Second Edition

Measuring the User Experience

Collecting, Analyzing, and Presenting Usability Metrics

MORGAN KAUFMANN

TOM TULLIS • BILL ALBERT

Measuring
the User Experience

This page intentionally left blank

Measuring
the User Experience
Collecting, Analyzing, and Presenting Usability Metrics

Second Edition

Tom Tullis
Bill Albert

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK
OXFORD • PARIS • SAN DIEGO • SAN FRANCISCO
SINGAPORE • SYDNEY • TOKYO
Morgan Kaufmann is an imprint of Elsevier

Acquiring Editor: Meg Dunkerley
Editorial Project Manager: Heather Scherer
Project Manager: Priya Kumaraguruparan
Cover Designers: Greg Harris, Cheryl Tullis

Morgan Kaufmann is an imprint of Elsevier
225 Wyman Street, Waltham, MA, 02451, USA

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods or professional practices, may become necessary. Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information or methods described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility. To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability,negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
Tullis, Tom (Thomas)
Measuring the user experience : collecting, analyzing, and presenting usability metrics / William Albert, Thomas Tullis.
pages cm
Revised edition of: Measuring the user experience / Tom Tullis, Bill Albert. 2008.
Includes bibliographical references and index.
ISBN 978-0-12-415781-1
1. User interfaces (Computer systems) 2. User interfaces (Computer systems)—Measurement.
3. Measurement. 4. Technology assessment. I. Albert, Bill (William) II. Title.
QA76.9.U83T95 2013
005.4'37—dc23

2013005748

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

Printed in the United States of America
13 14 15 16 17 10 9 8 7 6 5 4 3 2 1

Dedication

Tom:
To my wife, Susan, and my daughters, Cheryl and Virginia

Bill:
To my late father, Lee Albert, and late mother-in-law, Gita Mitra

This page intentionally left blank

Contents

Preface to the Second Edition

Welcome to the second edition of "Measuring the User Experience!" The world of user experience, or UX as it's often abbreviated, has changed quite a bit since the first edition of this book was published. In early 2008 the iPad didn't exist yet nor did Android-powered smartphones. The iPhone was still in its infancy; Pinterest, as well as many other social-networking sites, hadn't even been thought of yet; and the Google Chrome browser was just a rumor. We mention those devices and services because they have helped shape users' experiences with and expectations of the technology many people use on a daily basis. Users expect to be able to pick up a new application or device and start using it right away. Making sure they can is where this book comes in.

UX covers all aspects of someone's interaction with a product, application, or system. Many people seem to think of the user experience as some nebulous quality that can't be measured or quantified. We think it can be. And the tools for measuring it are metrics like the following:

- Can users use their smartphone successfully to find the nearest doctor that's in their health plan?
- How long does it take users to make a flight reservation using a travel website?
- How many errors do users make in trying to log in to a new system?
- How many users are successful in using their new tablet application to instruct their digital video recorder to record all episodes of their favorite TV show?
- How many users get into a new "destination-based" elevator without first choosing their desired floor, only to discover there are no floor buttons?
- How many users get frustrated trying to read the tiny serial number hidden under the battery of their new mobile phone when registering for service?
- How many users are delighted by how easy it was to assemble their new bookcase that came with wordless instructions?

These are all examples of behaviors and attitudes that can be measured. Some may be easier to measure than others, but they can all be measured. Success rates, times, number of mouse clicks or keystrokes, self-reported ratings of frustration or delight, and even the number of visual fixations on a link on a web page are all examples of UX metrics. And these metrics can give you invaluable insight into the user experience.

Why would you want to measure the user experience? The answer is simple: to help you improve it. With most consumer products, apps, and websites these days, if you're not improving, you're falling behind. UX metrics can help you determine where you stand relative to your competition and help you pinpoint where you should focus your improvement efforts—the areas that users find the most confusing, inefficient, or frustrating.

This book is a how-to guide, not a theoretical treatise. We provide practical advice about what metrics to collect in what situations, how to collect them, how to make sense of the data using various analysis techniques, and how to present the results in the clearest and most compelling way. We're sharing practical lessons we've learned from our 40+ combined years of experience in this field.

This book is intended for anyone interested in improving the user experience for any type of product, whether it's a consumer product, computer system, application, website, or something else altogether. If it's something people use, then you can measure the user experience associated with it. Those who are interested in improving the user experience and who could benefit from this book come from many different perspectives and disciplines, including usability and UX professionals, interaction designers, information architects, product designers, web designers and developers, software developers, graphic designers, and marketing and market-research professionals, as well as project and product managers.

So what's new in this second edition of the book? Here are some of the highlights:

- New technologies for measuring emotional engagement, including wrist sensors and automated analyses of facial expressions.
- Advances in eye-tracking technology, including remote webcam-based eye tracking.
- New case studies, including examples of what people in the UX field are working on right now. (Chapter 10, Case Studies, is entirely new.)
- New methods and tools for collecting and analyzing UX data, including a variety of online tools.
- Many new examples throughout the book.

We hope that you find this book helpful in your quest to improve the user experience for your products. We'd like to hear about your successes (and failures!). We really value the feedback and suggestions that many readers have given us about the first edition. Much of that feedback helped shape the changes and additions we made in this edition. You can contact us through our website, www.MeasuringUserExperience.com. You will also find supplementary material there, such as the actual spreadsheets and graphs for many of the examples in this book, as well as information about tools that can help in measuring the user experience.

Acknowledgments

First of all, we thank Meg Dunkerley from Elsevier. We appreciate all your hard work and knowing when to push us. We also thank Joe Dumas, Bob Virzi, and Karen Hitchcock for your review of the manuscript. Your suggestions helped make the book more relevant and focused. We owe a debt of gratitude to all the authors who contributed case studies: Erin Bradner, Mary Theofanos, Yee-Yin Choong, Brian Stanton, Tanya Payne, Grant Baldwin, Tony Haverda, Viki Stirling, Caroline Jarrett, Amanda Davis, Elizabeth Rosenzweig, and Fiona Tranquada. The book is far more useful because of your willingness to share your experiences with our readers. Also, we thank Daniel Bender, Sven Krause, and Ben van Dongen for sharing information about your organizations and the exciting technologies you are using. Finally, we are very grateful to our colleagues at Fidelity Investments and the Bentley University Design and Usability Center. We have learned so much from you, and feel very fortunate to work with such an amazing group of UX researchers.

Tom:

I thank my wife, Susan, for your support and assistance in more ways than I can count. You're helping me become a better writer. And I thank my daughters, Cheryl and Virginia, for your continuing encouragement. And for laughing at my stupid jokes.

Bill:

I thank my family. Devika, I am inspired and awed by your love of writing. Arjun, your fascination for numbers helps me think about data and metrics in new ways. To my wife Monika, thank you for all the support and encouragement you had in me while writing this book. I could not have done this without you.

This page intentionally left blank

Biographies

Thomas S. (Tom) Tullis is Vice President of User Experience Research at Fidelity Investments and Adjunct Professor in Human Factors in Information Design at Bentley University. He joined Fidelity in 1993 and was instrumental in the development of the company's User Research department, whose facilities include state-of-the-art Usability Labs. Prior to joining Fidelity, he held positions at Canon Information Systems, McDonnell Douglas, Unisys Corporation, and Bell Laboratories. He and Fidelity's usability team have been featured in a number of publications, including *Newsweek*, *Business 2.0*, *Money*, *The Boston Globe*, *The Wall Street Journal*, and *The New York Times*. Tullis received his B.A. from Rice University, M.A. in Experimental Psychology from New Mexico State University, and Ph.D. in Engineering Psychology from Rice University. With more than 35 years of experience in human–computer interface studies, Tullis has published over 50 papers in numerous technical journals and has been an invited speaker at national and international conferences. He also holds eight United States patents. He coauthored (with Bill Albert and Donna Tedesco) "Beyond the Usability Lab: Conducting Large-Scale Online User Experience Studies," published by Elsevier/Morgan Kauffman in 2010. Tullis was the 2011 recipient of the Lifetime Achievement Award from the User Experience Professionals Association (UXPA) and in 2013 was inducted into the CHI Academy by the ACM Special Interest Group on Computer–Human Interaction (SIGCHI). Follow Tom as @TomTullis.

William (Bill) Albert is currently Executive Director of the Design and Usability Center at Bentley University and Adjunct Professor in Human Factors in Information Design at Bentley University. Prior to joining Bentley University, he was Director of User Experience at Fidelity Investments, Senior User Interface Researcher at Lycos, and Post-Doctoral Researcher at Nissan Cambridge Basic Research. Albert has published and presented his research at more than 30 national and international conferences. In 2010 he coauthored (with Tom Tullis and Donna Tedesco), "Beyond the Usability Lab: Conducting Large-Scale Online User Experience Studies," published by Elsevier/Morgan Kauffman. He is co-Editor in Chief of the *Journal of Usability Studies*. Bill has been awarded prestigious fellowships through the University of California Santa Barbara and the Japanese government for his research in human factors and spatial cognition. He received his B.A. and M.A. degrees from the University of Washington (Geographic Information Systems) and his Ph.D. from Boston University (Geography-Spatial Cognition). He completed a Post-Doc at Nissan Cambridge Basic Research. Follow Bill as @UXMetrics.

This page intentionally left blank

CHAPTER 1

Introduction

CONTENTS

The goal of this book is to show how user experience (UX) metrics can be a powerful tool for evaluating and improving the design of any product. When some people think about user experience metrics, they feel overwhelmed by complicated formulas, contradictory research, and advanced statistical methods. We hope to demystify much of the research and focus on the practical application of UX metrics. We'll walk you through a step-by-step approach to collecting, analyzing, and presenting UX metrics. We'll help you choose the right metrics for each situation or application and show you how to use them to produce reliable, actionable results without breaking your budget. We will introduce some new UX metrics that you might want to consider adding to your toolkit. We'll give you guidelines and tips for analyzing a wide range of metrics and provide many different examples of how to present UX metrics to others in simple and effective ways.

Measuring the User Experience. DOI: http://dx.doi.org/10.1016/B978-0-12-415781-1.00001-7

Our intention is to make this book a practical, how-to guide about measuring the user experience of any product. We aren't going to give you a lot of formulas; in fact, there are very few. The statistics are fairly limited, and the calculations can be done easily in Excel or some other common software package or web application. Our intention is to give you the tools you need to evaluate the user experience of nearly any type of product, without overwhelming you with unnecessary details.

This book is both product and technology neutral. The UX metrics we describe can be used for practically any type of product utilizing nearly any type of technology. This is one of the great features of UX metrics: they aren't just for websites or any single technology. For example, task success and satisfaction are equally valid whether you evaluate a website, a smartphone, or a microwave oven.

The "half-life" of UX metrics is much greater than any specific design or technology. Despite all the changes in technology, the metrics essentially stay the same. Some metrics may change with the development of new technologies to measure the user experience, but the underlying phenomena being measured don't change. Eye tracking is a great example. Many researchers wanted a method for determining where exactly someone is looking at any point in time. Now, with the latest advances in eye-tracking technology, measurement has become much easier and far more accurate. The same can be said for measuring emotional engagement. New technologies in affective computing allow us to measure levels of arousal through very unobtrusive skin conductance monitors as well as facial recognition software. This has offered glimpses into the emotional state of users as they interact with different types of products. These new technologies for measurement are no doubt extremely useful; however, the underlying questions we are all trying to answer don't change that much at all.

So why did we write this book? There's certainly no shortage of books on human factors, statistics, experimental design, and usability methods. Some of those books even cover the more common UX metrics. Does a book that focuses entirely on UX metrics even make sense? Obviously, we think so. In our (humble) opinion, this book makes five unique contributions to the realm of user experience research:

- We take a *comprehensive* look at UX metrics. No other books review so many different metrics. We provide details on collecting, analyzing, and presenting a diverse range of UX metrics.
- This book takes *a practical approach*. We assume you're interested in applying UX metrics as part of your job. We don't waste your time with unnecessary details. We want you to be able to use these metrics easily every day.
- We provide help in making the *right decisions* about UX metrics. One of the most difficult aspects of a UX professional's job is deciding whether to collect metrics and, if so, which ones to use. We guide you through the decision process so that you find the right metrics for *your* situation.

- We provide many *examples* of how UX metrics have been applied within different organizations and how they have been used to answer specific research questions. We also provide in-depth case studies to help you determine how best to use the information revealed by the UX metrics.
- We present UX metrics that can be used with *many different products or technologies*. We take a broad view so that these metrics can be helpful throughout your career even as technology evolves and products change.

This book is organized into three main parts. The first part (Chapters 1–3) provides background information needed to get up to speed on UX metrics.

- Chapter 1 provides an *overview* of user experience and metrics. We define user experience, discuss the value of measuring the user experience, share some of the emerging trends, dispel some of the common myths about UX metrics, and introduce some of the newest concepts in UX measurement.
- Chapter 2 includes *background* information on UX data and some basic statistical concepts. We also provide a guide for performing common statistical procedures related to different UX methods.
- Chapter 3 focuses on *planning* a study involving metrics, including defining participant goals and study goals and choosing the right metrics for a wide variety of situations.

The second part (Chapters 4–9) reviews five general types of UX metrics, as well as some special topics that don't fall neatly into any single type. For each metric, we explain what it is, when to use it, and when not to use it. We show you how to collect data and different ways to analyze and present it. We provide examples of how it has been used in real-world user experience research.

- Chapter 4 covers various *types of performance metrics*, including task success, time on task, errors, efficiency, and ease of learning. These metrics are grouped under an "umbrella" of performance because they measure different aspects of the user's behavior.
- Chapter 5 looks at *measuring usability issues*. Usability issues can be quantified easily by measuring the frequency, severity, and type of issue. We also discuss some of the debates about appropriate sample sizes and how to capture usability issues reliably.
- Chapter 6 focuses on *self-reported metrics*, such as satisfaction, expectations, ease-of-use ratings, confidence, usefulness, and awareness. Self-reported metrics are based on what users share about their experiences, not what the UX professional measures about their actual behaviors.
- Chapter 7 is devoted to *behavioral and physiological metrics*. These metrics include eye tracking, emotional engagement, facial expressions, and various measures of stress. All of these metrics capture something about how the body behaves as a result of the experience of interacting with a user interface.
- Chapter 8 discusses *how to combine different types of metrics and derive new metrics*. Sometimes it's helpful to get an overall assessment of the user

experience of any product. This global assessment is achieved by combining different types of metrics into a single UX score, summarizing them in a UX scorecard, or comparing them to expert performance.
- Chapter 9 presents *special topics* that we believe are important but that don't fit squarely into one of the five general categories. These include A/B testing on a live website, card-sorting data, accessibility data, and return on investment (ROI).

