as a web page).

Testers can use eye tracking data in several ways. Eye trackers can be set up to allow observers to follow the participant's gaze during the test session. The test moderator can then tailor post-test debriefing questions based on patterns observed during the test. For example, if the participant spent a lot of time looking at a feature, the test moderator can ask what the participant thought of that feature.

Testers can also use the quantitative data generated by the eye tracker in post-test analyses. These eye tracking results can provide additional insights into participants' behaviors. The data can answer questions such as "Which areas of the page did participants look at most?" and "Were there areas they did not see it all?" For example, Albert and Tedesco (2010) used eye tracking measures to determine if self-reported awareness of items on a screen are reliable.

Running a usability test with eye tracking is not difficult but does require some additional planning. For example, testers will have to adjust their screening process slightly. Eye trackers may have difficulty tracking certain people, such as those who wear some styles of the bifocals. Also, the screeners should inform potential participants about the eye tracking and encourage them to bring whatever vision correction

Downloaded by [RMIT University] at 23:36 01 March 2015

they need to see the screen easily. Participants who do not bring proper vision correction often sit too close or too far from the screen, where the equipment cannot track them.

In running a usability test with eye tracking, there are certain issues to consider:

- The informed consent should mention the eye tracking.
- The test protocol will have to include about 5–10 minutes at the beginning of each test session to calibrate the eye tracker to the participant.
- Scenarios for eye tracking tasks should not use a think aloud protocol. Participants look at the screen differently when they are thinking aloud (Bojko 2005). The tester may use eye tracking with some scenarios but not others to get a variety of information.

The analysis software for eye trackers can display the data in a variety of ways. Testers should understand the types of data they will be collecting and determine which are the most appropriate to address their issues (Bojko 2009; Poole and Ball 2005). Testers who want to use quantitative data should be sure to have enough participants to warrant statistical analyses (Goldberg and Wichansky 2002).

Eye tracking also has some disadvantages. It is difficult to conduct eye tracking studies with dynamic content, which includes not just video, but also objects such as cascading menus or pop-up message windows. The analysis software may present the results as if all the activity occurred on the original stimulus page.

Testing can be expensive, not just in terms of equipment, but also in terms of additional time to recruit participants, calibrate them during the test session, and analyze the results afterwards. In addition, because some participants cannot be tracked, the pool of possible participants becomes more limited. Testers may have to plan for additional participants in case some participants cannot be tracked (Schnipke and Todd 2000).

Despite these costs and challenges, eye tracking data can be very helpful in understanding participants' behavior. The data can help testers identify areas of confusion or point out objects participants missed entirely. Thus, although eye tracking is not standard usability laboratory equipment now, given the benefits of eye tracking, along with advancements in the technology, it is likely that the use of eye tracking will increase in the future.

53.8.2 FirstClick Testing

FirstClick usability testing is a method for evaluating the structure of a website. Wolfson et al. (2008) developed FirstClick testing as a way to conduct card sorting within the context of the actual website. They felt that the standard form of card sorting, using only labels and possibly brief descriptions for each "card," did not provide the same context as the website itself. They used it as a closed card sort, after designing wireframe options based on a more traditional open card sort.

In FirstClick testing, participants are given a task to complete. However, the scenario ends after they click on their first link. Researchers record the link selected and the time required in making a selection. Wolfson et al. also suggest having the participants rate their confidence after each selection. By aggregating data across participants, researchers can determine where users expect to start specific tasks. Researchers can see whether participants correctly selected the first link and whether the expectations were consistent.

To conduct a FirstClick test, researchers will need at least a somewhat functional wireframe of the homepage. The links must be active, but the second-level pages can just have a "task complete" message. With just a wireframe, researchers can conduct FirstClick testing fairly early in the development process, before the organization of the site has been established.

53.9 BASELINE AND COMPARISON TESTS

Some tests have a measurement focus, either for benchmarking a product's usability or comparing the usability of different products or versions. These performance-based tests tend to be summative and more like research experiments than a typical diagnostic test.

At present, the usability specialist's interpretation of summative usability test data plays a large role in evaluating the product's usability. Experienced usability professionals believe that they can make a relatively accurate and reliable assessment of a product's usability when considering the following:

- The product is stable.
- The number of participants is sufficiently large (larger than for most diagnostic tests).
- Participants are discouraged from making lengthy comments or evaluative statements in their think aloud protocol.
- The test administrator makes minimal interruptions to the flow of tasks.

The primary objective of a baseline test is to establish a standard against which other products or future versions of the product tested can be compared. By testing with the same set of tasks, a company can measure whether a new design has improved the usability of the product.

An important variation on the benchmark test is one focused primarily on comparing usability. Here the intention is to measure how usable a product is relative to some other product or to an earlier version of itself.

There are two variations:

- A diagnostic comparison test focused on finding as much as possible about a product's usability relative to a comparison product

Downloaded by [RMIT University] at 23:36 01 March 2015

- A summative comparison test intended to produce results that measure comparative usability and/or to find the winner

In both these tests, there are two important considerations:

- The test design must provide a valid comparison between the products.
- The selection of test participants, the tasks, and the way the test administrator interacts with participants must not favor any of the products.

As soon as the purpose of the test moves from diagnosis to comparison, the test design moves toward becoming more like a research experiment. In considering the design for the comparison, there are two important decisions:

- Will each participant use all the products, some of the products, or only one product?
- How many participants are enough to detect a statistically significant difference?

In the research methods literature, a design in which participants use all the products is called a "within-subjects" design, while in a "between-subjects" design, each participant uses only one product. If one uses a between-subjects design, one avoids having any contamination from product to product, but one needs to make sure that the groups who use each product are equivalent in important ways, and the sample size must increase. Because it is difficult to match groups on all the relevant variables, between-subject designs need to have enough participants in each group to wash out any minor differences. An important concern to beware of in the between-subjects design is the situation in which one of the participants in a group is especially good or bad at performing tasks; Gray and Salzman (1998) called this the "wildcard effect." If the group sizes are small, one superstar or dud could dramatically affect the comparison. With larger numbers of participants in a group, the wildcard has a smaller impact on the overall results. This phenomenon is one of the reasons that summative tests have larger sample sizes than diagnostic tests. The exact number of participants depends on the design and the variability in the data. Sample sizes in summative tests are closer to 20 in a group than the 5–8 that is common in diagnostic tests.

If one uses a within-subjects design in which each participant uses all the products, it eliminates the effect of groups not being equivalent and can have a smaller sample. However, one then has to worry about other problems, the most important being order and sequence effects and the length of the test session. (See Dumas (1998) for rules on counterbalancing.) One also has to be concerned about the test session becoming so long that participants get tired.

