CHAPTER

Quantifying User Research

2

WHAT IS USER RESEARCH?

For a topic with only two words, "*user research*" implies different things to different people. Regarding "user" in user research, Edward Tufte (Bisbort, 1999) famously said: "Only two industries refer to their customers as 'users': computer design and drug dealing."

This book focuses on the first of those two types of customers. This user can be a paying customer, internal employee, physician, call-center operator, automobile driver, cell phone owner, or any person attempting to accomplish some goal—typically with some type of software, website, or machine.

The "research" in user research is both broad and nebulous—a reflection of the amalgamation of methods and professionals that fall under its auspices. Schumacher (2010, p. 6) offers one definition:

> *User research is the systematic study of the goals, needs, and capabilities of users so as to specify design, construction, or improvement of tools to benefit how users work and live.*

Our concern is less with defining the term and what it covers than with quantifying the behavior of users, which is in the purview of usability professionals, designers, product managers, marketers, and developers.

DATA FROM USER RESEARCH

Although the term *user research* may eventually fall out of favor, the data that come from user research won't. Throughout this book we will use examples from usability testing, customer surveys, A/B testing, and site visits, with an emphasis on usability testing. There are three reasons for our emphasis on usability testing data:

1. Usability testing remains a central way of determining whether users are accomplishing their goals.
2. Both authors have conducted and written extensively about usability testing.
3. Usability testing uses many of the same metrics as other user research techniques (e.g., completion rates can be found just about everywhere).

USABILITY TESTING

Usability has an international standard definition in ISO 9241 pt. 11 (ISO, 1998), which defined *usability* as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use. Although there are no specific guidelines

Quantifying the User Experience. DOI: 10.1016/B978-0-12-384968-7.00002-3

on how to measure effectiveness, efficiency, and satisfaction, a large survey of almost 100 summative usability tests (Sauro and Lewis, 2009) reveals what practitioners typically collect. Most tests contain some combination of completion rates, errors, task times, task-level satisfaction, test-level satisfaction, help access, and lists of usability problems (typically including frequency and severity).

There are generally two types of usability tests: finding and fixing usability problems (formative tests) and describing the usability of an application using metrics (summative tests). The terms *formative* and *summative* come from education (Scriven, 1967) where they are used in a similar way to describe tests of student learning (formative—providing immediate feedback to improve learning, versus summative—evaluating what was learned).

The bulk of usability testing is formative. It is often a small-sample qualitative activity where the data take the form of problem descriptions and design recommendations. Just because the goal is to find and fix as many problems as you can does not mean there is no opportunity for quantification. You can quantify the problems in terms of frequency and severity, track which users encountered which problems, measure how long it took them to complete tasks, and determine whether they completed the tasks successfully.

There are typically two types of summative tests: benchmark and comparative. The goal of a benchmark usability test is to describe how usable an application is relative to a set of benchmark goals. Benchmark tests provide input on what to fix in an interface and also provide an essential baseline for the comparison of postdesign changes.

A comparative usability test, as the name suggests, involves more than one application. This can be a comparison of a current with a prior version of a product or comparison of competing products. In comparative tests, the same users can attempt tasks on all products (within-subjects design) or different sets of users can work with each product (between-subjects design).

Sample Sizes

There is an incorrect perception that sample sizes must be large (typically above 30) to use statistics and interpret quantitative data. We discuss sample sizes extensively in Chapters 6 and 7, and throughout this book show how to reach valid statistical conclusions with sample sizes less than 10. Don't let the size of your sample (even if you have as few as 2–5 users) preclude you from using statistics to quantify your data and inform your design decisions.

Representativeness and Randomness

Somewhat related to the issue of sample sizes is that of the makeup of the sample. Often the concern with a small sample size is that the sample isn't "representative" of the parent population. Sample size and representativeness are actually different concepts. You can have a sample size of 5 that is representative of the population and you can have a sample size of 1,000 that is not representative. One of the more famous examples of this distinction comes from the 1936 *Literary Digest* Presidential Poll. The magazine polled its readers on who they intended to vote for and received 2.4 million responses but incorrectly predicted the winner of the presidential election. The problem was not one of sample size but of representativeness. The people who responded tended to be individuals with higher incomes and education levels—not representative of the ultimate voters (see *http://en.wikipedia.org/wiki/The_Literary_Digest*).