The third part (Chapters 10 and 11) shows how UX metrics are put into practice. In this part, we highlight how UX metrics are actually used within different types of organizations and how to promote the use of metrics within an organization.

- Chapter 10 presents five *case studies*. Each case study reviews how different types of UX metrics were used, how data were collected and analyzed, and the results. These case studies were drawn from UX professionals in various types of organizations, including consulting, government, industry, and not-for-profit/education.
- Chapter 11 provides 10 *steps to help you move forward in using metrics* within your organization. We discuss how UX metrics can fit within different types of organizations, practical tips for making metrics work within your organization, and recipes for success.

1.1 WHAT IS USER EXPERIENCE

Before we try to measure user experience, we should know what it is and what it isn't. While many UX professionals have their own ideas of what constitutes a "user experience," we believe the user experience includes three main defining characteristics:

- A *user* is involved
- That user is interacting with a product, system, or really anything with an interface
- The users' experience is of interest, and observable or measurable

In the absence of a user doing something, we might just be measuring attitudes and preferences, such as in a political poll or survey about your favorite flavor of ice cream. There has to be behavior, or at least potential behavior, to be considered user experience. For example, we might show a screenshot of a website and ask participants what they *would* click if it were interactive.

You might also note that we never defined any characteristics of the product or system. We believe that any system or product can be evaluated from a user experience perspective, as long as there is some type of interface between the system or product and the user. We are hard-pressed to think of any examples of a product that don't have some type of human interface. We think that's a good thing, as it means that we can study almost any product or system from a UX perspective.

Some people distinguish between the terms *usability* and *user experience*. *Usability* is usually considered the ability of the user to use the thing to carry out a task successfully, whereas *user experience* takes a broader view, looking at the individual's entire interaction with the thing, as well as the thoughts, feelings, and perceptions that result from that interaction.

In any casual conversation about usability, most people would agree that it's good to have something that works well and isn't confusing to use. On the other side of the coin, some companies may intentionally design products to be confusing or frustrating. Fortunately, this is a rare occurrence. For the purposes of this book, we will be somewhat idealistic and make the assumption that both users and designers want products to be easy to use, efficient, and engaging.

User experience can sometimes mean the difference between life and death. For example, the health industry is not immune to poor usability. Usability issues abound in medical devices, procedures, and even diagnostic tools. Jakob Nielsen (2005) cites one study that found 22 separate usability issues that contributed to patients receiving the wrong medicine. Even more troubling is that, on average, 98,000 Americans die every year due to medical error (Kohn et al., 2000). While there are no doubt many factors behind this, some speculate that usability and human factors are at least partially to blame.

In some very compelling research, Anthony Andre looked at the design of automatic external defribulators (AEDs)(2003). An AED is a device used to resuscitate an individual experiencing cardiac arrest. AEDs are found in many public spaces, such as shopping malls, airports, and sporting events. An AED is intended to be used by the general public with no background or experience in life-saving techniques such as CPR. The design of an AED is critical, as most individuals who are actually using an AED are experiencing it for the first time, under a tremendous amount of stress. An AED must have simple and clear instructions, and deliver them in a way that is time sensitive and also mitigates user errors. Andre's research compared four different AED manufacturers. He was interested in how each of them performed in terms of users being able to deliver a shock successfully within a specified time limit. He was also interested in identifying specific usability issues that were impacting user performance with each of the machines.

In his 2003 study, he assigned 64 participants to one of four different machines. Participants were asked to enter a room and save a victim (a mannequin lying on the floor) with the AED they were assigned. The results he found were shocking (no pun intended!). While two machines performed as expected (0% errors from a sample of 16 participants for each machine), two other machines did not fare so well. For example, 25% of the participants who used one of the AEDs were not able to deliver a shock to the victim successfully. There were many reasons for this outcome. For example, participants were confused by the instructions on how to remove the packaging for the pads that adhere to the bare chest. Also, the instructions on where to place the electrodes were somewhat confusing.

After Andre shared his research findings with his client, they agreed to address these issues as part of a product redesign effort.

Similar situations can arise on a regular basis in the workplace or in the home. Just think of the written instructions for such actions as lighting the pilot light on a furnace, installing a new lighting fixture, or trying to figure out a tax form. An instruction that's misunderstood or misread can easily result in property damage, personal injury, or even death. User experience plays a much wider role in our lives than most people realize. It's not just about using the latest technology. User experience impacts everyone, every day. It cuts across cultures, age, gender, and economic class. It also makes for some very funny stories!

Saving lives is, of course, not the only motivation for a good user experience. Championing user experience in a business setting is often geared toward increasing revenues and/or decreasing costs. Stories abound of companies that lost money because of the poor user experience of a new product. Other companies have made ease of use a key differentiator as part of their brand message.

The Bentley University Design and Usability Center had the opportunity to work with a not-for-profit organization on the redesign of their charitable-giving website. They were concerned that visitors to their website would have difficulty finding and making donations to the charitable foundation. Specifically, they were interested in increasing the number of recurring donations, as it was an excellent way to build a more continuous relationship with the donor. Our research included a comprehensive usability evaluation with current and potential donors. We learned a great deal about how to not only improve navigation, but simplify the donation form and highlight the benefits of recurring donations. Soon after the launch of the new website, we learned that the redesign effort was a success. Overall donations had increased by 50%, and recurring donations increased from 2, up to 19 (a 6,715% increase!). This was a true usability success story, and one that also benefits a great cause.

User experience takes on an ever-increasing role in our lives as products become more complex. As technologies evolve and mature, they tend to be used by an increasingly diverse set of users. But this kind of increasing complexity and evolution of technology doesn't necessarily mean that the technologies are becoming easier to use. In fact, just the opposite is likely to happen unless we pay close attention to the user experience. As the complexity of technology grows, we believe that user experience must be given more attention and importance, and UX metrics will become a critical part of the development process to provide complex technology that's efficient, easy to use, and engaging.

1.2 WHAT ARE USER EXPERIENCE METRICS?

A *metric* is a way of measuring or evaluating a particular phenomenon or thing. We can say something is longer, taller, or faster because we are able to measure or quantify some attribute of it, such as distance, height, or speed. The process requires agreement on how to measure these things, as well as a consistent and

reliable way of doing it. An inch is the same length regardless of who is measuring it, and a second lasts for the same amount of time no matter what the time-keeping device is. Standards for such measures are defined by a society as a whole and are based on standard definitions of each measure.

Metrics exist in many areas of our lives. We're familiar with many metrics, such as time, distance, weight, height, speed, temperature, and volume. Every industry, activity, and culture has its own set of metrics. For example, the auto industry is interested in the horsepower of a car, its gas mileage, and the cost of materials. The computer industry is concerned with processor speed, memory size, and power requirements. At home, we're interested in similar measurements: how our weight changes when we step on the bathroom scale, where to set our thermostat in the evening, and how to interpret our water bill every month.

The user experience field is no different. We have a set of metrics specific to our profession: task success, user satisfaction, and errors, among others. This book gathers all the UX metrics in one place and explains how to use these metrics to provide maximum benefit to you and your organization.

So what is a UX metric and how does it compare to other types of metrics? Like all other metrics, UX metrics are based on a reliable system of measurement: Using the same set of measurements each time something is measured should result in comparable outcomes. All UX metrics must be *observable* in some way, either directly or indirectly. This observation might be simply noting that a task was completed successfully or noting the time required to complete the task. All UX metrics must be *quantifiable*—they have to be turned into a number or counted in some way. All UX metrics also require that the thing being measured represents some aspect of the user experience, presented in a numeric format. For example, a UX metric might reveal that 90% of the users are able to complete a set of tasks in less than 1 minute or 50% of users failed to notice a key element on the interface.

What makes a UX metric different from other metrics? UX metrics reveal something about the user experience—about the personal experience of the human being using a product or system. A UX metric reveals something about the interaction between the user and the product: some aspect of *effectiveness* (being able to complete a task), *efficiency* (the amount of effort required to complete the task), or *satisfaction* (the degree to which the user was happy with his or her experience while performing the task).

Another difference between UX metrics and other metrics is that they measure something about *people* and their behavior or attitudes. Because people are amazingly diverse and adaptable, we sometimes encounter challenges in our UX metrics. For this reason, we discuss *confidence intervals* with most of the UX metrics discussed in order to reflect the variability in the data. We will also discuss what metrics we consider relevant (and less relevant) in a UX context.

Certain things are not considered UX metrics, such as overall preferences and attitudes not tied to an actual experience of using something. Think of some

standard metrics such as the Presidential Approval Ratings, the Consumer Price Index, or the frequency of purchasing specific products. Although these metrics are all quantifiable and may reflect some type of behavior, they are not based on actually using something in order to reflect the variability in the data.

UX metrics are not an end unto themselves; rather, they are a means to help you reach an informed decision. UX metrics provide answers to questions that are critical to your organization and that can't be answered by other means. For example, UX metrics can answer these critical questions:

- Will the users recommend the product?
- Is this new product more efficient to use than the current product?
- How does the user experience of this product compare to the competition?
- Do the users feel good about the product or themselves after using it?
- What are the most significant usability problems with this product?
- Are improvements being made from one design iteration to the next?

1.3 THE VALUE OF UX METRICS

We think UX metrics are pretty amazing. Measuring the user experience offers so much more than just simple observation. Metrics add structure to the design and evaluation process, give insight into the findings, and provide information to the decision makers. Without the insight provided by metrics, important business decisions may be made based on incorrect assumptions, "gut feelings," or hunches. As a result, some of these decisions are not the best ones.

During a typical usability evaluation, it's fairly easy to spot some of the more obvious usability issues. But it's much harder to estimate the size or magnitude of the issues. For example, if all eight participants in a study have the same exact problem, you can be quite certain it is a common problem. But what if only two or three of the eight participants encounter the problem? What does that mean for the larger population of users? UX metrics offer a way to estimate the number of users likely to experience this problem. Knowing the magnitude of the problem could mean the difference between delaying a major product launch and simply adding an additional item to the bug list with a low priority. Without UX metrics, the magnitude of the problem is just a guess.

User experience metrics show whether you're actually improving the user experience from one product to the next. An astute manager will want to know as close to certain as possible that the new product will actually be better than the current product. UX metrics are the only way to really know if the desired improvements have been realized. By measuring and comparing the current with a new, "improved" product and evaluating the potential improvement, you create a win–win situation. There are three possible outcomes:

- The new version tests better than the current product: Everyone can sleep well at night knowing that improvements were made.

- The new version tests worse than the current version: Steps can be taken to address the problem or put remediation plans into place.
- No difference between the current product and the new product is apparent: The impact on the user experience does not affect the success or failure of the new product. However, improvements in other aspects of the product could make up for the lack of improvement in the user experience.

User experience metrics are a key ingredient in calculating a ROI. As part of a business plan, you may be asked to determine how much money is saved or how revenue increases as a result of a new product design. Without UX metrics, this task is impossible. With UX metrics, you might determine that a simple change in a data input field on an internal website could reduce data entry errors by 75%, reduce the time required to complete the customer service task, increase the number of transactions processed each day, reduce the backlog in customer orders, cut the delay in customer shipments, and increase both customer satisfaction and customer orders, resulting in an overall rise in revenue for the company.

User experience metrics can help reveal patterns that are difficult or even impossible to see. Evaluating a product with a very small sample size (without collecting any metrics) usually reveals the most obvious problems. However, many more subtle problems require the power of metrics. For example, sometimes it's difficult to see small inefficiencies, such as the need to reenter user data whenever a transaction displays a new screen. Users may be able to complete their tasks—and maybe even say they like it—but many small inefficiencies can eventually build up to impact the user experience and slow down the process. UX metrics help you gain new insights and lead toward a better understanding of user behavior.

1.4 METRICS FOR EVERYONE

We've been teaching a class on UX metrics, in one form or another, for almost a decade. During this time, we have met many UX and non-UX professionals who have little-to-no background in statistics, and even a few who were terrified of anything that looks like a number. Despite this, we have continually been impressed and inspired by how these folks are able to learn the basics on how to collect, analyze, and present UX metrics quickly and easily. UX metrics are a very powerful tool, but also easily accessible to almost anyone. The key is simply to try, and learn from your mistakes. The more metrics you collect and analyze, the better you will get! In fact, we even see some individuals who use this book simply as a guide to what types of UX metrics make the most sense for their organization or project and then go off and ask someone else to actually do the dirty work and collect/analyze the data. So, even if you don't want to get your hands dirty, there isn't an excuse for incorporating UX metrics into your work.

We've written this book to be easy and approachable to the broadest possible audience. In fact, we probably favor simplification rather than a deep dive into

heavy statistical analysis. We feel this will help attract as many UX and non-UX people as possible. Of course, we strongly encourage everyone to go beyond this book by creating new metrics tailored to your organization, product, or research practice.

1.5 NEW TECHNOLOGIES IN UX METRICS

Earlier we stated that UX metrics apply to a vast array of products, designs, and technologies. In fact, even with new tchnologies emerging every day, UX metrics still remain highly relevant. However, what does change (and quite rapidly) are the technologies themselves that better allow us to collect and analyze UX data. Throughout the book you will get a sense of some of the newest technologies that might make your job a little easier, and certainly more interesting. We wanted just to highlight a few of the latest technologies that have emerged in the last few years.

There are some exciting new advances in the world of eye tracking. For decades, eye tracking was restricted to the lab. This is no longer the case. Within the last couple of years, two major vendors in eye tracking (Tobii and SMI) have released goggles that can be used to track eye movements in the field. So, as your participant is walking down the aisle at the supermarket, you can gather data on what he/she is looking at and for how long. Of course, it is a little tricky when different objects occur in approximately the same location and at different depths. But, no doubt they are improving these goggles with each new release.