Perhaps the most important factor in the fairness of the comparison is the selection of tasks. The participants must perform the same tasks with the products. Anyone familiar with the products being compared is capable of selecting a sample of tasks that would favor one product. Consequently, some third party, perhaps an industry expert, who is not familiar with the details of the products but is familiar with the typical tasks users perform may be asked to select the tasks. Or a company conducting an internal comparison might ask a team independent of the test team to select the tasks.

The focus of the data analysis in a baseline or comparison task is usually on measures of performance and standardized subjective ratings rather than on qualitative measures that point to usability flaws.

53.10 TESTING WITH SPECIAL POPULATIONS

There is a growing literature about testing with special populations, including the following:

- International participants
- People with physical disabilities
- The elderly
- Children

This literature has been summarized by Dumas and Loring (2008). This section presents a brief summary of findings relevant to usability testing.

53.10.1 International Participants

Many manufacturers look for new customers across the globe. However, preparing a product for a new market may involve more than simply translating the language. Cultural differences can also impact appropriate design decisions such as color selections and the use of images. These differences can also impact the appropriate structure for web applications. Because of the significant differences across cultures, it is important to conduct usability testing with participants from all the target cultures.

International usability testing follows the principles and theories of generic usability testing. However, there are a variety of challenges with testing participants in other cultures that generally do not apply when testing in one's own culture. The challenges of communication and cultural differences are described below.

53.10.1.1 Communication

One of the most significant challenges with international usability tests is communication. Often, there are different languages. Sometimes the testers are bilingual, but often the tester must have helped recruiting participants, preparing test materials, conducting the test, analyzing the results, and writing the report. Nielsen (1996) and Vatrapu and Pérez-Quinones (2004) offer several suggestions including the following:

- Use employees of the company who live and work in that country. This may require training the employees to facilitate a usability test.

Downloaded by [RMIT University] at 23:36 01 March 2015

- Conduct the test in the participant's language using an interpreter.
- Hire a local usability firm.
- Run the test remotely.
- As a last resort, conduct the test yourself in your language, though this method is likely to be unnatural for the participant.

Tests that are conducted in the participant's language must be translated. Some testers prefer to have the translator work real-time during the test. The translator can either serve as a liaison between the tester and the participant (adding significant time to the test) or between the test administrator and participant (who are both speaking the same language) and the observers. The tester may also have to make arrangements to provide the test report in more than one language.

53.10.1.2 Cultural Differences

Other cultural differences may also impact a usability test. As noted earlier, Evers (2004) conducted think aloud tests and post-test interviews with a sample of 130 high school students from England, North America, the Netherlands, and Japan. There were several key differences including the finding that participants from Japan and the United Kingdom were uncomfortable thinking out loud. There may be gestures considered natural or friendly in one culture, but offensive in another. Vatrapu and Pérez-Quinones (2004) report that when both the participant and the test administrator were from the same culture, the participants engaged in more think aloud behavior and the usability tests revealed more problems.

53.10.2 Disabled Participants

Usability tests with disabled participants require careful planning. Testers must understand the participants' disabilities and adjust their procedures accordingly. Several researchers have published "lessons learned" from their experience with disabled participants (Coyne 2005; Grossnickle 2004; the Information Technology Technical Assistance and Training Center (ITTATC) 2004; Lepistö and Ovaska 2004; and Swierenga and Guy 2003). Some of these lessons include the following:

- Recruiting disabled participants is more time consuming than recruiting general population participants. Local organizations and support groups may be willing to help.
- Disabled participants may need assistance getting to the usability lab.
- Consent forms must be accessible to all participants.
- Blind participants may require electronic or Braille versions.
- Participants with learning or cognitive disabilities may require special consideration to ensure they understand the test and their rights.
- Deaf participants may require a sign language interpreter, who needs to be informed about the goals of the study.
- Participants with disabilities may require extra assistance understanding the tasks and may have trouble thinking aloud. Strain, Shaikh, and Boardman (2007) conducted concurrent think aloud tests with blind participants and found the audio from the screen reader interfered with the conversation.
- Participants with physical disabilities may require adaptive technology to interact with the computer. Be sure the devices are working before participants arrive.
- Because of the great variability in disabilities, it may take more participants than typical usability tests.
- It can be especially difficult to observe participants who use Braille readers, as there is currently no good way to follow what the participant is reading.

Overall, tests with disabled participants may take longer than expected; testers should schedule enough time so that participants are not rushed. Further, participation may be more taxing than for general population users, and so the test should limit the number of tasks evaluated (Coyne 2005). Finally, testers should ask participants before the test whether they need any special accommodations.

53.10.3 Elderly Participants

As the population ages, manufacturers are looking to expand their market to this growing population. Seniors are more active than ever. As a result, many manufacturers are working to ensure that their products are usable by their older users.

As people age, the diversity in their abilities increases. They may also have disabilities, such as those mentioned in the previous section. Many of the concerns and issues mentioned earlier also apply with elderly participants. In general, testers should be prepared for each participant, leaving plenty of time for each person.

There may also be generational issues. Testers should be aware of what their participants expect regarding social interaction and courtesy. Chisnell, Lee, and Redish (2005), Coyne (2005), and Tedesco, McNulty, and Tullis (2005) provide some guidance based on their experiences running usability test with older participants.

53.10.4 Children as Participants

When designing a product for children, usability tests must target children. Although the process is generally the same as with adult participants, there are a few important differences.

Recruiting children actually involves recruiting their parents. Patel and Paulsen (2002) suggest several good sources for recruiting. They recommend building rapport with organization leaders and parents. It is important to pay attention

Downloaded by [RMIT University] at 23:36 01 March 2015

to the needs of both the parents and the child. Sometimes it is necessary to have the parents in the room during the test, especially for very young children. Investigators should be sure that the parents do not unnecessarily interfere with the test. However, investigators should be flexible, as each family will be different.

Investigators may want to alter the usability laboratory itself to be a better environment for children. Most usability laboratories use a standard office layout and décor. Although this is fine for testing adults, it is not the most welcoming to children. Making the room more "child friendly" can make children more comfortable and willing to participate.

The tasks should accommodate the abilities of children in the target age group. Investigators should consider (1) the wording of the instructions to be sure they are at an appropriate grade level and (2) whether the participants are old enough to complete the tasks. For example, children may not be able to perform a task and think aloud simultaneously. As mentioned earlier, Als, Jensen, and Skov (2005) used a technique with children called constructive interaction, in which children work in pairs on tasks. The pairs who used constructive interaction exposed more usability problems than the children who used thinking aloud.

Finally, what motivates adults does not always motivate children. Hanna, Risden, and Alexander (1997) suggest age-appropriate approaches for motivating young participants to continue. Most likely, the best approach for a preschooler is very different from that for a teenager.