The most important thing in user research, whether the data are qualitative or quantitative, is that the sample of users you measure represents the population about which you intend to make statements. Otherwise, you have no logical basis for generalizing your results from the sample to the population. No amount of statistical manipulation can correct for making inferences about one population if you observe a sample from a different population. Taken to the extreme, it doesn't matter how many men are in your sample if you want to make statements about female education levels or salaries. If you want to gain insight into how to improve the design of snowshoes, it's better to have a sample of 5 Arctic explorers than a sample of 1,000 surfers. In practice, this means if you intend to draw conclusions about different types of users (e.g., new versus experienced, older versus younger) you should plan on having all groups represented in your sample.

One reason for the confusion between sample size and representativeness is that if your population is composed of, say, 10 distinct groups and you have a sample of 5, then there aren't enough people in the sample to have a representative from all 10 groups. You would deal with this by developing a sampling plan that ensures drawing a representative sample from every group that you need to study—a method known as *stratified sampling*. For example, consider sampling from different groups if you have reason to believe:

- There are potential and important differences among groups on key measures (Dickens, 1987).
- There are potential interactions as a function of a group (Aykin and Aykin, 1991).
- The variability of key measures differs as a function of a group.
- The cost of sampling differs significantly from group to group.

Gordon and Langmaid (1988) recommended the following approach to defining groups:

1. Write down all the important variables.
2. If necessary, prioritize the list.
3. Design an ideal sample.
4. Apply common sense to combine groups.

For example, suppose you start with 24 groups, based on the combination of six demographic locations, two levels of experience, and the two levels of gender. You might plan to (1) include equal numbers of males and females over and under 40 years of age in each group, (2) have separate groups for novice and experienced users, and (3) drop intermediate users from the test. The resulting plan requires sampling for 2 groups. A plan that did not combine genders and ages would require sampling 8 groups.

Ideally, your sample is also selected randomly from the parent population. In practice this can be very difficult. Unless you force your users to participate in a study you will likely suffer from at least some form of nonrandomness. In usability studies and surveys, people decide to participate and this group can have different characteristics than people who choose not to participate. This problem isn't unique to user research. Even in clinical trials in which life and death decisions are made about drugs and medical procedures, people have to participate or have a condition (like cancer or diabetes). Many of the principles of human behavior that fill psychology textbooks disproportionally come from college undergrads—a potential problem of both randomness and representativeness.

It's always important to understand the biases in your data and how that limits your conclusions. In applied research we are constrained by budgets and user participation, but products still must

ship, so we make the best decisions we can given the data we are able to collect. Where possible seek to minimize systematic bias in your sample but remember that representativeness is more important than randomness. In other words, you'll make better decisions if you have a less-than-perfectly random sample from the right population than if you have a perfectly random sample from the wrong population.

Data Collection

Usability data can be collected in a traditional lab-based moderated session where a moderator observes and interacts with users as they attempt tasks. Such test setups can be expensive and time consuming and require collocation of users and observers (which can prohibit international testing). These types of studies often require the use of small-sample statistical procedures because the cost of each sample is high.

More recently, remote moderated and unmoderated sessions have become popular. In moderated remote sessions, users attempt tasks on their own computer and software from their location while a moderator observes and records their behavior using screen-sharing software. In unmoderated remote sessions, users attempt tasks (usually on websites), while software records their clicks, page views, and time. For an extensive discussion of remote methods, see *Beyond the Usability Lab* (Albert et al., 2010).

For a comprehensive discussion of usability testing, see the chapter "Usability Testing" in the *Handbook of Human Factors and Ergonomics* (Lewis, 2012). For practical tips on collecting metrics in usability tests, see *A Practical Guide to Measuring Usability* (Sauro, 2010) and *Measuring the User Experience* (Tullis and Albert, 2008).

In our experience, although the reasons for human behavior are difficult to quantify, the outcome of the behavior is easy to observe, measure, and manage. Following are descriptions of the more common metrics collected in user research, inside and outside of usability tests. We will use these terms extensively throughout the book.

Completion Rates

Completion rates, also called success rates, are the most fundamental of usability metrics (Nielsen, 2001). They are typically collected as a binary measure of task success (coded as 1) or task failure (coded as 0). You report completion rates on a task by dividing the number of users who successfully complete the task by the total number who attempted it. For example, if 8 out of 10 users complete a task successfully, the completion rate is 0.8 and usually reported as 80%. You can also subtract the completion rate from 100% and report a failure rate of 20%.