Eye tracking is even moving beyond hardware. For example, EyeTrackShop has developed technology that collects eye movement data through the participant's webcam. So, no longer are you restricted to being in the same location as your participants, now you can literally collect eye-tracking data with anyone in the world, assuming they have an Internet connection and a webcam. This is a very exciting development, and it is certainly going to open up the market for eye-tracking data to many UX professionals who did not have, or could not afford the hardware.

Another exciting new technology is in the area of affective computing. For decades, UX professionals have gained insight into a user's emotional state by listening to and observing the participant, and of course asking all the right questions. These qualitative data have been, and will always be, extremly valuable. However, advances in affective computing have added a new dimension to measuring emotional engagement. Companies such as Affectiva combine data from sensors that measure skin conductance, along with facial recognition software that anayzes different facial expressions. Together, these two pieces of data tell the researcher something about not only the level of arousal, but the valence (whether it is a positive or negative emotion).

There are a host of new unmoderated usability testing tools that make data collection very easy and affordable. Some tools such as UserZoom and Loop11 are

powerful and affordable for collecting a lot of usability data very efficiently. Other tools such as Usabilla and Userlytics do a very nice job of integrating both qualitative and quantitative data for a reasonable price. Other tools, such as UsabilityTesting.com, allow you to essentially run qualitative-based, self-guided usabilty studies very easily and quickly. And, of course, there are some very specialized tools that help track clicks or mouse movements. It is very exciting that there are so many new technologies that the UX researcher can add to his/her suite of tools.

Analyzing open-ended responses has always been very luborious and inprecise. It is all too common for researchers to disregard verbatim comments or just randonly select a small sample for quotes. In the last few years verbatim analysis software has improved greatly to the point that researchers now have the ability to analyze open-ended responses.

1.6 TEN MYTHS ABOUT UX METRICS

There are many common myths about UX metrics. Some of these myths may come from of a lack of experience with using metrics. Perhaps these myths arose from a negative experience (such as someone from marketing screaming about your sample size) or even other UX professionals complaining about the hassles and costs associated with using metrics. Ultimately the source of these myths doesn't matter. What matters is to separate fact from fiction. We've listed 10 of the most common myths surrounding UX metrics and a few examples that dispel these myths.

Myth 1: Metrics Take Too Much Time to Collect

At best, UX metrics can speed up the design process and, at worst, should not impact the overall timeline. Metrics are collected quickly and easily as part of a normal iterative usability evaluation. Project team members may assume incorrectly that full-blown surveys need to be launched or that you have to be testing in the lab for two straight weeks to collect even basic UX metrics. In fact, there are some fairly simple UX metrics you can collect as part of your everyday testing. Adding a few extra questions at the beginning or end of each usability session will not impact the length of the session. Participants can quickly answer a few key questions as part of either a typical background questionnaire or follow-up activities.

Participants can also rate tasks for ease of use or satisfaction after each task or at the end of all tasks. If you have easy access to a large group of target users or a user panel, you can send out an e-mail blast with a few key questions, perhaps with some screenshots. It's possible to collect data from hundreds of users in just 1 day. Some data can also be collected quickly without even involving the user. For example, you can report the frequency and severity of specific issues quickly and easily with each new design iteration. The time it takes to collect metrics doesn't have to be weeks or even days. Sometimes it's just a few extra hours or even minutes.

Myth 2: UX Metrics Cost Too Much Money

Some people believe that the only way to get reliable UX data is to outsource the study to a market research firm or UX/design consultancy. Although this may be helpful in some situations, it can also be quite costly. Many reliable metrics don't cost an arm and a leg. Even as part of your everyday testing, you can collect incredibly valuable data on the frequency and severity of different usability issues. You can also collect huge amounts of quantitative data by sending out short e-mail surveys to fellow employees or a panel of targeted users. Also, some of the best analysis tools are actually free on the web. Although money does help in certain situations, it is by no means necessary to get some great metrics.

Myth 3: UX Metrics are not Useful When Focusing on Small Improvements

Some project team members may question the usefulness of metrics when they are interested in only some fairly small improvements. They may say it's best to focus on a narrow set of improvements and not worry about metrics. They may not have any extra time or budget to collect any UX metrics. They may say that metrics have no place in a rapid-pace iterative design process. Analyzing usability issues is an obvious and incredibly valuable solution. For example, looking at the severity and frequency of usability issues and why they occur is an excellent way to focus resources during the design process. This approach saves the project both money and time. You can easily derive UX metrics based on previous studies that might help you answer key usability questions. UX metrics are useful for large and small projects alike.

Myth 4: UX Metrics Don't Help us Understand Causes

Some people argue that metrics don't help us understand the root cause of user experience problems. They assume (incorrectly) that metrics serve only to highlight the magnitude of the problem. But if they concentrate on only success rates or completion time data, it's easy to see why some might have this perception. Metrics, however, can tell you much more about the root cause of usability issues than you might initially think. You can analyze verbatim comments to reveal the source of the problem and how many users experience it. You can identify where in the system users experience a problem and use metrics to tell where and even why some problems occur. Depending on how the data are coded and the methods used, there is a wealth of UX data that can help reveal the root cause of many UX issues.

Myth 5: UX Metrics are Too Noisy

One big criticism of UX metrics is that the data are too "noisy." Too many variables prevent getting a clear picture of what's going on. The classic example of "noisy" data is measuring task completion time in an automated usability study when the participant goes out for a cup of coffee or, worse, home for the weekend. Although this may happen on occasion, it should not deter you from

collecting task time data or any other type of usability data. There are some simple things that can be done to minimize or even remove noise in the data. UX data can be cleaned up so that extreme values are not used in the analysis. Also, specific metrics can be chosen carefully to mitigate noisy data. Well-defined procedures can be used to ensure that appropriate levels of consistency are achieved in evaluating tasks or usability issues. Many standard questionnaires have already been widely validated by many researchers. The bottom line is that with some careful thought and a few simple techniques, a lot of the noise in UX data can be reduced significantly to show a clear picture of user behavior and attitudes.

Myth 6: You Can Just Trust Your Gut

Many usability decisions are made on a "gut level." There's always someone on the project team who proclaims, "This decision just feels right!" One of the beauties of metrics is that having data takes a lot of the guesswork out of usability decisions. Some design options are truly borderline cases, but they might actually have an impact on a large population. Sometimes the right design solutions are counterintuitive. For example, a design team may ensure that all the information on a web page is above the fold, thereby eliminating the need to scroll. However, usability data (perhaps in the form of task completion times) may reveal longer task completion times because there's not enough white space between the various visual elements. Intuition is certainly important, but data are better.

Myth 7: Metrics Don't Apply to New Products

Some people shy away from metrics when evaluating a new product. They may argue that since there is no point of comparison, metrics don't make sense. We would argue just the opposite. When evaluating a new product, it's critical to establish a set of baseline metrics against which future design iterations can be compared. It's the only way to really know if the design is improving or not. In addition, it's helpful to establish target metrics for new products. Before a product is released, it should meet basic UX metrics around task success, satisfaction, and efficiency.

Myth 8: No Metrics Exist for the Type of Issues We are Dealing with

Some people believe that there aren't any metrics related to the particular product or project they are working on. Whatever the goal of the project, at least a couple of metrics should tie directly to the business goals of the project. For example, some people say they are only interested in the emotional response of users and not in actual task performance. In this case, several well-established ways of measuring emotional responses are available. In other situations, someone might be concerned only with awareness. Very simple ways to measure awareness also exist, even without investing in eye-tracking equipment. Some people say that they are only interested in more subtle reactions of users, such as their level of frustration. There are ways to measure stress levels without actually

asking the user. In our years of UX research, we have yet to come across a business or user goal that was not measurable in some way. You may have to be creative in how you collect the data, but it's always possible.

Myth 9: Metrics are not Understood or Appreciated by Management

Although some managers view user experience research as providing only qualitative feedback about a design or product, most managers see the value of measurement. It has been our experience that UX metrics are not only understood but very much appreciated by upper-level management. They can relate to metrics. Metrics provide credibility to the team, the product, and the design process. Metrics can be used to calculate ROI. Most managers love metrics, and UX metrics are one type of metric they will embrace quickly. UX metrics can also be real attention grabbers with management. It's one thing to say there's a problem with the online checkout process, but it's an entirely different thing to say that 52% of users are unable to purchase a product online successfully once they've found it.

Myth 10: It's Difficult to Collect Reliable Data with a Small Sample Size

A widely held belief is that a large sample size is required to collect any reliable UX metrics. Many people assume that you need at least 30 participants to even start looking at UX data. Although having a larger sample size certainly helps increase the confidence level, smaller sample sizes of 8 or 10 participants can still be meaningful. We will show you how to calculate a confidence interval that takes into account the sample size when making any conclusion. Also, we will show you how to determine the sample size you need to identify usability issues. Most of the examples in this book are based on fairly small sample sizes (fewer than 20 participants). So not only are metrics possible to analyze with fairly small sample sizes, doing so is quite common!

CHAPTER 2

Background

CONTENTS

This chapter covers background information about data, statistics, and graphs that apply to just about any user experience metrics. Specifically, we address the following:

- The basic *types of variables and data* in any user experience study, including independent and dependent variables, and nominal, ordinal, interval, and ratio data.

Measuring the User Experience. DOI: http://dx.doi.org/10.1016/B978-0-12-415781-1.00002-9

- Basic *descriptive statistics* such as the mean and median, standard deviation, and the concept of *confidence intervals*, which reflect how accurate your estimates of measures such as task times, task success rates, and subjective ratings actually are.
- Simple *statistical tests* for comparing means and analyzing relationships between variables.
- Tips for *presenting your data visually* in the most effective way.

We use Microsoft Excel 2010 for all of the examples in this chapter (and really in most of this book) because it is so popular and widely available. Most of the analyses can also be done with other readily available spreadsheet tools such as Google Docs or OpenOffice.org.

2.1 INDEPENDENT AND DEPENDENT VARIABLES

At the broadest level, there are two types of variables in any usability study: independent and dependent. Independent variables are the things you manipulate or control for, such as designs you're testing or the ages of your participants. Dependent variables are the things you measure, such as success rates, number of errors, user satisfaction, completion times, and many more. Most of the metrics discussed in this book are dependent variables.

When designing a user experience study, you should have a clear idea of what you plan to manipulate (independent variables) and what you plan to measure (dependent variables). The most interesting outcomes of a study are at the intersection of the independent and dependent variables, such as whether one design resulted in a higher task success rate than the other.

2.2 TYPES OF DATA

Both independent and dependent variables can be measured using one of four general types of data: nominal, ordinal, interval, and ratio. Each type of data has its own unique characteristics and, most importantly, supports specific types of analyses and statistics. When collecting and analyzing user experience data, you should know what type of data you're dealing with and what you can and can't do with each type.

2.2.1 Nominal Data

Nominal (also called categorical) data are simply unordered groups or categories. Without order between the categories, you can say only that they are different, not that one is any better than the other. For example, consider apples, oranges, and bananas. They are just different; no one fruit is inherently better than any other.

In user experience, nominal data might be characteristics of different types of users, such as Windows versus Mac users, users in different geographic locations, or males vs females. These are typically independent variables that allow you to segment data by these different groups. Nominal data also include some

commonly used dependent variables, such as task success, the number of users who clicked on link A instead of link B, or users who chose to use a remote control instead of the controls on a DVD player itself.

Among the statistics you can use with nominal data are simple descriptive statistics such as counts and frequencies. For example, you could say that 45% of the users are female, there are 200 users with blue eyes, or 95% were successful on a particular task.

CODING NOMINAL DATA

One important thing to consider when working with nominal data is how to code it. In analyzing nominal data, it's not uncommon to represent the membership in each group using numbers. For example, you might code males as group "1" and females as group "2." But remember that those figures are not data to be analyzed as numbers: An average of these values would be meaningless. (You could just as easily code them as "F" and "M.") The software you're using for your analysis can't distinguish between numbers used strictly for coding purposes, like these, and numbers whose values have true meaning. One useful exception to this is task success. If you code task success as a "1" and a failure as "0," the average will represent the proportion of users who were successful.

2.2.2 Ordinal Data

Ordinal data are ordered groups or categories. As the name implies, data are organized in a certain way. However, the intervals between measurements are not meaningful. Some people think of ordinal data as ranked data. For example, the list of the top 100 movies, as rated by the American Film Institute (AFI), shows that their 10th best movie of all time, *Singing in the Rain*, is better than their 20th best movie of all time, *One Flew Over the Cuckoo's Nest*. But these ratings don't say that *Singing in the Rain* is *twice* as good as *One Flew Over the Cuckoo's Nest*. One film is just *better* than the other, at least according to the AFI. Because the distance between the ranks is not meaningful, you cannot say one is twice as good as the other. Ordinal data might be ordered as better or worse, more satisfied or less satisfied, or more severe or less severe. The relative ranking (the order of the rankings) is the only thing that matters.

In user experience, the most common examples of ordinal data come from self-reported data. For example, a user might rate a website as excellent, good, fair, or poor. These are relative rankings: The distance between excellent and good is not necessarily the same distance between good and fair. Or if you were to ask the participants in a usability study to rank order four different designs for a web page according to which they prefer, that would also be ordinal data. There's no reason to assume that the distance between the page ranked first by a participant and the page ranked second is the same as the distance between the

page ranked second and the one ranked third. It could be that the participant really loved one page and hated all three of the others.

The most common way to analyze ordinal data is by looking at frequencies. For example, you might report that 40% of the users rated the site as excellent, 30% as good, 20% percent as fair, and 10% as poor. Calculating an average ranking may be tempting but it's statistically meaningless.

2.2.3 Interval Data

Interval data are continuous data where differences between the values are meaningful, but there is no natural zero point. An example of interval data familiar to most of us is temperature. Defining 0° Celsius or 32° Fahrenheit based on when water freezes is completely arbitrary. The freezing point of water does not mean the absence of heat; it only identifies a meaningful point on the scale of temperatures. But the differences between the values are meaningful: the distance from 10° to 20° is the same as the distance from 20° to 30° (using either scale). Dates are another common example of interval data.