Children can be unpredictable, so one or more members of the test team must understand the skills, abilities, and expectations of the children in the target user population. This will help testers to respond appropriately to unexpected situations.

53.11 HOW TESTS ARE ACTUALLY CONDUCTED

While there are many books and articles that describe how usability testing ought to be practiced, there have been few studies of how tests actually are conducted. The Comparative Usability Evaluations (CUE), especially the first two, inspected test reports from commercial usability laboratories (Molich et al. 1998; Molich et al. 2004). By reviewing the reports, the study authors saw the procedures used as well as the quality of the reports. There have been two other studies in which test sessions at commercial laboratories were observed and recorded (Boren and Ramey 2000; Norgaard and Hornbaek 2006). The results of these studies taken together are not encouraging. There is a large discrepancy between what testers actually do and what didactic texts say they should be doing.

The CUE studies looked at test reports from 13 organizations. No two reports were alike. They described tests from 4 to 50 participants with widely varying sets of tasks for the same product tested, leading or poorly designed task scenarios, different measures taken, and reports with few descriptions of the profiles of participants or procedures used.

Norgaard and Hornbaek watched and recorded 14 test sessions from seven different companies. They also recorded many discussions, analyses, and informal conversations among the usability evaluators before and after the sessions. They found that evaluators asked questions that were leading, questions asking participants to predict future outcomes, and questions that put words into the participants' mouths, such as "So … you feel more secure now … or?" (p. 215) There were two additional findings that are cause for concern. First, there was no systematic analysis while the results of a session were still fresh in evaluators' minds. Evaluators did not discuss findings during or directly after the sessions. Second, the behavior of evaluators indicated that they were confirming usability problems that they had found by inspecting the design themselves before the test started. Their tasks, questions, and probes were designed to support their own preconceived opinions about what the problems were. When participants' ratings disagreed with the evaluator's opinions, they were dismissed without further analysis.

While the Boren and Ramey and Noorgaard and Horbaek studies did not make other measures of participant's performance, a recent study has (Olmsted-Hawala et al. 2010). In that study, participants were assigned to various think-allowed conditions, ranging from "silent," with no think allowed or interaction with the test administrator, to "coaching," where the test administrator asked direct questions about the participant's thoughts and behaviors, which is what moderator's typically do in diagnostic testing. The results showed that when moderators are free to probe and ask questions, participants complete significantly more tasks and rate the product as more usable. That study is the first evidence that the moderator's behavior can change participants' behavior as well as the participants' perception of the product.

These studies suggest that usability testing as actually practiced is another important source of variability in usability measurement.

53.12 RELIABILITY OF USABILITY TESTING

Prior to about 1998, practitioners assumed that two equally competent teams conducting independent tests on the same product would have a large degree of overlap in the problems they detected, especially for problems judged to be severe. Jacobsen, Hertzum, and John (1998) were the first to study the reliability of testing in a laboratory experiment. They looked at how evaluators differ when analyzing the same usability test sessions. Four usability testers independently analyzed the same set of videotapes of four usability test sessions. Each session involved a user thinking aloud while solving tasks. Forty six percent of the problems were uniquely reported by single evaluators and all four evaluators agreed on only 20% of the problems. Furthermore, none of the top 10 most severe problems appeared on all four evaluators' lists.

Downloaded by [RMIT University] at 23:36 01 March 2015

In that same year, the first of the CUE studies appeared (Molich et al. 1998). It reported that of 141 unique problems found by four professional testing teams, only one problem appeared on all four lists. Subsequent CUE studies have also reported low levels of agreement (Molich et al. 2004; Molich and Dumas 2008).

Hertzum and Jacobsen (2001) conducted the first meta-analysis of reliability studies and termed the lack of agreement on problems "the evaluator effect." They also clarified the metric of agreement, recommending the any-2 agreement method. Any-2 agreement is the average of $|P_i \cap P_j| / |P_i \cup P_j|$ for all ½ $n(n-1)$ pairs of evaluators—the total problems found in common divided by total problems found between two evaluators. Any-2 agreement has become the most commonly reported metric of reliability in assessments of usability evaluation. Using that metric, Hertzum and Jacobsen (2001) found only a 11% agreement among independent evaluators of a usability test.

To date, there have been more than two dozen papers with data on the reliability of testing, and they all show relatively low agreement rates. Several factors have been proposed to explain the low agreement:

- Evaluators use different tasks and task scenarios.
- Users explore different parts of the product.
- Participants are chosen based on different qualifications.
- Evaluators have different skills, experience, and training.
- Evaluators bring different biases to the test.
- There are no objective problem criteria.
- There is no metric for determining when two problems are the same or different.

53.12.1 Severe, Serious, or Just "Show Stoppers"

Several practitioners have proposed scheme for rating the severity of usability problems: Dumas and Redish (1999), Nielsen (1992), Rubin (1994), and Wilson and Coyne (2001). The schemes differ on a number of dimensions. In addition, many organizations have created their own scales. The reliability of severity scales has been questioned by several studies. Jacobsen, Hertzum, and John (1998) asked four experienced usability testers to watch tapes of the same usability test and then identify problems, including the Top 10 problems in terms of severity. None of the Top 10 severe problems appeared on all four evaluators' lists. Lesaigle and Biers (2000) reported a disappointing correlation coefficient (+0.16) among professional testers' ratings of the severity of the same usability problems in a usability test. Molich and Dumas (2008) found that 25% of the problems reported in common by two or more evaluation teams were classified into different severity categories.

The results of these studies strike a blow at one of the most often mentioned strengths of usability testing—its ability to uncover the most severe usability problems. At this point in time, we do not know whether the inconsistencies in severity judgments are the result of the poorly designed scales, the differing perceptions of usability specialists, the lack of training in how to make severity judgments, or all three.

53.12.2 Testing Is No Longer a Gold Standard

Several authors have proposed that usability testing be used as a gold standard against which to compare other evaluation methods (Andre, Williges and Hartson 2003; Sears 1997; Bailey, Allan, and Raiello 1992; Desurvire, Kondziela, and Atwood 1992). Their argument is that only problems identified by testing are true problems or hits. When other evaluation methods identify problems not found by testing, those problems are by that very fact not considered to be true problems. They are considered to be false positives, sometimes called false alarms. Those papers have been particularly harsh in their criticism of inspection by experts, such as heuristic evaluation.

There are two reasons to reject testing as a standard. First, as we have just described, independent tests find different subsets of problems. Second, Molich and Dumas (2008, p. 263) compared problem detection with testing and with expert inspection. They reported, "…there was no practical difference between the results obtained from usability testing and expert reviews for the issues identified. It was not possible to prove the existence of either missed problems or false alarms in expert reviews."