It is possible to define criteria for partial task success, but we prefer the simpler binary measure because it lends itself better for statistical analysis. When we refer to completion rates in this book, we will be referring to binary completion rates.

The other nice thing about a binary rate is that they are used throughout the scientific and statistics literature. Essentially, the presence or absence of anything can be coded as 1's and 0's and then reported as a proportion or percentage. Whether this is the number of users completing tasks on software, patients cured from an ailment, number of fish recaptured in a lake, or customers purchasing a product, they can all be treated as binary rates.

Usability Problems

If a user encounters a problem while attempting a task and it can be associated with the interface, it's a user interface problem (UI problem). UI problems, typically organized into lists, have names, a description, and often a severity rating that takes into account the observed problem frequency and its impact on the user.

The usual method for measuring the frequency of occurrence of a problem is to divide the number of occurrences within participants by the number of participants. A common technique (Rubin, 1994; Dumas and Redish, 1999) for assessing the impact of a problem is to assign impact scores according to whether the problem (1) prevents task completion, (2) causes a significant delay or frustration, (3) has a relatively minor effect on task performance, or (4) is a suggestion.

When considering multiple types of data in a prioritization process, it is necessary to combine the data in some way. One approach is to combine the data arithmetically. Rubin (1994) described a procedure for combining four levels of impact (using the criteria previously described with 4 assigned to the most serious level) with four levels of frequency (4: frequency $\geq$ 90%; 3: 51–89%; 2: 11–50%; 1: $\leq$ 10%) by adding the scores. For example, if a problem had an observed frequency of occurrence of 80% and had a minor effect on performance, its priority would be 5 (a frequency rating of 3 plus an impact rating of 2). With this approach, priority scores can range from a low of 2 to a high of 8.

A similar strategy is to multiply the observed percentage frequency of occurrence by the impact score (Lewis, 2012). The range of priorities depends on the values assigned to each impact level. Assigning 10 to the most serious impact level leads to a maximum priority (severity) score of 1,000 (which can optionally be divided by 10 to create a scale that ranges from 1 to 100). Appropriate values for the remaining three impact categories depend on practitioner judgment, but a reasonable set is 5, 3, and 1. Using those values, the problem with an observed frequency of occurrence of 80% and a minor effect on performance would have a priority of 24 (80 $\times$ 3/10).

From an analytical perspective, a useful way to organize UI problems is to associate them with the users who encountered them, as shown in Table 2.1.

Knowing the probability with which users will encounter a problem at each phase of development can become a key metric for measuring usability activity impact and return on investment (ROI). Knowing which user encountered which problem allows you to better estimate sample sizes, problem discovery rates, and the number of undiscovered problems (as described in detail in Chapter 7).

Table 2.1 Example of a UI Problem Matrix

	User 1	User 2	User 3	User 4	User 5	User 6	Total	Proportion
Problem 1	X	X			X	X	4	0.67
Problem 2	X						1	0.167
Problem 3	X	X	X	X	X	X	6	1
Problem 4				X	X		2	0.33
Problem 5					X		1	0.167
Total	3	2	1	2	4	2	**14**	**p = 0.47**

Note: The X's represent users who encountered a problem. For example, user 4 encountered problems 3 and 4.

Task Time

Task time is how long a user spends on an activity. It is most often the amount of time it takes users to successfully complete a predefined task scenario, but it can be total time on a web page or call length. It can be measured in milliseconds, seconds, minutes, hours, days, or years, and is typically reported as an average (see Chapter 3 for a discussion on handling task-time data). There are several ways of measuring and analyzing task duration:

1. *Task completion time*: Time of users who completed the task successfully.
2. *Time until failure*: Time on task until users give up or complete the task incorrectly.
3. *Total time on task*: The total duration of time users spend on a task.

Errors

Errors are any unintended action, slip, mistake, or omission a user makes while attempting a task. Error counts can go from 0 (no errors) to technically infinity (although it is rare to record more than 20 or so in one task in a usability test). Errors provide excellent diagnostic information on why users are failing tasks and, where possible, are mapped to UI problems. Errors can also be analyzed as binary measures: the user either encountered an error (1 = yes) or did not (0 = no).

Satisfaction Ratings

Questionnaires that measure the perception of the ease of use of a system can be completed immediately after a task (post-task questionnaires), at the end of a usability session (post-test questionnaires), or outside of a usability test. Although you can write your own questions for assessing perceived ease of use, your results will likely be more reliable if you use one of the currently available standardized questionnaires (Sauro and Lewis, 2009). See Chapter 8 for a detailed discussion of standardized usability questionnaires.