In usability, the System Usability Scale (SUS) is one example of interval data. SUS (described in detail in Chapter 6) is based on self-reported data from a series of questions about the overall usability of any system. Scores range from 0 to 100, with a higher SUS score indicating better usability. The distance between each point along the scale is meaningful in the sense that it represents an incremental increase or decrease in perceived usability.

Interval data allow you to calculate a wide range of descriptive statistics (including averages and standard deviation). There are also many inferential statistics that can be used to generalize about a larger population. Interval data provide many more possibilities for analysis than either nominal or ordinal data. Much of this chapter will review statistics that can be used with interval data.

One of the debates you can get into with people who collect and analyze subjective ratings is whether you must treat the data as purely ordinal or if you can treat it as being interval. Consider these two rating scales:

o Poor o Fair o Good o Excellent

Poor o o o o Excellent

At first glance, you might say those two scales are the same, but the difference in presentation makes them different. Putting explicit labels on items in the first scale makes the data ordinal. Leaving the intervening labels off in the second scale and only labeling the end points make the data more "interval-like," which is why most subjective rating scales only label the ends, or "anchors," and not every data point. Consider a slightly different version of the second scale:

Poor o o o o o o o o o Excellent

Presenting it that way, with 9 points along the scale, makes it even more obvious that the data can be treated as if it were interval data. The reasonable

interpretation of this scale by a user is that the distances between all the data points along the scale are equal. A question to ask yourself when deciding whether you can treat some data like this as interval or not is whether a point halfway between any two of the defined data points makes sense. If it does, then it makes sense to analyze the data as interval data.

2.2.4 Ratio Data

Ratio data are the same as interval data but with the addition of an absolute zero. This means that the zero value is not arbitrary, as with interval data, but has some inherent meaning. With ratio data, differences between the measurements are interpreted as a ratio. Examples of ratio data are age, height, and weight. In each example, zero indicates the absence of age, height, or weight.

In user experience, the most obvious example of ratio data is time. Zero seconds left to complete a task would mean no time or duration remaining. Ratio data let you say something is twice as fast or half as slow as something else. For example, you could say that one user is twice as fast as another user in completing a task.

There aren't many additional analyses you can do with ratio data compared to interval data in usability. One exception is calculating a geometric mean, which might be useful in measuring differences in time. Aside from that calculation, there really aren't many differences between interval and ratio data in terms of the available statistics.

2.3 DESCRIPTIVE STATISTICS

Descriptive statistics are essential for any interval or ratio-level data. Descriptive statistics, as the name implies, describe the data, without saying anything about the larger population. Inferential statistics let you draw some conclusions or infer something about a larger population above and beyond your sample.

The most common types of descriptive statistics are measures of central tendency (such as the mean), measures of variability (such as the standard deviation), and confidence intervals, which pull the other two together. The following sections use the sample data shown in Table 2.1 to illustrate these statistics. These data represent the time, in seconds, that it took each of 12 participants in a usability study to complete the same task.

2.3.1 Measures of Central Tendency

Measures of central tendency are simply a way of choosing a single number that is in some way representative of a set of numbers. The three most common measures of central tendency are the mean, median, and mode.

The mean is what most people think of as the average: the sum of all values divided by how many values there are. The mean of most user experience metrics is extremely useful and is probably the most common statistic cited in a usability report. For the data in Table 2.1, the mean is 35.1 seconds.

Participant	Task Time (seconds)
P1	34
P2	33
P3	28
P4	44
P5	46
P6	21
P7	22
P8	53
P9	22
P10	29
P11	39
P12	50

Table 2.1 Time to complete a task, in seconds, for each of 12 participants in a usability study.

EXCEL TIP

The mean of any set of numbers in Excel can be calculated using the "=AVERAGE" function. The median can be calculated using the "=MEDIAN" function, and the mode can be calculated using the "=MODE" function. If the mode can't be calculated (which happens when each value occurs an equal number of times), Excel returns "#N/A".

The median is the middle number if you put them in order from smallest to largest: half the values are below the median and half are above the median. If there is no middle number, the median is halfway between the two values on either side of the middle. For the data in Table 2.1, the median is equal to 33.5 seconds (halfway between the middle two numbers, 33 and 34). Half of the users were faster than 33.5 seconds and half were slower. In some cases, the median can be more revealing than the mean. For example, let's assume the task time for P12 had been 150 seconds rather than 50.That would change the mean to 43.4 seconds, but the median would be unchanged at 33.5 seconds. It's up to you to decide which is a more representative number, but this illustrates the reason that the median is sometimes used, especially when larger values (or so-called "outliers") may skew the distribution.

The mode is the most commonly occurring value in the set of numbers. For the data in Table 2.1, the mode is 22 seconds, because two participants completed the task in 22 seconds. It's not common to report the mode in usability test results. When data are continuous over a broad range, such as the task times shown in Table 2.1, the mode is generally less useful. When data have a more limited set of values (such as subjective rating scales), the mode is more useful.

NUMBER OF DECIMAL PLACES TO USE WHEN REPORTING DATA

One of the most common mistakes many people make is reporting data from a usability test (mean times, task completion rates, etc.) with more precision than it really deserves. For example, the mean of the times in Table 2.1 is technically 35.08333333 seconds. Is that the way you should report the mean? Of course not. That many decimal places may be mathematically correct, but it's ridiculous from a practical standpoint. Who cares whether the mean was 35.083 or 35.085 seconds? When you're dealing with tasks that took *about* 35 seconds to complete, a few milliseconds or a few hundredths of a second make no difference whatsoever.

So how many decimal places should you use? There's no universal answer, but some of the factors to consider are accuracy of the original data, its magnitude, and its variability. The original data in Table 2.1 appear to be accurate to the nearest second. One rule of thumb is that the number of significant digits you should use when reporting a statistic, such as the mean, is no more than one additional significant digit in comparison to the original data. So in this example, you could report that the mean was 35.1 seconds.

2.3.2 Measures of Variability

Measures of variability reflect how much the data are spread or dispersed across the range of values. For example, these measures help answer the question, "Do most users have similar task completion times or is there a wide range of times?" In most usability studies, variability is caused by individual differences among your participants. There are three common measures of variability: range, variance, and standard deviation.

The range is the distance between the minimum and maximum values. For the data in Table 2.1, the range is 32, with a minimum time of 21 seconds and a maximum time of 53 seconds. The range can vary wildly depending on the metric. For example, in many kinds of rating scales, the range is usually limited to five or seven, depending on the number of values used in the scales. When you study completion times, the range is very useful because it will help identify "outliers" (data points that are at the extreme top and bottom of the range). Looking at the range is also a good check to make sure that the data are coded properly. If the range is supposed to be from one to five, and the data include a seven, you know there is a problem.

EXCEL TIP

The minimum of any set of numbers in Excel can be determined using the "=MIN" function and the maximum using the "=MAX" function. The range can then be determined by MAX-MIN. The variance can be calculated using the "=VAR" function and the standard deviation using the "=STDEV" function.

Variance tells you how spread out the data are relative to the average or mean. The formula for calculating variance measures the difference between each individual data point and the mean, squares that value, sums all of those squares, and then divides the result by the sample size minus 1. For the data in Table 2.1, the variance is 126.4.

Once you know the variance, you can calculate the standard deviation easily, which is the most commonly used measure of variability. The standard deviation is simply the square root of the variance. The standard deviation of the data shown in Table 2.1 is 11.2 seconds. Interpreting the standard deviation is a little easier than interpreting the variance, as the unit of the standard deviation is the same as the original data (seconds, in this example).

EXCEL TIP: DESCRIPTIVE STATISTICS TOOL

An experienced Excel user might be wondering why we didn't just suggest using the "Descriptive Statistics" tool in the Excel Data Analysis ToolPak. (You can add the Data Analysis ToolPak using "Excel Options">"Add-Ins".) This tool will calculate the mean, median, range, standard deviation, variance, and other statistics for any set of data you specify. It's a very handy tool. However, it has what we consider a significant limitation: the values it calculates are static. If you go back and update the original data, the statistics don't update. We like to set up our spreadsheets for analyzing the data from a usability study before we actually collect the data. Then we update the spreadsheet as we're collecting the data. This means we need to use formulas that update automatically, such as MEAN, MEDIAN, and STDEV, instead of the "Descriptive Statistics" tool. But it can be a useful tool for calculating a whole batch of these statistics at once. Just be aware that it won't update if you change the data.

2.3.3 Confidence Intervals

A confidence interval is an estimate of a range of values that includes the true population value for a statistic, such as a mean. For example, assume that you need to estimate the true population mean for a task time whose sample times are shown in Table 2.1. You could construct a confidence interval around that mean to show the range of values that you are reasonably certain will include the true population mean. The phrase "reasonably certain" indicates that you will need to choose how certain you want to be or, put another way, how willing you are to be wrong in your assessment. This is what's called the confidence level that you choose or, conversely, the alpha level for the error that you're willing to accept. For example, a confidence level of 95%, or an alpha level of 5%, means that you want to be 95% certain, or that you're willing to be wrong 5% of the time.

There are three variables that determine the confidence interval for a mean:

- The sample size, or the number of values in the sample. For the data in Table 2.1, the sample size is 12, as we have data from 12 participants.
- The standard deviation of the sample data. For our example, that is 11.2 seconds.

- The alpha level we want to adopt. The most common alpha levels (primarily by convention) are 5 and 10%. Let's choose an alpha of 5% for this example, which is a 95% confidence interval.

The 95% confidence interval is then calculated using the following formula:

$$\text{Mean} \pm 1.96 * [\text{standard deviation}/\text{sqrt}(\text{sample size})]$$

The value "1.96" is a factor that reflects the 95% confidence level. Other confidence levels have other factors. This formula shows that the confidence interval will get smaller as the standard deviation (the variability of data) decreases or as the sample size (number of participants) increases.

EXCEL TIP

You can calculate the confidence interval quickly for any set of data using the CONFIDENCE function in Excel. The formula is easy to construct:

$$= \text{CONFIDENCE}(\text{alpha, standard deviation, sample size})$$

Alpha is your significance level, which is typically 5% (0.05) or 10% (0.10). The standard deviation can be calculated using the STDEV function. The sample size is simply the number of cases or data points you are examining, which can be calculated using the COUNT function. Figure 2.1 shows an example. For the data in Table 2.1, the result of this calculation is 6.4 seconds. Since the mean is 35.1 seconds, the 95% confidence interval for that mean is 35.1±6.4, or 28.7 to 41.5 seconds. So you can be 95% certain that the true population mean for this task time is between 28.7 and 41.5 seconds.

C4 fx =CONFIDENCE(0.05,STDEV(B2:B13),COUNT(B2:B13))

	A	B	C	D
1	Participant	Task Time (seconds)		
2	P1	34		
3	P2	33	95% confidence interval:	
4	P3	28	6.4	
5	P4	44		
6	P5	46		
7	P6	21		
8	P7	22		
9	P8	53		
10	P9	22		
11	P10	29		
12	P11	39		
13	P12	50		

Figure 2.1 Example of how to calculate a 95% confidence interval using the "confidence" function in Excel.

Confidence intervals are incredibly useful. We think you should calculate and display them routinely for just about any means that you report from a usability study. When displayed as error bars on a graph of means, they make it visually obvious how accurate the measures actually are.

WHAT CONFIDENCE LEVEL SHOULD YOU USE?

How should you decide what confidence level to use? Traditionally, the three commonly used confidence levels are 99, 95, and 90% (or their corresponding alpha levels of 1, 5, and 10%). The history behind the use of these three levels goes back to the days before computers and calculators, when you had to look up values for confidence levels in printed tables. The people printing these tables didn't want to print a bunch of different versions, so they made just these three. Today, of course, all of these calculations are done for us so we could choose any confidence level we want. But because of their longstanding use, most people choose one of these three.

The level you choose really does depend on how certain you need to be that the confidence interval contains the true mean. If you're trying to estimate how long it will take someone to administer a life-saving shock using an automated external defibrillator, you probably want to be very certain of your answer, and would likely choose at least 99%. But if you're simply estimating how long it will take someone to upload a new photo to their Facebook page, you probably would be satisfied with a 90% confidence level. In our day-to-day use of confidence intervals, we find that we use a 90% confidence level most of the time, sometimes a 95% level, and rarely a 99% level.

2.3.4 Displaying Confidence Intervals as Error Bars

Let's now consider the data in Figure 2.2, which shows the checkout times for two different designs of a prototype website. In this study, 10 participants performed the checkout task using Design A and another 10 participants performed the checkout task using Design B. Participants were assigned randomly to one group or the other. The means and 90% confidence interval for both groups have been calculated using the AVERAGE and CONFIDENCE functions. The means have been plotted as a bar graph, and the confidence intervals are shown as error bars on the graph. Even just a quick glance at this bar graph shows that the error bars for these two means don't overlap with each other. When that is the case, you can safely assume that the two means are significantly different from each other.

EXCEL TIP

Once you've created a bar graph showing the means, such as Figure 2.2, you then want to add error bars to represent the confidence intervals. First, click on the chart to select it. Then, in the Excel button bar, choose the "Layout" tab under "Chart Tools." On the "Layout" tab, choose "Error Bars>More Error Bars Options." In the resulting dialog box, select the "Custom" option near the bottom of the dialog box. Then click on the

	A	B	C
1		Design A	Design B
2		83	145
3		102	89
4		86	182
5		106	128
6		56	139
7		112	99
8		82	125
9		60	120
10		104	154
11		71	91
12	Means	86.2	127.2
13	90% Confidence Interval	10.2	15.3

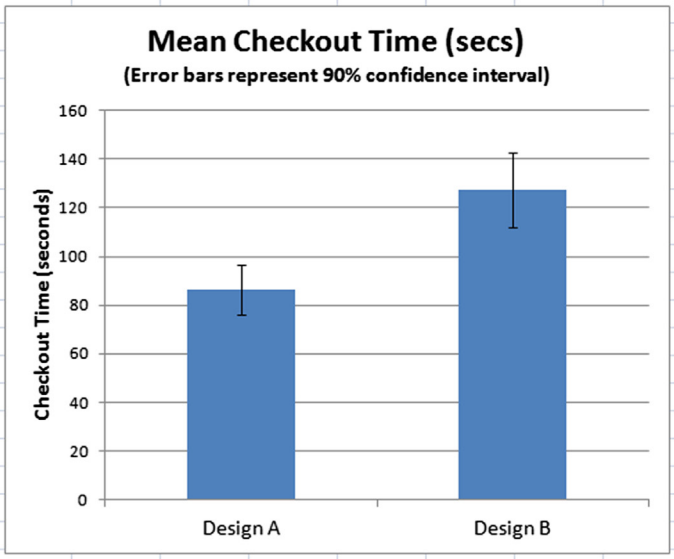


Figure 2.2 Illustration of displaying confidence intervals as error bars on bar graph.