An issue not discussed in the literature comparing evaluation method is whether one should expect experts in usability evaluation to agree. There is a large body of literature on expertise showing that agreement among experts in most fields is low. It may be that disagreement among usability specialists is not any worse than it is among experts in medicine, biological, and social science disciplines (Shanteau 2001; Aboraya et al. 2006).

53.13 VALIDITY OF A USABILITY TEST

While there has been a good deal of research and analysis about the reliability of testing, there has been almost nothing written about its validity. Validity always has to do with whether a method does what it is supposed to do. There has never been a published study questioning whether testing finds problems. Perhaps the validity of usability testing has been ignored because, no matter how they are designed, tests always find strengths and weaknesses in a product. It has strong face validity. Testers believe when they have finished a test that they have uncovered the most important design flaws. But is finding problems enough?

To truly assess the validity of usability testing, we must first agree on what a usability test is supposed to do. Prior to the mid-1990s, the usability community used diagnostic tests primarily to uncover usability problems. The more problems found, the better and, of course, the tests should find the most severe ones. Because testing has never been viewed as the only usability evaluation method to apply during development and because, ideally, there are iterative tests performed,

Downloaded by [RMIT University] at 23:36 01 March 2015

it was not essential or expected that one test would find all the problems.

The RITE method, discussed earlier, suggests two additional possibilities for goals:

1. A test should provide the data for and confirm the usability of an improved design.
2. A test should increase the commitment of a development team to user-centered design and its willingness to pursue it for future projects.

Fifteen years ago, Sawyer, Flanders, and Wixon (1996) proposed that the measure of validity for usability inspections should be how many of the problems that it identifies are actually fixed in the design. That criterion also could be applied to usability testing.

Until we sort out the importance of these goals (finding problems, creating an improved design, team building, and problems fixed in the design), we cannot fully understand the validity of what is arguably our most powerful usability assessment tool.

53.14 TESTING ETHICS

Informed consent is a method testers used to ensure that usability test participants have the information they need to decide whether to participate in the session. Millett, Friedman, and Felten (2001) state that "informed" requires the tester to be disclosing the necessary information in a manner that the participant can comprehend. They define "consent" to be the voluntary agreement to participate, made by someone competent to make such a decision.

Participants complete informed consent forms at the beginning of the test session. The forms themselves vary widely across organizations, but are generally expected to include the following information (Burmeister 2001):

- A brief description of what the participant will be expected to do
- A statement that participation is voluntary and that the participant can withdraw at any time without penalty
- Any potential risks the participant will be exposed to
- A description of any benefits either to the participant directly (including incentives) or to the population at large
- The name and contact information for the person responsible for the test
- How the test will handle all records from the test session (i.e., the extent to which data will be kept confidential), including the following:
  - Measures resulting from the test
  - Direct quotes from the participant
  - Video and/or audio recordings of the session (including whether the video will show the participant's face)
  - Eye tracking data

Sometimes, testers use forms that allow participants to choose whether or not they will allow the testers to release video, quotes, and so on.

Completing the informed consent form usually only requires a participant to read and sign the form. Some testers follow the good practice of reviewing the form with the participants to be sure they understand and are aware of all the information.

However, in some cases, the informed consent process is not as straightforward.

Some disabilities make it difficult for participants to read, understand, and/or sign the form. When testing low-vision and blind users, the form should be presented in an accessible format. This may mean sending the form to participants ahead of time or providing Braille or large print versions (Henry 2007; Swierenga and Guy 2003). Testers may need to help physically disabled participants sign the form. In addition, testers may also need consent forms for sign language interpreters if they appear in any recordings (Henry 2007). When participants have cognitive disabilities, testers should be sure to provide the informed consent in a manner that each participant can understand.

When testing minors under the age of legal consent, the testers must get a signed consent form from a legal guardian, often a parent (Ellis, Quigley, and Power 2008). The guidelines from the U.S. Department of Health and Human Services (2008) allow guardians to fax in their forms. When the participants are old enough, it might be beneficial to have them sign a form as well, being sure to use age-appropriate language. This will help ensure that the minors understand their participation is voluntary.

When conducting one-on-one remote testing, it can be difficult to get signed consent forms before the test session starts. Dumas and Loring (2008) provide a sample electronic form that can be e-mailed to remote participants. For online testing, the testers usually do not know who the participants are and there is no audio or video recording. There still may be emotional or psychological risks to online participants, but that issue has not been explored in the literature.

When testing international populations, it is important to be sure the consent form is in a language each participant can understand. Also, there may be special requirements for information contained in the forms based on local regulations. As with disabled participants, you may also need consent forms for interpreters.

53.14.1 Additional Ethical Principles

The principles of informed consent and confidentiality that have been discussed in the HCI literature have been borrowed from ethical practices in biomedical research. We believe that, on the whole, testers have followed practices borrowed from biomedical research appropriately but have not been aware of some additional principles from social science research (House 1990).

Because of the often dramatic harm that biological and medical experimentation have caused in human history,

Downloaded by [RMIT University] at 23:36 01 March 2015

ethical principles to protect participants have focused on costs and benefits that result from the application of procedures that occur during research studies in those areas. By analogy, the usability test has been treated as a variation on the research experiment. Consequently, the focus has been on informed consent being voluntary and knowledgeable and on confidentiality restricting the use of participants, name and, sometimes, video image. It has been assumed that risks of physical harm to participants in usability tests are minimal.

On the positive side, the sample informed consent forms in the literature describe the activities participants will be asked to perform, their right to withdraw at any time without penalty, the methods used for recording, and the restrictions on the use of data including the use of participants' names and images. But the possible risks of psychological or emotional harm and challenges to self-esteem are seldom mentioned. Perhaps, testers are afraid that mentioning those possibilities will bias participants to have a negative attitude toward the product being tested. The analogous situation in biomedical research would be not to mention a potentially harmful side effect because it might bias patients to expect such effects.

There is a large volume of literature that stresses the differences between biomedical and social science. There are at least two areas that are relevant to usability tests. First, in the social sciences, the researcher and the participant are often presented as equal partners in the investigation. They work together as colleagues (Murphy et al. 1998). This is the way diagnostic usability testing typically is framed. For example, in formative tests, the test administrator is more engaging and active toward participants. While this approach is intended to make participants feel empowered and more comfortable, it can do just the opposite when participants struggle and fail at tasks. When that happens, it presents the tester and participants with a situation that is not covered by the typical informed consent statement. The hidden assumption behind letting participants fail is that, in a utilitarian accounting of harm, it is better that a few participants fail so that potentially many future users will not fail (see Dumas and Loring 2008). This utilitarian approach to ethics runs counter to a different approach that says that it is unethical to use a harmful means to achieve a beneficial end (Macklin 1982). According to that approach, knowingly causing distress and possibility lowering self-esteem cannot be justified without informed consent. At a minimum, informed consent forms for usability testing should describe the possibility of emotional distress.