Combined Scores

Although usability metrics significantly correlate (Sauro and Lewis, 2009), they don't correlate strongly enough that one metric can replace another. In general, users who complete more tasks tend to rate tasks as easier and to complete them more quickly. Some users, however, fail tasks and still rate them as being easy, or others complete tasks quickly and report finding them difficult. Collecting multiple metrics in a usability test is advantageous because this provides a better picture of the overall user experience than any single measure can. However, analyzing and reporting on multiple metrics can be cumbersome, so it can be easier to combine metrics into a single score. A combined usability metric can be treated just like any other metric and can be used advantageously as a component of executive dashboards or for determining statistical significance between products (see Chapter 5). For more information on combining usability metrics into single scores, see Sauro and Kindlund (2005), Sauro and Lewis (2009), and the "Can You Combine Usability Metrics into Single Scores?" section in Chapter 9.

A/B TESTING

A/B testing, also called split-half testing, is a popular method for comparing alternate designs on web pages. In this type of testing, popularized by Amazon, users randomly work with one of two deployed design alternatives. The difference in design can be as subtle as different words on a button or a different product image, or can involve entirely different page layouts and product information.

Clicks, Page Views, and Conversion Rates

For websites and web applications, it is typical practice to automatically collect clicks and page views, and in many cases these are the only data you have access to without conducting your own study. Both these measures are useful for determining conversion rates, purchase rates, or feature usage, and are used extensively in A/B testing, typically analyzed like completion rates.

To determine which design is superior, you count the number of users who were presented with each design and the number of users who clicked through. For example, if 1,000 users experienced Design A and 20 clicked on "Sign-Up," and 1,050 users saw Design B and 48 clicked on "Sign-Up," the conversion rates are 2% and 4.5%, respectively. To learn how to determine if there is a statistical difference between designs, see Chapter 5.

SURVEY DATA

Surveys are one of the easiest ways to collect attitudinal data from customers. Surveys typically contain some combination of open-ended comments, binary yes/no responses, and Likert-type rating scale data.

Rating Scales

Rating scale items are characterized by closed-ended response options. Typically, respondents are asked to agree or disagree to a statement (often referred to as Likert-type items). For numerical analysis, the classic five-choice Likert response options can be converted into numbers from 1 to 5 (as shown in Table 2.2).

Once you've converted the responses to numbers you can compute the mean and standard deviation and generate confidence intervals (see Chapter 3) or compare responses to different products (see Chapter 5). See Chapter 8 for a detailed discussion of questionnaires and rating scales specific to usability, and the "Is It Okay to Average Data from Multipoint Scales?" section in Chapter 9 for a discussion of the arguments for and against computing means and conducting standard statistical tests with this type of data.

Table 2.2 Mapping of the Five Classic Likert Response Options to Numbers

***This* →**	**Strongly Disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
Becomes This →	1	2	3	4	5

Net Promoter Scores®

Even though questions about customer loyalty and future purchasing behavior have been around for a long time, a recent innovation is the net promoter question and scoring method used by many companies and in some usability tests (Reichheld, 2003, 2006). The popular net promoter score (NPS) is based on a single question about customer loyalty: How likely is it that you'll recommend this product to a friend or colleague? The response options range from 0 to 10 and are grouped into three segments:

Promoters: Responses from 9 to 10
Passives: Responses from 7 to 8
Detractors: Responses from 0 to 6

By subtracting the percentage of detractor responses from the percentage of promoter responses you get the net promoter score, which ranges from −100% to 100%, with higher numbers indicating a better loyalty score (more promoters than detractors). Although the likelihood-to-recommend item can be analyzed just like any other rating scale item (using the mean and standard deviation), the segmentation scoring of the NPS requires slightly different statistical treatments (see Chapter 5).

Note: Net Promoter, NPS, and Net Promoter Score are trademarks of Satmetrix Systems, Inc., Bain & Company, and Fred Reichheld.

Comments and Open-ended Data

Analyzing and prioritizing comments is a common task for a user researcher. Open-ended comments take all sorts of forms, such as:

- Reasons why customers are promoters or detractors for a product.
- Customer insights from field studies.
- Product complaints to calls to customer service.
- Why a task was difficult to complete.