"Specify Value" button. The resulting small window allows you to specify the values for the positive and negative portions of the error bars, which will both be the same. Click on the button to specify the Positive Error Value and then select **both** of the values for the 90% confidence interval on the spreadsheet (cells B13 and C13 in Figure 2.2). Then click on the button for the Negative Error Value and select the exact same cells again. Close both windows and your error bars should be on the graph.

2.4 COMPARING MEANS

One of the most useful things you can do with interval or ratio data is to compare different means. If you want to know whether one design has higher satisfaction ratings than another or if the number of errors is higher for one group of users compared to another, your best approach is through statistics.

There are several ways to compare means, but before jumping into the statistics, you should know the answers to a couple of questions:

1. Is the comparison *within* the same set of users or *across* different users? For example, if you are comparing some data for men vs women, it is highly likely that these are different users. When comparing different samples like this, it's called independent samples. But if you're comparing the *same* group of users on different products or designs, you will use something called paired samples.

2. How many samples are you comparing? If you are comparing two samples, use a *t* test. If you are comparing three or more samples, use an analysis of variance (also called ANOVA).

2.4.1 Independent Samples

Perhaps the simplest way to compare means from independent samples is using confidence intervals, as shown in the previous section. In comparing the confidence intervals for two means, you can draw the following conclusions:

- If the confidence intervals *don't overlap*, you can safely assume the two means are significantly different from each other (at the confidence level you chose).
- If the confidence intervals *overlap slightly*, the two means might still be significantly different. Run a *t* test to determine if they are different.
- If the confidence intervals *overlap widely*, the two means are not significantly different.

Let's consider the data in Figure 2.3 to illustrate running a *t* test for independent samples. This shows the ratings of ease of use on a 1 to 5 scale for two different designs as rated by two different groups of participants (who were assigned randomly to one group or the other). We've calculated the means and confidence intervals and graphed those. But note that the two confidence intervals overlap slightly: Design 1's interval goes up to 3.8, whereas Design 2's goes down to 3.5. This is a case where you should run a *t* test to determine if the two means are significantly different.

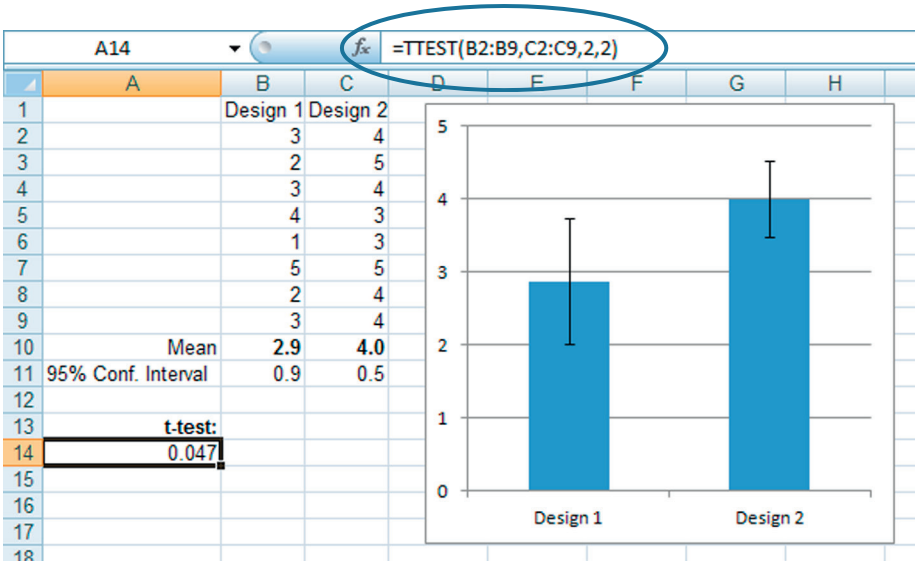


Figure 2.3 Example of a *t* test on independent samples.

EXCEL TIP

As illustrated in Figure 2.3, you can use the TTEST function in Excel to run a *t* test:

=TTEST(Array 1, Array 2, Tails, Type)

Array 1 and Array 2 refer to the sets of values that you want to compare. In Figure 2.3, Array 1 is the set of ratings for Design 1 and Array 2 is the set of ratings for Design 2. Tails refer to whether your test is one-tailed or two-tailed. This relates to the tails (extremes) of the normal distribution and whether you're considering one end or both ends. From a practical standpoint, this is asking whether it is theoretically possible for the difference between these two means to be in either direction (i.e., Design 1 *either* higher or lower than Design 2). In almost all cases that we deal with, the difference could be in either direction, so the correct choice is "2" for two tailed. Finally, Type indicates the type of *t* test. For these independent samples (not paired), the Type is 2.

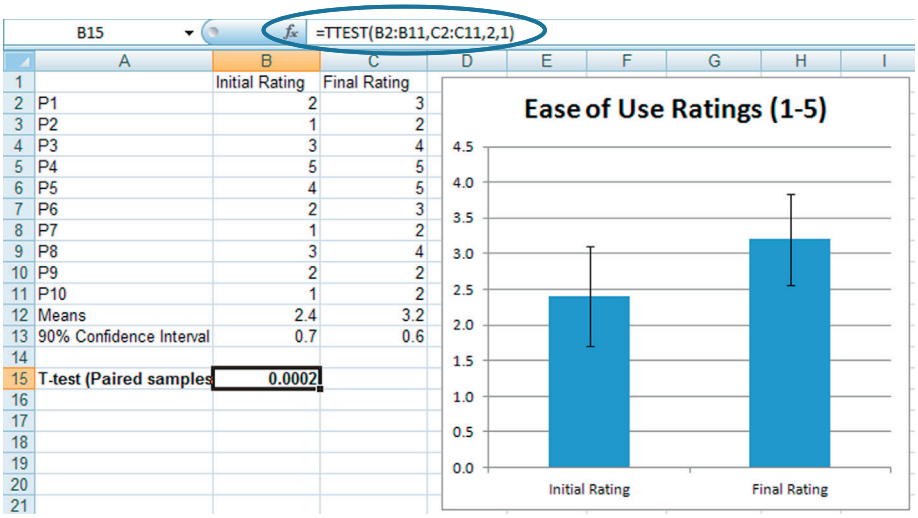
This *t* test returns a value of 0.047. So how do you interpret that? It's telling you the probability that the difference between these two means is simply due to chance. So there's a 4.7% chance that this difference is *not* significant. Since we were dealing with a 95% confidence interval, or a 5% alpha level, and this result is less than 5%, we can say that the difference is statistically significant at that level.

2.4.2 Paired Samples

A paired samples *t* test is used when comparing means within the same set of users. For example, you may be interested in knowing whether there is a difference between two prototype designs. If you have the same set of users perform tasks using prototype A and then prototype B, and you are measuring variables such as self-reported ease of use and time, you will use a paired samples *t* test.

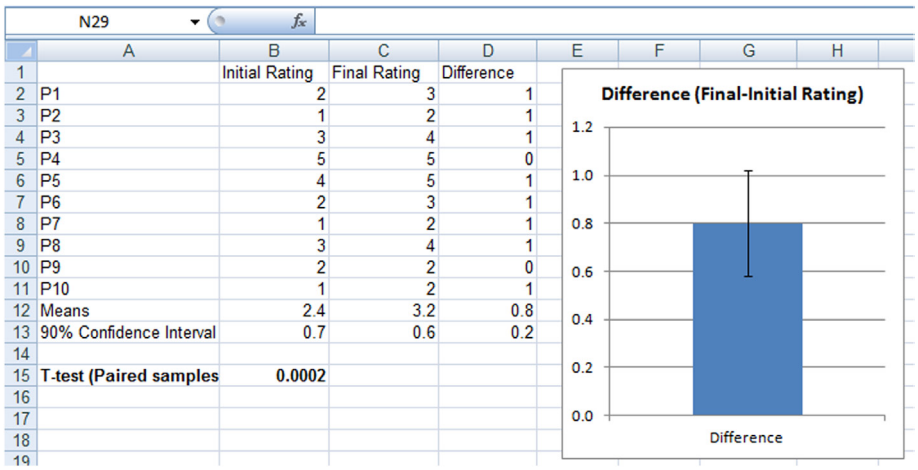
With paired samples like these, the key is that you're comparing each person to themselves. Technically, you're looking at the difference in each person's data for the two conditions you're comparing. Let's consider the data shown in Figure 2.4, which shows "Ease of Use" ratings for an application after their initial use and then again at the end of the session. So there were 10 participants who gave two ratings each. The means and 90% confidence intervals are shown and have been graphed. Note that the confidence intervals overlap pretty widely. If these were independent samples, you could conclude that the ratings are not significantly different from each other. However, because these are paired samples, we've done a *t* test on paired samples (with the "Type" as "1"). That result, 0.0002, shows that the difference is highly significant.

Let's look at the data in Figure 2.4 in a slightly different way, as shown in Figure 2.5. This time we've simply added a third column to the data in which the initial rating was subtracted from the final rating for each participant. Note that for 8 of the 10 participants, the rating increased by one point, whereas for 2 participants it stayed the same. The bar graph shows the mean of those differences (0.8) as well as the confidence interval for that mean difference. In a



	A	B	C
1		Initial Rating	Final Rating
2	P1	2	3
3	P2	1	2
4	P3	3	4
5	P4	5	5
6	P5	4	5
7	P6	2	3
8	P7	1	2
9	P8	3	4
10	P9	2	2
11	P10	1	2
12	Means	2.4	3.2
13	90% Confidence Interval	0.7	0.6
14			
15	**T-test (Paired samples**	**0.0002**	
16			
17			
18			
19			
20			
21			

Figure 2.4 Data showing paired samples in which each of 10 participants gave an ease of use rating (on a 1–5 scale) to an application after an initial task and at the end of the study.



	A	B	C	D
1		Initial Rating	Final Rating	Difference
2	P1	2	3	1
3	P2	1	2	1
4	P3	3	4	1
5	P4	5	5	0
6	P5	4	5	1
7	P6	2	3	1
8	P7	1	2	1
9	P8	3	4	1
10	P9	2	2	0
11	P10	1	2	1
12	Means	2.4	3.2	0.8
13	90% Confidence Interval	0.7	0.6	0.2
14				
15	**T-test (Paired samples**	**0.0002**		
16				
17				
18				
19				

Figure 2.5 Same data as in Figure 2.4, but also showing the difference between initial and final ratings, the mean of those differences, and the 90% confidence interval.

paired-samples test like this, you're basically testing to see if the confidence interval for the mean difference includes 0 or not. If not, the difference is significant.

Note that in a paired samples test, you should have an equal number of values in each of the two sets of numbers that you're comparing (although it is possible to have missing data). In the case of independent samples, the number of values does not need to be equal. You might happen to have more participants in one group than the other.

2.4.3 Comparing More Than Two Samples

We don't always compare only two samples. Sometimes we want to compare three, four, or even six different samples. Fortunately, there is a way to do this without a lot of pain. An ANOVA lets you determine whether there is a significant difference across more than two groups.

Excel lets you perform three types of ANOVAs. We will give an example for just one type of ANOVA, called a single-factor ANOVA. A single-factor ANOVA is used when you just have one variable you want to examine. For example, you might be interested in comparing task completion times across three different prototypes.

Let's consider the data shown in Figure 2.6, which shows task completion times for three different designs. There were a total of 30 participants in this study, with each using only one of the three designs.

EXCEL TIP

To run an ANOVA in Excel requires the Analysis ToolPak. From the "Data" tab, choose the "Data Analysis" button, which is probably on the far right of the button bar. Then choose "ANOVA: Single Factor." This just means that you are looking at one variable (factor). Next, define the range of data. In our example (Figure 2.6), the data are in columns B, C, and D. We have set an alpha level to 0.05 and have included our labels in the first row.

Results are shown in two parts (the right-hand portion of Figure 2.6). The top part is a summary of the data. As you can see, the average time for Design 2 is quite a bit slower, and Designs 1 and 3 completion times are faster. Also, the variance is greater for Design 2 and less for Designs 1 and 3. The second part of the output lets us know whether this difference is significant. The *p* value of 0.000003 reflects the statistical significance of this result. Understanding exactly what this means is important: It means that there is a significant effect of the "designs" variable.

	A	B	C	D	E	F	G	H	I	J	K	L
1		Design 1	Design 2	Design 3		Anova: Single Factor						
2		34	49	22								
3		33	54	28		SUMMARY						
4		28	52	21		*Groups*	*Count*	*Sum*	*Average*	*Variance*		
5		44	39	30		Design 1	10	335	33.5	43.16667		
6		21	60	32		Design 2	10	490	49	63.33333		
7		40	58	36		Design 3	10	302	30.2	38.62222		
8		36	49	27								
9		29	34	40								
10		32	46	37		ANOVA						
11		38	49	29		*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
12	Means	33.5	49.0	30.2		Between Groups	2015.267	2	1007.633	20.83003	0.000003	3.354131
13	90% Conf Interval	3.4	4.1	3.2		Within Groups	1306.1	27	48.37407			
14												
15						Total	3321.367	29				

Figure 2.6 Task completion times for three different designs (used by different participants) and results of a single-factor ANOVA.