A second difference is that violations to participants' confidentiality may come during the reporting phase, which may occur long after the test sessions (Hammersley and Atkinson 1995). While test reports almost never mention participants by name, their user role is often described. A common strategy for emphasizing the priority of a usability problem is to quote participants' negative descriptions of the product or even the company developing it. In tests with small populations that are performed on internal rather than commercial products such quotes may be attributable to particular individuals who then face the embarrassment of exposure. In addition to quotes, it is now technically easy to attach a segment of tape to a slide presentation showing the quote or task failures. In these situations, participants are used as ammunition in the battle between testers and developers over whether changes will be made to the product. Testers need to be aware of the ethics of these situations and take extra precautions to ensure that the identity of participants is not revealed.

53.15 CONCLUSION

Usability testing has evolved in line with changes in the user experience field. For example, practitioners have been exploring ways to work faster and cheaper and to be less formal in their preparation and reporting. In addition, we are just beginning to understand the impact of long-accepted think aloud methods. We now have a better understanding of the standard usability measures, but we also have new technologies, such as eye tracking, which provide new sources of data. Some issues, such as the number and types of participants to use, continue to be debated with no clear resolution. As evidence of this, there has been a push to find ways to conduct tests with both local, convenient participants (e.g., hallway testing) and diverse participants (remote testing). So although there has been progress on many fronts, there are still many areas left to explore.

REFERENCES

Aboraya, A., E. Rankin, C. France, A. El-Missiry, and C. John. 2006. The reliability of psychiatric diagnosis revisited: The clinician's guide to improve the reliability of psychiatric diagnosis. *Psychiatry* 3(1):41–50.

Albert, W., and D. Tedesco. 2010. Reliability of self-reported awareness measures based on eye tracking. *J Usability Stud* 5(2):50–64.

Albert, W., T. Tullis, and D. Tedesco. 2010. *Beyond the Usability Lab: Conducting Large-Scale Online User Experience Studies*. Burlington, MA: Morgan Kaufmann Publishers.

Als, B., J. Jensen, and M. Skov. 2005. Comparison of think-aloud and constructive interaction in usability testing with children. In *Proceedings of the Conference on Interaction Design and Children*, 9–16. (Boulder, CO), New York: The Association for Computing Machinery.

Andre, T., R. Williges, and H. Hartson. 2003. The effectiveness of usability evaluation methods: Determining the appropriate criteria. In *Proceedings of the Human Factors and Ergonomics Society, 43rd Annual Meeting*, 1090–4. (Denver, CO), Santa Monica, CA: The Human Factors and Ergonomics Society.

Bailey, R. W., R. W. Allan, and P. Raiello. 1992. Usability testing vs. heuristic evaluation: A head-to-head comparison. In *Proceedings of the Human Factors Society, 36th Annual Meeting*, 409–13. (Atlanta, GA), Santa Monica, CA: The Human Factors Society.

Betiol, A. H., and W. Cybis. 2005. Usability testing of mobile devices: A comparison of three approaches. In *Proceedings of the Tenth IFIP TC13 International Conference on Human-Computer Interaction*, 470–81. (Rome, Italy), The IFIP Technical Committee on Human-Computer Interaction.

Birru, M. S., V. M. Monaco, L. Charles, H. Drew, V. Njie, T. Bierria, E. Detlefsen, and R. A. Steinman. 2004. Internet usage by low-literacy adults seeking health information: An observational analysis. *J Med Internet Res* 6(3):e25. www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1550604 (accessed November 29, 2011).

Downloaded by [RMIT University] at 23:36 01 March 2015

Downloaded by [RMIT University] at 23:36 01 March 2015

Bojko, A. 2005. Eye tracking in user experience testing: How to make the most of it. In *Proceedings of the Usability Professionals' Association Annual Meeting*, 1–9. (Montreal, Canada), Bloomingdale, IL: The Usability Professionals' Association.

Bojko, A. 2009. Informative or misleading? Heatmaps deconstructed. In *Human-Computer Interaction*, ed. J. Jacko, 30–9. Heidelberg, Germany: Springer-Verlag.

Boren, M., and J. Ramey. 2000. Thinking aloud: Reconciling theory and practice. *IEEE Trans Prof Commun* 43(3):261–78.

Branaghan, R. 1997. Ten tips for selecting usability test participants. *Common Ground* 7:3–6.

Burmeister, O. K. 2001. Usability testing: Revisiting informed consent procedures for testing Internet sites. In *Second Australian Institute Conference on Computer Ethics*, 3–9. (Sydney, Australia), Darlinghurst, Australia: The Australian Computer Society, Inc.

Castillo, J. C., H. R. Hartson, and D. Hix. 1998. Remote usability evaluation: Can users report their own critical incidents? In *Proceedings of Human Factors in Computing Systems*, 253–354. (Los Angeles, CA), New York: The Association for Computing Machinery.

Catani, M. B. 2003. Observation methodologies for usability tests of handheld devices. In *Proceedings of The Usability Professionals' Association Annual Meeting*, 1–6. (Scottsdale, AZ), Bloomingdale, IL: The Usability Professionals' Association.

Chisnell, D., A. Lee, and J. Redish. 2005. *Recruiting and Working with Older Participants*, *American Association of Retired Persons*. www.aarp.org/olderwiserwired/oww-features/Articles/a2004-03-03-recruiting-participants.html (accessed October 13, 2005).

Coyne, K. P. 2005. Conducting simple usability studies with users with disabilities. In *Proceedings of HCI International*, 890–3. (Las Vegas, NV), Mahwah, NJ: Lawrence Erlbaum Associates.

Desurvire, H. W., J. M. Kondziela, and M. E. Atwood. 1992. What is gained and lost when using evaluation methods other than empirical testing. In *People and Computers VII*, ed. A. Monk, D. Diaper, and M. D. Harrison, 89–102. Cambridge, MA: Cambridge University Press.

Douglass, R., and K. Hylton. 2010. Get it RITE. *User Exp* 9:12–3.

Duh, H. B.-L., G. C. B. Tan, and V. H. Chen. 2006. Usability evaluation for mobile device: A comparison of laboratory and field tests. In *Proceedings of the 8th Conference on Human-Computer Interaction with Mobile Devices and Services*, 181–6. (Helsinki, Finland), New York: The Association for Computing Machinery.

Dumas, J. 1998. Usability testing methods: Using test participants as their own controls. *Common Ground* 8:3–5.