Just as usability problems can be counted, comments and most open-ended data can be turned into categories, quantified and subjected to statistical analysis (Sauro, 2011). You can then further analyze the data by generating a confidence interval to understand what percent of all users likely feel this way (see Chapter 3).

REQUIREMENTS GATHERING

Another key function of user research is to identify features and functions of a product. While it's rarely as easy as asking customers what they want, there are methods of analyzing customer behaviors that reveal unmet needs. As shown in Table 2.3, these behaviors can be observed at home or the workplace and then quantified in the same way as UI problems. Each behavior gets a name and description, and then you record which users exhibited the particular behavior in a grid like the one shown in the table.

You can easily report on the percentage of customers who exhibited a behavior and generate confidence intervals around the percentage in the same way you do for binary completion rates

Table 2.3 Example of a UI Behavior Matrix

	User 1	User 2	User 3
Behavior 1	X	X	
Behavior 2	X		
Behavior 3	X	X	X

(see Chapter 3). You can also apply statistical models of discovery to estimate required sample sizes, requirement discovery rates, and the number of undiscovered requirements (see Chapter 7).

KEY POINTS FROM THE CHAPTER

- User research is a broad term that encompasses many methodologies that generate quantifiable outcomes, including usability testing, surveys, questionnaires, and site visits.
- Usability testing is a central activity in user research and typically generates the metrics of completion rates, task times, errors, satisfaction data, and user interface problems.
- Binary completion rates are both a fundamental usability metric and a metric applied to all areas of scientific research.
- You can quantify data from small sample sizes and use statistics to draw conclusions.
- Even open-ended comments and problem descriptions can be categorized and quantified.

References

Albert, W., Tullis, T., Tedesco, D., 2010. Beyond the Usability Lab. Morgan Kaufmann, Boston.

Aykin, N.M., Aykin, T., 1991. Individual differences in human–computer interaction. Comput. Ind. Eng. 20, 373–379.

Bisbort, A., 1999. Escaping Flatland. Available at http://www.edwardtufte.com/tufte/advocate_1099 (accessed July 17, 2011).

Dickens, J., 1987. The fresh cream cakes market: The use of qualitative research as part of a consumer research programme. In: Bradley, U. (Ed.), Applied Marketing and Social Research. Wiley, New York, pp. 23–68.

Dumas, J., Redish, J.C., 1999. A Practical Guide to Usability Testing. Intellect, Portland.

Gordon, W., Langmaid, R., 1988. Qualitative Market Research: A Practitioner's and Buyer's Guide. Gower Publishing, Aldershot, England.

ISO, 1998. Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs)—Part 11: Guidance on Usability (ISO 9241-11:1998(E)). ISO, Geneva.

Lewis, J.R., 2012. Usability testing. In: Salvendy, G. (Ed.), Handbook of Human Factors and Ergonomics. Wiley, New York, pp. 1267–1312.

Nielsen, J., 2001. Success Rate: The Simplest Usability Metric. Available at http://www.useit.com/alertbox/20010218.html (accessed July 10, 2011).

Reichheld, F.F., 2003. The one number you need to grow. Harvard Bus. Rev. 81, 46–54.

Reichheld, F., 2006. The Ultimate Question: Driving Good Profits and True Growth. Harvard Business School Press, Boston.

Rubin, J., 1994. Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. Wiley, New York.

Sauro, J., Kindlund, E., 2005. A method to standardize usability metrics into a single score. In: Proceedings of CHI 2005. ACM, Portland, pp. 401–409.

Sauro, J., 2010. A Practical Guide to Measuring Usability. Measuring Usability LLC, Denver.

Sauro, J., 2011. How to Quantify Comments. Available at http://www.measuringusability.com/blog/quantify-comments.php (accessed July 15, 2011).

Sauro, J., Lewis, J.R., 2009. Correlations among prototypical usability metrics: Evidence for the construct of usability. In: Proceedings of CHI 2009. ACM, Boston, pp. 1609–1618.

Schumacher, R., 2010. The Handbook of Global User Research. Morgan Kaufmann, Boston.

Scriven, M., 1967. The methodology of evaluation. In: Tyler, R.W., Gagne, R.M., Scriven, M. (Eds.), Perspectives of Curriculum Evaluation. Rand McNally, Chicago, pp. 39–83.

Tullis, T., Albert, B., 2008. Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics. Morgan Kaufmann, Boston.