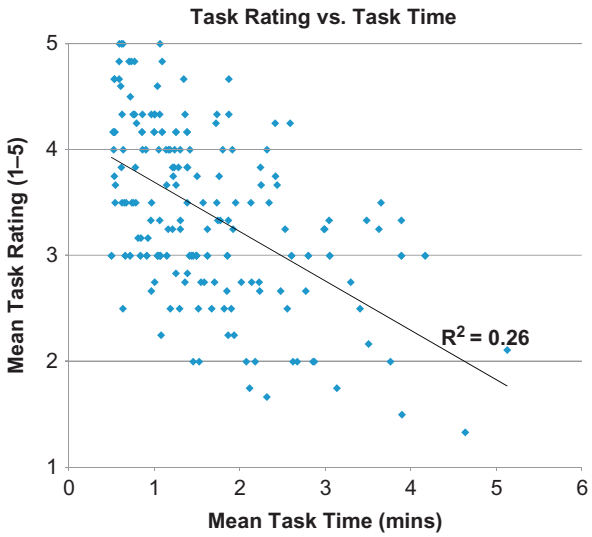


Figure 2.7 An example of a scatterplot (with trend line) in Excel.

It does not necessarily mean that each of the design means is significantly different from each of the others—only that there *is* an effect overall. To see if any two means are significantly different from each other, you could do a two sample *t* test on just those two sets of values.

2.5 RELATIONSHIPS BETWEEN VARIABLES

Sometimes it's important to know about the relationship between different variables. We've seen many cases where someone observing a usability test for the first time remarks that what users say and what they do don't always correspond with each other. Many users will struggle to complete just a few tasks with a prototype, but when asked to rate how easy or difficult it was, they often give it good ratings. This section provides examples of how to perform analyses that investigate these kinds of relationships (or lack thereof).

2.5.1 Correlations

When you first begin examining the relationship between two variables, it's important to visualize what the data look like. That's easy to do in Excel using a scatterplot. Figure 2.7 is an example of a scatterplot of actual data from an online usability study. The horizontal axis shows mean task time in minutes, and the vertical axis shows mean task rating (1–5, with higher numbers being better). Note that as the mean task time increases, the average task rating drops. This is called a negative relationship because as one variable increases (task time), the other variable decreases (task rating). The line that runs through the data is called a trend line and is added easily to the chart in Excel by right-clicking on any one of the data points and selecting "Add Trend Line." The trend line helps you better visualize the relationship between the two variables. You can also have Excel display the R^2 value (a measure of the strength of the relationship) by right-clicking on the trend line, choosing "Format Trend Line," and checking the box next to "Display R-squared value on chart."

EXCEL TIP

You can calculate the strength of the relationship between any two variables (such as task time and task rating) using the CORREL function in Excel:

= CORREL(Array 1, Array 2)

Array 1 and Array 2 are the two sets of numbers to be correlated. The result will be a correlation coefficient, or "r". For the data represented in Figure 2.7, $r = -0.53$. A correlation

coefficient is a measure of the strength of the relationship between the two variables and has a range from −1 to +1. The stronger the relationship, the closer the value is to −1 or +1. The weaker the relationship, the closer the correlation coefficient is to 0. The negative value for "r" signifies the negative relationship between the two variables. If you square the correlation coefficient you get the same value as the R^2 value shown on the scatterplot (0.28).

2.6 NONPARAMETRIC TESTS

Nonparametric tests are used for analyzing nominal and ordinal data. For example, you might want to know if a significant difference exists between men and women for success and failure on a particular task. Or perhaps you're interested in determining whether there is a difference among experts, intermediates, and novices on how they ranked different websites. To answer questions that involve nominal and ordinal data, you will need to use some type of nonparametric test.

Nonparametric statistics make different assumptions about the data than the statistics we've reviewed for comparing means and describing relationships between variables. For instance, when we run *t* tests and correlation analysis, we assume that data are distributed normally and the variances are approximately equal. The distribution is not normal for nominal or ordinal data. Therefore, we don't make the same assumptions about the data in nonparametric tests. For example, in the case of (binary) success, when there are only two possibilities, the data are based on the binomial distribution. Some people like to refer to nonparametric tests as "distribution-free" tests. There are a few different types of nonparametric tests, but we will just cover the χ^2 test because it is probably the most commonly used.

2.6.1 The χ^2 Test

The χ^2 (pronounced "chi square") test is used when you want to compare nominal (or categorical) data. Let's consider an example. Assume you're interested in knowing whether there is a significant difference in task success among three different groups: novice, intermediates, and experts. You run a total of 60 people in your study, 20 in each group. You measure task success or failure on a single task. You count the number of people who were successful in each group. For novices, only 6 out of 20 were successful, 12 out of 20 intermediates were successful, and 18 out of 20 experts were successful. You want to know if there is a statistically significant difference among the groups.

EXCEL TIP

To perform a χ^2 test in Excel, you use the "CHITEST" function. This function calculates whether differences between observed and expected values are simply due to chance. The function is relatively easy is to use:

= CHITEST(actual_range, expected_range)

The actual range is the number of people successful on the task for each group. The expected range is the total number of people successful (33) divided by the number of groups (3), or 11 in this example. The expected value is what you would expect if there were no differences among any of the three groups.

C7 =CHITEST(B2:B4,C2:C4)

	A	B	C	D
1	**Group**	**Observed**	**Expected**	
2	Novice	6	11	
3	Intermediate	9	11	
4	Experts	18	11	
5	**Total**	**33**	**33**	
6				
7		Chi Test	0.029	

Figure 2.8 Output from a χ^2 test in Excel.

C13 =CHITEST(B3:C5,B9:C11)

	A	B	C	D
1		***Observed***	***Observed***	
2	**Group**	**Design A**	**Design B**	
3	Novice	4	2	
4	Intermediate	6	3	
5	Expert	12	6	
6				
7		***Expected***	***Expected***	
8	**Group**	**Design A**	**Design B**	
9	Novice	5.5	5.5	
10	Intermediate	5.5	5.5	
11	Expert	5.5	5.5	
12				
13		Chi Test	0.003	

Figure 2.9 Output from a χ^2 test with two variables.

Figure 2.8 shows what the data look like and output from the CHITEST function. In this example, the likelihood that this distribution is due to chance is about 2.9% (0.029). Because this number is less than 0.05 (95% confidence), we can reasonably say that there is a difference in success rates among the three groups.

In this example we were just examining the distribution of success rates across a single variable (experience group). There are some situations in which you might want to examine more than one variable, such as experience group and design prototype. Performing this type of evaluation works the same way. Figure 2.9 shows data based on two different variables: group and design. For a more detailed example of using χ^2 to test for differences in live website data for two alternative pages (so-called A/B tests), see Chapter 9.

2.7 PRESENTING YOUR DATA GRAPHICALLY

You might have collected and analyzed the best set of usability data ever, but it's of little value if you can't communicate it effectively to others. Data tables are certainly useful in some situations, but in most cases you'll want to present your data graphically. A number of excellent books on the design of effective data graphs are available, including those written by Edward Tufte (1990, 1997, 2001, 2006), Stephen Few (2006, 2009, 2012), and Dona Wong (2010). Our intent in this section is simply to introduce some of the most important principles in the design of data graphs, particularly as they relate to user experience data.

We've organized this section around tips and techniques for five basic types of data graphs:

Column or bar graphs
Line graphs
Scatterplots
Pie or donut charts
Stacked bar or column graphs

We begin each of the following sections with one good example and one bad example of that particular type of data graph.

GENERAL TIPS FOR DATA GRAPHS

Label the axes and units. It might be obvious to you that a scale of 0 to 100% represents the task completion rate, but it may not be obvious to your audience. You might know that the times being plotted on a graph are minutes, but your audience may be left pondering whether they could be seconds or even hours. Sometimes the labels on an axis make it clear what the scale is (e.g., "Task 1," "Task 2," etc.), in which case adding a label for the axis itself would be redundant.

Don't imply more precision in your data than it deserves. Labeling your time data with "0.00" seconds to "30.00" seconds is almost never appropriate, nor is labeling your task completion data with "0.0%" to "100.0%." Whole numbers work best in most cases. Exceptions include some metrics with a very limited range and some statistics that are almost always fractional (e.g., correlation coefficients).

Don't use color alone to convey information. Of course, this is a good general principle for the design of any information display, but it's worth repeating. Color is used commonly in data graphs, but make sure it's supplemented by positional information, labels, or other cues that help someone who can't clearly distinguish colors to interpret the graph.

Show confidence intervals whenever possible. This mainly applies to bar graphs and line graphs that are presenting means of individual participant data (times, ratings, etc.). Showing 95 or 90% confidence intervals for means via error bars is a good way to visually represent the variability in data.

Don't overload your graphs. Just because you *can* create a single graph that shows the task completion rate, error rate, task times, and subjective ratings for each of 20 tasks, broken down by novice versus experienced users, doesn't mean you *should*.

Be careful with 3D graphs. If you're tempted to use a 3D graph, ask yourself whether it really helps. In many cases, the use of 3D makes it harder to see the values being plotted.

2.7.1 Column or Bar Graphs

Column graphs and bar graphs (Figure 2.10) are the same thing; the only difference is their orientation. Technically, column graphs are vertical and bar graphs are horizontal. In practice, most people refer to both types simply as bar graphs, which is what we will do.

Bar graphs are probably the most common way of displaying usability data. Almost every presentation of data from a usability test that we've seen has included at least one bar graph, whether it was for task completion rates, task times, self-reported data, or something else. The following are some of the principles used for bar graphs.

- Bar graphs are appropriate when you want to present the values of continuous data (e.g., times, percentages) for discrete items or categories (e.g., tasks, participants, designs). If both variables are continuous, a line graph is appropriate.

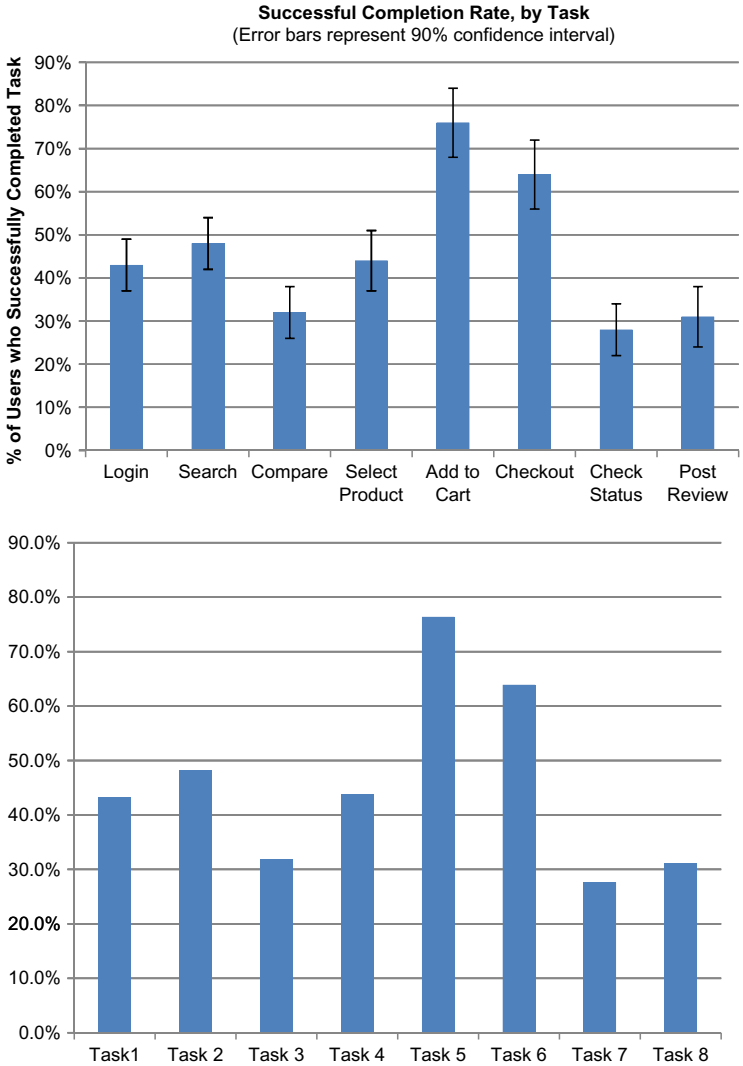


Figure 2.10 Good (top) and bad (bottom) examples of bar graphs for the same data. Mistakes in the bad version include failing to label data, not starting the vertical axis at 0, not showing confidence intervals when you can, and showing too much precision in the vertical axis labels.

- The axis for the continuous variable (the vertical axis in Figure 2.10) should normally start at 0. The whole idea behind bar graphs is that the lengths of the bars represent the values being plotted. By not starting the axis at 0, you're manipulating their lengths artificially. The bad example in Figure 2.10 gives the impression that there's a larger difference between the tasks than there really is. A possible exception is when you include error bars, making it clear which differences are real and which are not.

- Don't let the axis for the continuous variable go any higher than the maximum value that's theoretically possible. For example, if you're plotting percentages of users who completed each task successfully, the theoretical maximum is 100%. If some values are close to that maximum, Excel and other packages will tend to automatically increase the scale beyond the maximum, especially if error bars are shown.

2.7.2 Line Graphs

Line graphs (Figure 2.11) are used most commonly to show trends in continuous variables, often over time. Although not as common as bar graphs in presenting usability data, they certainly have their place. The following are some of the key principles for using line graphs.

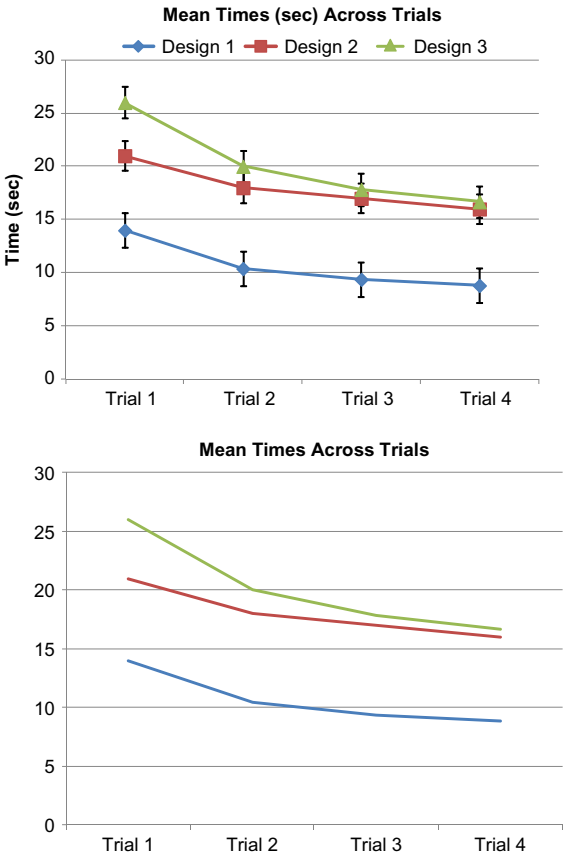


Figure 2.11 Good (top) and bad (bottom) examples of line graphs for the same data. Mistakes in the bad version include failing to label the vertical axis, not showing data points, not including a legend, and not showing confidence intervals.