Dumas, J. 2001. Usability testing methods: Think aloud protocols. In *Design by People for People: Essays on Usability*, ed. R. Branaghan, 119–30. Chicago, IL: Usability Professionals' Association.

Dumas, J. 2003. Usability evaluation from your desktop. *Assn Inf Syst (AIS) SIGCHI Newsletter* 2(2):7–8.

Dumas, J. 2007. The great leap forward: The birth of the usability profession (1988–1993). *J Usability Stud* 2(2):54–60.

Dumas, J., and J. Fox. 2007. Usability testing: Current practice and future directions. In *The Human-Computer Interaction Handbook*, ed. J. Jacko and A. Sears 2nd ed., 1129–49. Mahwah, NJ: Lawrence Erlbaum, Associates.

Dumas, J., and B. Loring. 2008. *Moderating Usability Tests: Principles and Practices for Interacting*. San Francisco, CA: Morgan Kaufman.

Dumas, J., and G. Redish. 1993. *A Practical Guide to Usability Testing (1st ed.)*. London: Intellect Books.

Dumas, J., and G. Redish. 1999. *A Practical Guide to Usability Testing* (Rev. ed.). London: Intellect Books.

Ebling, M., and B. John. 2000. On the contributions of different empirical data in usability testing. In *Proceedings of Designing Interactive Systems*, 289–96. (Brooklyn, NY), New York: The Association for Computing Machinery.

Eger, N., L. J. Ball, R. Stevens, and J. Dodd. 2007. Cueing retrospective verbal reports in usability testing through eye-movement replay. In *People and Computers XXI—HCI … but not as we know it: Proceedings of HCI 2007*, ed. L. J. Ball, M. A. Sasse, C. Sas, T. C. Ormerod, A. Dix, P. Bagnall, and T. McEwan. Swindon: The British Computer Society.

Ellis, K., M. Quigley, and M. Power. 2008. Experiences in ethical usability testing with children. *J Inf Technol Res* 1(3):1–13.

Ericsson, K. A., and H. A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data*. Cambridge, MA: MIT Press.

Evers, V. 2004. Cross-cultural applicability of user evaluation methods. A case study amongst Japanese, North-American, English and Dutch users. In *Proceedings of Human Factors in Computing Systems*, 740–1. (New Orleans, LA), New York: The Association for Computing Machinery.

Faulkner, L. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behav Res Methods Instrum Comput* 35(3):379–83.

Frishberg, N. 2010. Agile and UX. *User Exp* 9:4.

Frøkjær, E., M. Hertzum, and K. Hornbæk. 2000. Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of Human Factors in Computing Systems*, 345–52. (Fort Lauderdale, FL), New York: The Association for Computing Machinery.

Frøkjær, E., and K. Hornbæk. 2005. Cooperative usability testing: Complementing usability tests with user-supported interpretation sessions. In *Proceedings of Human Factors in Computing Systems*, 1383–6. (Denver, CO), New York: The Association for Computing Machinery.

Goldberg, J. H., and A. M. Wichansky. 2002. Eye tracking in usability evaluation: A practitioner's guide. In *The Mind's Eyes: Cognitive and Applied Aspects of Eye Movements*, ed. J. Hyönä, R. Radach, and H. Deubel, 493–516. Oxford: Elsevier Science.

Gray, W., and M. Salzman. 1998. Damaged merchandise? A review of experiments that compare usability methods. *Hum Comput Interact* 13:203–335.

Grossnickle, M. M. 2004. How many users with disabilities should you include when conducting a usability test for accessibility? Idea Market presented at *The Usability Professionals' Association Annual Meeting* www.upassoc.org/usability_resources/conference/2004/im_martinson.html (accessed September 13, 2005).

Hammersley, M., and P. Atkinson. 1995. *Ethnography: Principles in Practice*. London: Routledge.

Hancock, P., A. Pepe, and L. Murphy. 2005. Hedonomics: The power of positive and pleasurable ergonomics. *Ergon Des* 13(1):8–14.

Hanna, L., K. Risden, and K. J. Alexander. 1997. Guidelines for usability testing with children. *Interactions* 4:9–14.

Hartson, H. R., J. C. Castillo, J. Kelso, and W. Neale. 1996. Remote evaluation: The network as an extension of the usability laboratory. In *Proceedings of Human Factors in Computing Systems*, 228–35. (Vancouver, Canada), New York: The Association for Computing Machinery.

Henry, S. L. 2007. Just ask. www.uiaccess.com/accessucd/index.html (accessed February 12, 2010).

Hertzum, M., and N. E. Jacobsen. 2001. The evaluator effect: A chilling fact about usability evaluation methods. *Int J Hum Comput Interact* 13(4):421–43.

Hornbæk, K., and E. Law. 2007. Meta-analysis of correlations among usability measures. In *Proceedings of Human Factors in Computing Systems*, 617–26. (San Jose, CA), New York: The Association for Computing Machinery.

House, E. 1990. An ethics of qualitative field studies. In *The Paradigm Dialog*, ed. E. Gaba, 158–201. Newbury Park, CA: Sage.

Information Technology Technical Assistance and Training Center (ITTATC), Georgia Institute of Technology. 2004. *Planning Usability Testing for Accessibility*. www.ittatc.org/technical/access-ucd/ut_plan.php (accessed September 13, 2005).

Jacobsen, N., M. Hertzum, and B. E. John. 1998. The evaluator effect in usability studies: Problem detection and severity judgments. In *Proceedings of the Human Factors and Ergonomics Society, 42nd Annual Meeting*, 1336–40. (Chicago, IL), Santa Monica, CA: The Human Factors and Ergonomics Society.

Jordan, P. 2002. The personalities of products. In *Pleasure with Products*, ed. W. Green and P. Jordan, 19–48. London: Taylor & Francis.

Kaikkaner, A., A. Kekalainen, M. Canker, T. Kalliot, and A. Kankainen. 2005. Usability testing of mobile applications: A comparison between laboratory and field studies. *J Usability Stud* 1:4–16.

Karat, J. 2003. Beyond task completion: Evaluation of affective components of use. In *The Human-Computer Interaction Handbook*, ed. J. Jacko and A. Sears, 1152–64. Mahwah, NJ: Lawrence Erlbaum, Assoc.

Kjeldskov, J., and J. Stage. 2004. New techniques for usability evaluation of mobile systems. *Int J Hum Comput Stud* 60(5–6):599–620.

Krahmer, E., and N. Ummelen. 2004. Thinking about thinking aloud: A comparison of two verbal protocols for usability testing. *IEEE Trans Prof Commun* 47(2):105–17.

Krug, S. 2010. *Rocket Surgery Made Easy*. Berkeley, CA: New Riders.

Law, C., and G. Vanderheiden. 2000. Reducing sample sizes when user testing with people who have and who are simulating disabilities: Experiences with blindness and public information kiosks. In *Proceedings of the IEA 2000/HFES 2000 Congress*, 26, 157–60. (San Diego, CA), Santa Monica, CA: The Human Factors and Ergonomics Society.

Lepistö, A., and S. Ovaska. 2004. Usability evaluation involving participants with cognitive disabilities. In *Proceedings of NordiCHI*, 305–8. (Tempere, Finland), New York: The Association for Computing Machinery.

Lesaigle, E. M., and D. W. Biers. 2000. Effect of type of information on real-time usability evaluation: Implications for remote usability testing. In *Proceedings of the IEA 2000/HFES 2000 Congress*, 37, 585–8. (San Diego, CA), Santa Monica, CA: The Human Factors and Ergonomics Society.

Lewis, J. 1994. Sample size for usability studies: Additional considerations. *Hum Fact* 36:368–78.

Lewis, J. 2001. Evaluation of procedures of adjusting problem discovery rates estimates from small samples. *Int J Hum Comput Interact* 71(1):57–78.

Lindgaard, G., and J. Chattratichart. 2007. Usability testing: What have we overlooked? In *Proceedings of Human Factors in Computing Systems*, 1415–24. (San Jose, CA), New York: The Association for Computing Machinery.

Lu, C., T. Rauch, and L. Miller. 2010. Agile teams: Best practices for agile development. *User Exp* 9:6–10.

Luef, B., and W. Cunningham. 2001. *The Wiki Way: Quick Collaboration on the Web.* Reading, MA: Addison-Wesley, Inc.

Macklin, R. 1982. The problem of adequate disclosure in social science research. In *Ethical Issues in Social Science Research*, ed. T. Beauchamp, R. Faden, R. Wallace, and L. Walters, 193–214. Baltimore, MD: Johns Hopkins.

Medlock, M., D. Wixon, M. McGee, and D. Welsh. 2005. The rapid iterative test and evaluation method: Better products in less time. In *Cost-Justifying Usability: An Update for the Information Age*, 489–517. New York: Morgan Kaufman Publishers.

Millett, L. I., B. Friedman, and E. Felten. 2001. Cookies and web browser design: Toward realizing informed consent online. In *Proceedings of Human Factors in Computing Systems*, 46–52. (Seattle, WA), New York: The Association for Computing Machinery.

Molich, R., N. Bevan, I. Curson, S. Butler, E. Kindlund, D. Miller, and J. Kirakowski. 1998. Comparative evaluation of usability tests. In *Proceedings of the Usability Professionals' Association Annual Meeting.* Bloomingdale, IL: The Usability Professionals' Association.

Molich, R., and J. Dumas. 2008. Comparative usability evaluation (CUE-4). *Behav Inf Technol* 27(3):263–81.

Molich, R., R. Meghan, K. Ede, and B. Karyukin. 2004. Comparative usability evaluation. *Behav Inf Technol* 23:65–74.

Murphy, E., R. Dingwall, D. Greatbatch, S. Parker, and P. Watson. 1998. Qualitative research methods in health technology assessment: A review of the literature. *Health Technol Assess* 2(16):1–272.

Murphy, L., K. Stanney, and P. Hancock. 2003. The effect of affect: The hedonic evaluation of human-computer interaction. In *Proceedings of the Human Factors and Ergonimics Society 47th Annual Meeting*. 764–7. Santa Monica, CA: The Human Factors and Ergonomics Society.

Nielsen, J. 1992. Finding usability problems through heuristic evaluation. In *Proceedings of Human Factors in Computing Systems*, 373–80. (Monterey, CA), New York: The Association for Computing Machinery.

Nielsen, J. 1996. *International Usability Testing.* www.useit.com/papers/international_usetest.html (accessed September 13, 2005).

Nielsen. 2000. *Why you Only Need to Test with 5 Users*. www.useit.com/alertbox/20000319.html (accessed February 22, 2009).

Nielsen, J., T. Clemmensen, and C. Yssing. 2002. Getting access to what goes on in people's heads: Reflections on the think-aloud technique. In *Proceedings of NordiCHI*, 101–10. (Aarhus, Denmark), New York: The Association for Computing Machinery.

Norgaard, M., and K. Hornbaek. 2006. What do usability evaluators do in practice? An explorative study of think-aloud testing. In *Proceedings of Designing Interactive Systems*, 209–18. (University Park, PA), New York: The Association for Computing Machinery.

Olmsted-Hawala, E., S. Hawala, E. Murphy, and K. Ashenfelter. 2010. Think-aloud protocols: A comparison of three think-aloud protocols for use in testing data-dissemination web sites

Downloaded by [RMIT University] at 23:36 01 March 2015

for usability. In *Proceedings of Human Factors in Computing Systems*, 2381–90. (Atlanta, GA), New York: The Association for Computing Machinery.

Pagulayan, R., K. Keeker, D. Wixon, R. Romero, and T. Fuller. 2003. User-centered design in games. In *The Human-Computer Interaction Handbook*, ed. J. Jacko and A. Sears, 883–906. Mahwah, NJ: Lawrence Erlbaum, Assoc.

Patel, M., and C. A. Paulsen. 2002. Strategies for recruiting children for usability tests. In *Proceedings of the Usability Professionals' Association Annual Meeting*, 1–4. (Orlando, FL), Bloomingdale, IL: The Usability Professionals' Association.

Petrie, H., F. Hamilton, N. King, and P. Pavan. 2006. Remote usability evaluations with disabled people. In *Proceedings of Human Factors in Computing Systems*, 1133–41. (Montreal, Canada), New York: The Association for Computing Machinery.

Poole, A., and L. J. Ball. 2005. Eye tracking in human-computer interaction and usability research: Current status and future prospects. In *Encyclopedia of a Human-Computer Interaction*, ed. C. Ghaoui, 211–19. Hershey, PA: Idea Group.

Pruitt, J., and T. Adlin. 2005. *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. San Francisco, CA: Morgan Kaufman.

Quesenbery, W. 2004. "Balancing the 5Es: Usability." *Cutter IT J* 17(2):4–11.

Quesenbery, W. 2005. The five dimensions of usability. In *Content and Complexity: Information Design in Technical Communication*, ed. M. Albers, B. Mazur, 81–102. Mahwah, NJ: Lawrence Erlbaum & Associates.

Redish, J. 2007. Expanding usability testing to evaluate complex systems. *J Usability Stud* 2:102–11.

Redish, J. C., and J. Scholtz. 2007. Evaluating complex information systems for domain experts. Paper presented at *HCI and Information Design to Communicate Complex Information*, 1–23. Memphis, TN: University of Memphis.