- Line graphs are appropriate when you want to present the values of one continuous variable (e.g., percent correct, number of errors) as a function of another continuous variable (e.g., age, trial). If one of the variables is discrete (e.g., gender, participant, task), then a bar graph is more appropriate.
- Show your data points. Your actual data points are the things that really matter, not the lines. The lines are just there to connect the data points and make the trends more obvious. You may need to increase the default size of the data points in Excel.
- Use lines that have sufficient weight to be clear. Very thin lines are not only hard to see, but it's harder to detect their color and they may imply a greater precision in data than is appropriate. You may need to increase the default weight of lines in Excel.
- Include a legend if you have more than one line. In some cases, it may be clearer to move the labels manually from the legend into the body of the graph and put each label beside its appropriate line. It may be necessary to do this in PowerPoint or some other drawing program.
- As with bar graphs, the vertical axis normally starts at 0, but it's not as important with a line graph to always do that. There are no bars whose length is important, so sometimes it may be appropriate to start the vertical axis at a higher value. In that case, you should mark the vertical axis appropriately. The traditional way of doing this is with a "discontinuity" marker (≷) on that axis. Again, it may be necessary to do that in a drawing program.

LINE GRAPHS VERSUS BAR GRAPHS

Some people have a hard time deciding whether it's appropriate to use a line graph or a bar graph to display a set of data. Perhaps the most common data-graph mistake we see is using a line graph when a bar graph is more appropriate. If you're considering presenting some data with a line graph, ask yourself a simple question: Do the places along the line *between* the data points make sense? In other words, even though you don't have data for those locations, would they make sense if you did? If they don't make sense, a bar graph is more appropriate. For example, it's technically possible to show the data in Figure 2.10 as a line graph, as shown in Figure 2.12.However, you should ask yourself whether things such as "Task 1½" or "Task 6¾" make any sense, because the lines imply that they should. Obviously, they don't, so a bar graph is the correct representation. The line graph might make an interesting picture, but it's a misleading picture.

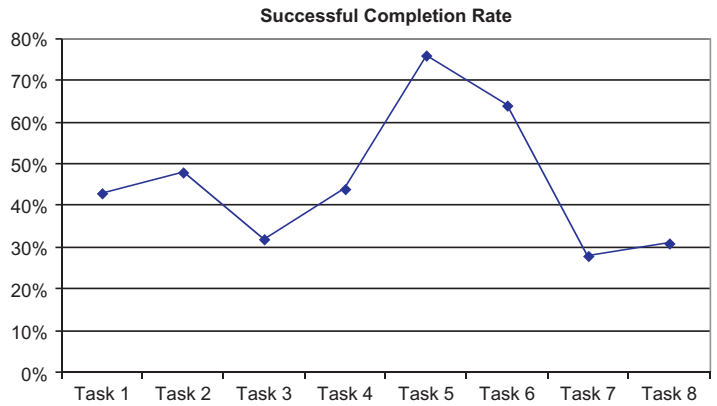


Figure 2.12 An inappropriate line graph of data shown in Figure 2.10. Lines imply that the tasks are a continuous variable, which they are not.

2.7.3 Scatterplots

Scatterplots (Figure 2.13), or X/Y plots, show pairs of values. Although they're not very common in usability reports, they can be very useful in certain situations, especially to illustrate relationships between two variables. Here are some of the key principles for using scatterplots.

- You must have paired values that you want to plot. A classic example is heights and weights of a group of people. Each person would appear as a data point, and the two axes would be height and weight.
- Normally, both of the variables would be continuous. In Figure 2.13, the vertical axis shows mean values for a visual appeal rating of 42 web pages (from Tullis & Tullis, 2007). Although that scale originally had only four

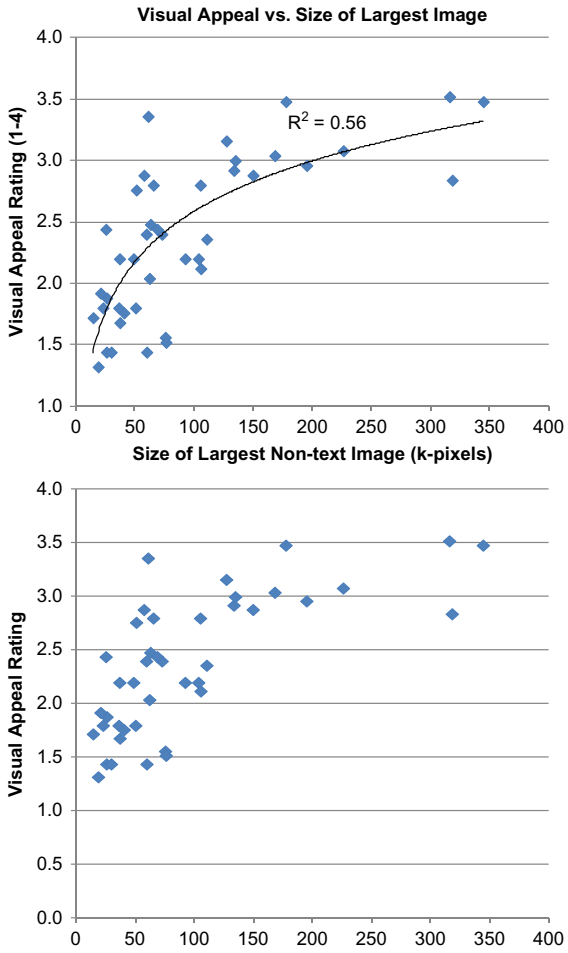


Figure 2.13 Good (top) and bad (bottom) examples of scatterplots for the same data. Mistakes in the bad version include an inappropriate scale for the vertical axis, not showing the scale for visual appeal ratings (1–4), not showing a trend line, and not showing goodness of fit (R^2).

values, the means come close to being continuous. The horizontal axis shows the size, in k pixels, of the largest nontext image on the page, which truly is continuous.

- You should use appropriate scales. In Figure 2.13, the values on the vertical axis can't be any lower than 1.0, so it's appropriate to start the scale at that point rather than 0.
- Your purpose in showing a scatterplot is usually to illustrate a relationship between the two variables. Consequently, it's often helpful to add a trend line to the scatterplot, as in the good example in Figure 2.13. You may want to include the R^2 value to indicate the goodness of fit.

2.7.4 Pie or Donut Charts

Pie or donut charts (Figure 2.14) illustrate the parts or percentages of a whole. They can be useful any time you want to illustrate the relative proportions of the parts of a whole to each other (e.g., how many participants in a usability test succeeded, failed, or gave up on a task). Here are some key principles for their use.

- Pie or donut charts are appropriate only when the parts add up to 100%. You have to account for all the cases. In some situations, this might mean creating an "other" category.
- Minimize the number of segments in the chart. Even though the bad example in Figure 2.14 is technically correct, it's almost impossible to make any sense out of it because it has so many segments. Try to use no more than six segments. Logically combine segments, as in the good example, to make the results clearer.
- In almost all cases, you should include the percentage and label for each segment. Normally these should be next to each segment, connected by leader lines if necessary. Sometimes you have to move the labels manually to prevent them from overlapping.

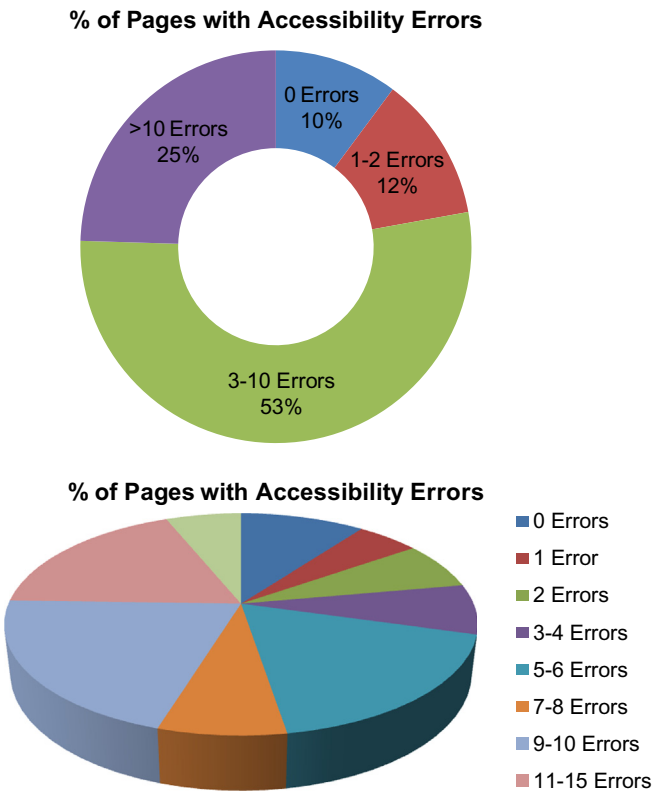


Figure 2.14 Good (top) and bad (bottom) examples of pie or donut charts for the same data. Mistakes in the bad version include too many segments, poor placement of the legend, not showing percentages for each segment, and using 3D, for which the creator of this pie chart should be pummeled with a wet noodle.

2.7.5 Stacked Bar or Column Graphs

Stacked bar graphs (Figure 2.15) are basically multiple pie charts shown in bar or column form. They're appropriate whenever you have a series of data sets, each of which represents parts of the whole. Their most common use in user experience data is to show different task completion states for each task. Here are some key principles for their use.

- Like pie charts, stacked bar graphs are only appropriate when the parts for each item in the series add up to 100%.
- The items in the series are normally categorical (e.g., tasks, participants).

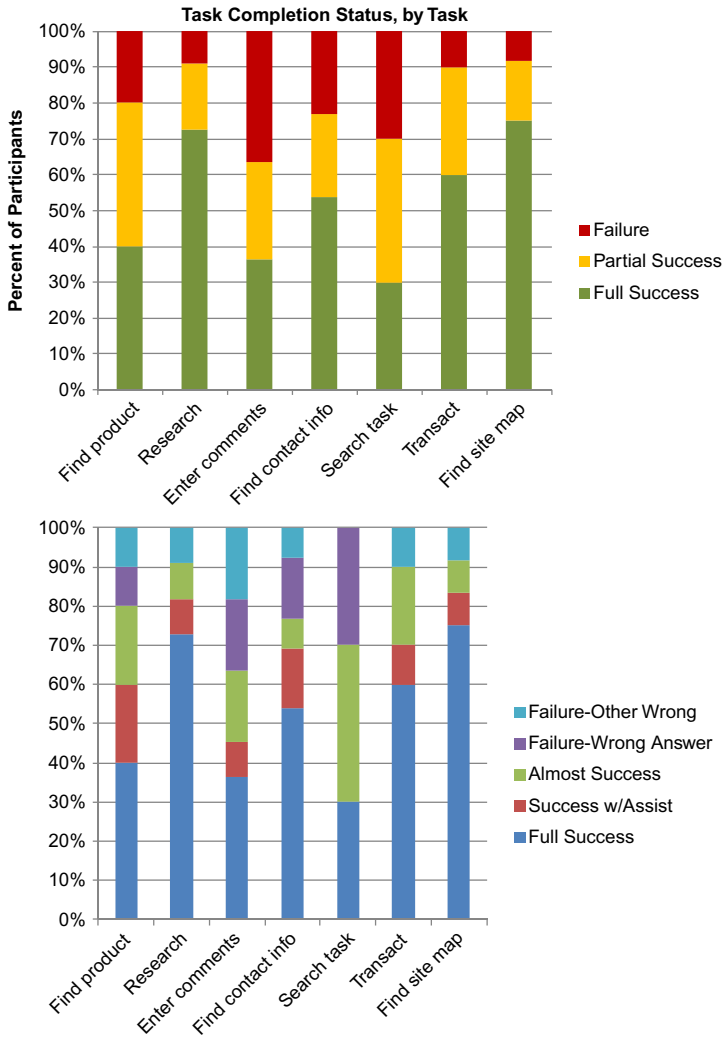


Figure 2.15 Good and bad examples of stacked bar graphs for the same data. Mistakes in the bad version include too many segments, poor color coding, and failing to label the vertical axis.

- Minimize the number of segments in each bar. More than three segments per bar can make it difficult to interpret. Combine segments as appropriate.
- When possible, make use of color-coding conventions that your audience is likely to be familiar with. For many U.S. audiences, green is good, yellow is marginal, and red is bad. Playing off of these conventions can be helpful, as in the good example in Figure 2.15, but don't rely solely on them.

2.8 SUMMARY

In a nutshell, this chapter is about knowing your data. The better you know your data, the more likely you are to answer your research questions clearly. The following are some of the key takeaways from this chapter.

1. When analyzing your results, it's critical to know your data. The specific type of data you have will dictate what statistics you can (and can't) perform.
2. Nominal data are categorical, such as binary task success or males and females. Nominal data are usually expressed as frequencies or percentages. χ^2 tests can be used when you want to learn whether the frequency distribution is random or there is some underlying significance to the distribution pattern.
3. Ordinal data are rank orders, such as a severity ranking of usability issues. Ordinal data are also analyzed using frequencies, and the distribution patterns can be analyzed with a χ^2 test.
4. Interval data are continuous data where the intervals between each point are meaningful but without a natural zero. The SUS score is one example. Interval data can be described by means, standard deviations, and confidence intervals. Means can be compared to each other for the same set of users (paired samples *t* test) or across different users (independent samples *t* test). ANOVA can be used to compare more than two sets of data. Relationships between variables can be examined through correlations.
5. Ratio data are the same as interval but with a natural zero. One example is completion times. Essentially, the same statistics that apply to interval data also apply to ratio data.
6. Any time you can calculate a mean, you can also calculate a confidence interval for that mean. Displaying confidence intervals on graphs of means helps the viewer understand the accuracy of the data and to see quickly any differences between means.
7. When presenting your data graphically, use the appropriate types of graphs. Use bar graphs for categorical data and line graphs for continuous data. Use pie charts or stacked bar graphs when data sum to 100%.