Rosenbaum, S., J. Rohn, and J. Humburg. 2000. A toolkit for strategic usability: Results from workshops, panels, and surveys. In *Proceedings of Human Factors in Computing Systems*, 337–44. (The Hague, Netherlands), New York: The Association for Computing Machinery.

Rubin, J. 1994. *Handbook of Usability Testing.* New York: John Wiley & Sons, Inc.

Sauro, J., and J. Dumas. 2009. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of Human Factors in Computing Systems*, 1599–608. (Boston, MA), New York: The Association for Computing Machinery.

Sauro, J., and J. Lewis. 2009. Correlations among prototypical usability metrics: Evidence for the construct of usability. In *Proceedings of Human Factors in Computing Systems*, 1609–18. (Boston, MA), New York: The Association for Computing Machinery.

Sawyer, P., A. Flanders, and D. Wixon. 1996. Making a difference—the impact of inspection. In *Proceedings of Human Factors in Computing Systems*, 378–82. (Vancouver, British Columbia, Canada), New York: The Association for Computing Machinery.

Schnipke, S. K., and M. W. Todd. 2000. Trials and tribulations of using an eye-tracking system. In *Proceedings of Human Factors in Computing Systems*, 185–6. (The Hague, Netherlands), New York: The Association for Computing Machinery.

Schusteritsch, R., C. Y. Wei, and M. LaRosa. 2007. Towards the perfect infrastructure for usability testing on mobile devices. In *Proceedings of Human Factors in Computing Systems*, 1839–44. (San Jose, CA), New York: The Association for Computing Machinery.

Shanteau, J. 2001. What does it mean when experts disagree? In *Linking Expertise and Naturalistic Decision Making*, ed. E. Salas and G. Klein, 229–44. Mahwah, NJ: Lawrence Erlbaum Associates.

Sears, A. 1997. Heuristic walkthroughs: Finding the problems without the noise. *Int J Hum Comput Interact* 9:213–34.

Spolsky, J. 2000. *The Joel Test: 12 Steps to Better Code*. www.joelonsoftware.com/articles/fog0000000043.html (accessed November 29, 2011).

Strain, P., A. D. Shaikh, and R. Boardman. 2007. Thinking but not seeing: Think-aloud for non-sighted users. In *Proceedings of Human Factors in Computing Systems*, 1851–6. (San Jose, CA), New York: The Association for Computing Machinery.

Swierenga, S. J., and T. Guy. 2003. Session logistics for usability testing of users with disabilities. In *Proceedings of the Usability Professionals' Association Annual Meeting*, 1–6. (Scottsdale, AZ), Bloomingdale, IL: The Usability Professionals' Association.

Teague, R., and H. Whitney. 2002. What's love got to do with it? *User Exp* 1:6–13.

Tedesco, D., M. McNulty, and T. Tullis. 2005. Usability testing with older adults. In *Proceedings of the Usability Professionals' Association Annual Meeting*, 1–8. (Montreal, Canada), Bloomingdale, IL: The Usability Professionals' Association.

Tedesco, D., and T. Tullis. 2006. A Comparison of methods for eliciting post-task subjective ratings in usability testing. In *Proceedings of the Usability Professionals Association Annual Meeting*, 1–9. (Broomfield, Colorado), Bloomingdale, IL: The Usability Professionals' Association.

Thompson, K., E. Rozanski, and A. Haake. 2004. Here, there, anywhere: Remote usability testing that works. In *Proceedings of SIGITE*, 132–7. (Salt Lake City, UT), New York: The Association for Computing Machinery.

Tullis, T., S. Flieschman, M. McNulty, C. Cianchette, and M. Bergel. 2002. An empirical comparison of lab and remote usability testing of web sites. In *Proceedings of the Usability Professionals' Association Annual Meeting*, 1–5. (Orlando, FL), Bloomingdale, IL: The Usability Professionals' Association.

Tullis, T., and J. Stetson. 2004. A comparison of questionnaires for assessing website usability. In *Proceedings of the Usability Professionals' Association Annual Meeting*, 1–12. (Minneapolis, MN), Bloomingdale, IL: The Usability Professionals' Association.

Turner, C., J. R. Lewis, and J. Nielsen. 2006. Determining usability test sample size. In *International Encyclopedia of Ergonomics and Human Factors*, ed. W. Karwowski, 3084–8. Boca Raton, FL: CRC Press.

U.S. Department of Health and Human Services. 2008. *Office for Human Research Protections (OHRP): OHRP Informed Consent Frequently Asked Questions*. www.answers.hhs.gov/ohrp/categories/1566 (accessed February 25, 2010).

van den Haak, M. J., M. D. T. de Jong, and P. J. Schellens. 2003. Retrospective vs. concurrent think aloud protocols: Testing the usability of an online library catalogue. *Behav Inf Technol* 22(5):339–51.

Downloaded by [RMIT University] at 23:36 01 March 2015

Vatrapu, R., and M. A. Pérez-Quiñones. 2004. *Culture and International Usability Testing: The Effects of Culture in Structured Interviews.* Technical Report cs.HC/0405045. Computing Research Repository (CoRR). http://arxiv.org/pdf/cs/0405045v1 (accessed October 4, 2005).

Virzi, R. A. 1990. Streamlining the design process: Running fewer subjects. In *Proceedings of the Human Factors Society, 34th Annual Meeting*, 291–4. (Orlando, FL), Santa Monica, CA: The Human Factors and Ergonomics Society.

Virzi, R. A. 1992. Refining the test phase of usability evaluation: How many subjects is enough? *Hum Fact* 34:457–68.

Virzi, R. A., J. F. Sorce, and L. B. Herbert. 1993. A comparison of three usability evaluation methods: Heuristic, think-aloud, and performance testing. In *Proceedings of the Human Factors and Ergonomics Society, 37th Annual Meeting*, 309–13. (Seattle, WA), Santa Monica, CA: The Human Factors and Ergonomics Society.

West, R., and K. R. Lehman. 2006. Automated summative usability studies: An empirical evaluation. In *Proceedings of Human Factors in Computing Systems*, 631–9. (Montreal, Canada), New York: The Association for Computing Machinery.

Wilson, C. E., and K. P. Coyne. 2001. Tracking usability issues: To bug or not to bug? *Interactions* 8:15–9.

Wolfson, C. A., R. W. Bailey, J. Nall, and S. Koyani. 2008. Contextual card sorting (or FirstClick testing): A new methodology for validating information architectures. In *Proceedings of the Usability Professionals' Association Annual Meeting*, 1–6. (Baltimore, MD), Bloomingdale, IL: The Usability Professionals' Association.

Downloaded by [RMIT University] at 23:36 01 March 2015

Downloaded by [RMIT University] at 23:36 01 March 2015