CHAPTER 3

Planning

CONTENTS

Preparation is the key to any successful user experience study. If nothing else, it is hoped this chapter convinces you to plan ahead, particularly where data collection is involved.

Measuring the User Experience. DOI: http://dx.doi.org/10.1016/B978-0-12-415781-1.00003-0

When planning any UX study, a few high-level questions should be answered. First, you need to understand the goals of the study. For example, are you trying to ensure optimal user experience for a new piece of functionality or are you benchmarking the user experience for an existing product? Next, you need to understand the goals of the users. Are users looking to simply complete a task and then stop using the product or will they use the product many times on a daily basis? Knowing both study goals and user goals will lead toward choosing the right metrics.

Many practical details come into play as well. For example, you must decide on the most appropriate evaluation method, how many participants are enough to get reliable feedback, how collecting metrics will impact the timeline and budget, what the best tool is to collect data, and how data will be analyzed. By answering these questions, you will be well prepared to carry out any UX study involving metrics. In the end, you will likely save time and money and have a greater impact on the product.

3.1 STUDY GOALS

The first decision to make when planning a study is how the data will ultimately be used within the product development life cycle. There are essentially two ways to use UX data: formative and summative.

3.1.1 Formative Usability

When running a formative study, a UX specialist is much like a chef who checks a dish periodically while it's being prepared and makes adjustments to impact the end result positively. The chef might add a little salt, then a few more spices, and finally a dash of chili pepper right before serving. The chef is evaluating, adjusting, and reevaluating periodically. The same is true in formative usability. A UX professional, like a chef, evaluates a product or design periodically while it is being created, identifies shortcomings, makes recommendations, and then repeats the process, until, ideally, the product comes out as close to perfect as possible.

What distinguishes formative usability is both the iterative nature of the testing and when it occurs. The goal is to make improvements in the design prior to release. This means identifying or diagnosing the problems, making and implementing recommendations, and then evaluating again. Formative usability is always done before the design has been finalized. In fact, the earlier the formative evaluation, the more impact the usability evaluations will have on the design.

Here are a few key questions you will be able answer with a formative approach:

- What are the most significant usability issues preventing users from accomplishing their goals or resulting in inefficiencies?
- What aspects of the product work well for the users? What do users find frustrating?
- What are the most common errors or mistakes users are making?

- Are improvements being made from one design iteration to the next?
- What usability issues can you expect to remain after the product is launched?

The most appropriate situation to run a formative usability study is when an obvious opportunity to improve the design presents itself. Ideally, the design process allows for multiple usability evaluations. If there's no opportunity to impact the design, then running a formative test is probably not a good use of time or money. Generally, though, selling the value of formative usability shouldn't be a problem. Most people will see the importance of it. The biggest obstacles tend to be a limited budget or time rather than a failure to see the value.

3.1.2 Summative Usability

Continuing with our cooking metaphor, summative usability is about evaluating the dish after it comes out of the oven. The usability specialist running a summative test is like a food critic who evaluates a few sample dishes at a restaurant or perhaps compares the same meal in multiple restaurants. The goal of summative usability is to evaluate how well a product or piece of functionality meets its objectives. Summative testing can also be about comparing several products to each other. Although formative testing focuses on identifying ways of making improvements, summative testing focuses on evaluating against a set of criteria. Summative usability evaluations answer these questions:

- Did we meet the usability goals of the project?
- What is the overall usability of our product?
- How does our product compare against the competition?
- Have we made improvements from one product release to the next?

Running a successful summative usability test should always involve some follow-up activities. Just seeing the metrics is usually not enough for most organizations. Potential outcomes of a summative usability test might be securing funding to enhance functionality on your product, launching a new project to address some outstanding usability issues, or even benchmarking changes to the user experience against which senior managers will be evaluated. We recommend that follow-up actions be planned along with any summative usability study.

FORMATIVE AND SUMMATIVE USABILITY TESTING

The terms *formative* and *summative* were borrowed from the classroom environment, where formative assessment is done on an ongoing basis by a teacher every day in the classroom (think informal observation and "pop quizzes"), while summative assessment is done at the end of some significant period of time (think "final exams"). The earliest application of these terms to usability testing appears to be in a paper presented by Tom Hewett at a conference at the University of York in the United Kingdom (Hewett, 1986). This was also when one of us (Tullis) first met Tom Hewett, mainly because we were the only two Americans at the conference! We've been friends ever since.

3.2 USER GOALS

When planning a usability study, you need to understand the users and what they are trying to accomplish. For example, are users required to use the product every day as part of their job? Are they likely to use the product only once or just a few times? Are they using it frequently as a source of entertainment? It's critical to understand what matters to the user. Does the user simply want to complete a task or is its efficiency the primary driver? Do users care at all about the design aesthetics of the product? All these questions boil down to measuring two main aspects of the user experience: performance and satisfaction.

3.2.1 Performance

Performance is all about what the user actually does in interacting with the product. It includes measuring the degree to which users can accomplish a task or set of tasks successfully. Many measures related to the performance of these tasks are also important, including the time it takes to perform each task, the amount of effort to perform each (such as number of mouse clicks or amount of cognitive effort), the number of errors committed, and the amount of time it takes to become proficient in performing the tasks (learnability). Performance measures are critical for many different types of products and applications, especially those where the user doesn't really have much choice in how they are used (such as a company's internal applications). If users can't perform key tasks successfully when using a product, it's likely to fail. Chapter 4 reviews different types of performance measures.

3.2.2 Satisfaction

Satisfaction is all about what the user says or thinks about his interaction with the product. The user might report that it was easy to use, that it was confusing, or that it exceeded his expectations. The user might have opinions about the product being visually appealing or untrustworthy. User satisfaction has many different aspects. Satisfaction, and many other self-reported metrics, is important for products where the user has some choice in their usage. This would certainly be true for most websites, software applications, and consumer products. Satisfaction metrics are reviewed in Chapter 6.

DO PERFORMANCE AND SATISFACTION ALWAYS CORRELATE?

Perhaps surprisingly, performance and satisfaction don't always go hand-in-hand. We've seen many instances of a user struggling to perform key tasks with an application and then giving it glowing satisfaction ratings. Conversely, we've seen users give poor satisfaction ratings to an application that worked perfectly. So it's important that you look at both performance and satisfaction metrics to get an accurate overall picture of

the user experience. We were curious about the correlations we've seen between two measures of performance (task success and task time) and one measure of satisfaction (task ease rating). We looked at data from 10 online usability studies we've run. The number of participants in each of these studies ranged from 117 to 1036. The correlations between task time and task rating were mostly negative, as you would expect (the longer it takes, the less satisfied you are), but ranged from −0.41 to +0.06. The correlations between task success and task rating were at least all positive, ranging from 0.21 to 0.65. Together, these results suggest that a relationship exists between performance and satisfaction, but not always.

3.3 CHOOSING THE RIGHT METRICS: TEN TYPES OF USABILITY STUDIES

Some of the issues you should consider when choosing metrics for a usability study include the goals of the study and the user, the technology that's available to collect the data, and the budget and time you have to turn around your findings. Because every usability study has unique qualities, we can't prescribe the exact metrics to use for every type of study. Instead, we've identified 10 prototypical categories of usability studies and developed recommendations about metrics for each. The recommendations we offer are simply suggestions that should be considered when running a usability study with a similar set of characteristics. Conversely, metrics that may be essential to your study may not be on the list. Also, we strongly recommend that you explore your raw data and develop new metrics that are meaningful to your project goals. Ten common usability study scenarios are listed in Table 3.1. The metrics that are used commonly or are appropriate for each of the scenarios are indicated. The following sections discuss each of the 10 scenarios.

3.3.1 Completing a Transaction

Many usability studies are aimed at making transactions run as smoothly as possible. These might take the form of a user completing a purchase, registering a new piece of software, or resetting a password. A transaction usually has a well-defined beginning and end. For example, on an e-commerce website, a transaction may start when a user places something in his shopping cart and ends when he has completed the purchase on the confirmation screen.

Perhaps the first metric that you will want to examine is task success. Each task is scored as a success or failure. Obviously the tasks need to have a clear end state, such as reaching a confirmation that the transaction was successful.

Reporting the percentage of participants who were successful is an excellent measure of the overall effectiveness of the transaction. If the transaction involves a website or some live website metrics, such a drop-off rate from the transaction can be very useful. By knowing where users are dropping off, you will be able to focus your attention on the most problematic steps in the transaction.

Usability Study Scenario	Task Success	Task Time	Errors	Efficiency	Learn-ability	Issues-based Metrics	Self-reported Metrics	Behavioral & Physiological Metrics	Combined & Comparative Metrics	Live Website Metrics	Card-Sorting Data
1. Completing a transaction	X			X		X	X			X	
2. Comparing products	X			X			X		X		
3. Evaluating frequent use of the same product	X	X		X	X		X				
4. Evaluating navigation and/or information architecture	X		X	X							X
5. Increasing awareness							X	X		X	
6. Problem discovery						X	X				
7. Maximizing usability for a critical product	X		X	X							
8. Creating an overall positive user experience							X	X			
9. Evaluating the impact of subtle changes										X	
10. Comparing alternative designs	X	X				X	X		X		

Table 3.1 Ten common usability study scenarios and the metrics that may be most appropriate for each.

Calculating issue severity can help narrow down the cause of specific usability problems with a transaction. By assigning a severity to each usability issue, you will be able to focus on the high-priority problems with any transaction. Two types of self-reported metrics are also very useful: likelihood to return and user expectations. In cases where users have a choice of where to perform their transactions, it's important to know what they thought of their experience. One of the best ways to learn this is by asking participants whether they would use the same product again and whether the product met or exceeded their expectations. Efficiency is an appropriate metric when a user has to complete the same transaction many times. Efficiency is often measured as task completion per unit of time.

3.3.2 Comparing Products

It's always useful to know how your product compares to the competition or to previous releases. By making comparisons, you can determine your product's strengths and weaknesses and whether improvements have been made from one release to another. The best way to compare different products or releases is through the use of various metrics. The type of metrics you choose should be based on the product itself. Some products aim to maximize efficiency, whereas others try to create an exceptional user experience.

For most types of products, we recommend three general classes of metrics to get an overall sense of the user experience. First, we recommend looking at some task success measures. Being able to complete a task correctly is essential for most products. It's also important to pay attention to efficiency. Efficiency might be task completion time, number of page views (in the case of some websites), or number of action steps taken. By looking at efficiency, you will get a good sense of how much effort is required to use the product. Some self-reported metrics of satisfaction provide a good summary of the user's overall experience. Satisfaction measures make the most sense with products where people have choices. Finally, one of the best ways to compare the user experience across products is by combined and comparative metrics. This will give an excellent big picture of how the products compare from a UX perspective.

3.3.3 Evaluating Frequent Use of the Same Product

Many products are intended to be used on a frequent or semifrequent basis. Examples might include microwave ovens, mobile phones, web applications used as part of your job, and even the software program we used to write this book. These products need to be both easy to use and highly efficient. The amount of effort required to send a text message or download an application needs to be kept to a minimum. Most of us have very little time or patience for products that are difficult and inefficient to use.

The first metric we would recommend is task time. Measuring the amount of time required to complete a set of tasks will reveal the effort involved. For most products, the faster the completion time, the better. Because some tasks

are naturally more complicated than others, it may be helpful to compare task completion times to expert performance. Other efficiency metrics, such as the number of steps or page views (in the case of some websites), can also be helpful. The time for each step may be short, but the separate decisions that must be made to accomplish a task can be numerous.

Learnability metrics assess how much time or effort is required to achieve maximum efficiency. Learnability can take the form of any of the previous efficiency metrics examined over time. In some situations, consider self-reported metrics, such as awareness and usefulness. By examining the difference between users' awareness and perceived usefulness, you will be able to identify aspects of the product that should be promoted or highlighted. For example, users may have low awareness for some parts of the product, but once they use it, they find out it is extremely useful.

3.3.4 Evaluating Navigation and/or Information Architecture

Many usability studies focus on improving the navigation and/or information architecture. This is probably most common for websites, software programs, mobile applications, consumer electronics, interactive voice responses, or devices. It may involve making sure that users can find what they are looking for quickly and easily, navigate around the product easily, know where they are within the overall structure, and know what options are available to them. Typically, these studies involve the use of wireframes or partially functional prototypes because the navigation and information mechanisms and information architecture are so fundamental to the design that they have to be figured out before almost anything else.

One of the best metrics to evaluate navigation is task success. By giving participants tasks to find key pieces of information (a "scavenger hunt"), you can tell how well the navigation and information architecture works for them. Tasks should touch on all the different areas of the product. An efficiency metric that's useful for evaluating navigation and information architecture is lostness, which looks at the number of steps the participant took to complete a task (e.g., web page visits) relative to the minimum number to complete the task.

Card sorting is a particularly useful method to understand how participants organize information. One type of card-sorting study is called a closed sort, which has participants put items into predefined categories. A useful metric to come from a closed card sort study is the percentage of items placed into the correct category. This metric indicates the intuitiveness of the information architecture. There are some helpful online tools to collect and analyze this type of data, such as Optimal Sort and Treejack (developed by Optimal Workshop in New Zealand).

3.3.5 Increasing Awareness

Not every design that goes through a usability evaluation is about making something easier or more efficient to use. Some design changes are aimed at increasing

awareness of a specific piece of content or functionality. This is certainly true for online advertisements, but it's also true for products that have important but underutilized functionality. There can be many reasons why something is not noticed or used, including some aspect of the visual design, labeling, or placement.

First, we recommend monitoring the number of interactions with the element in question. This is not foolproof, as a participant might notice something but not click on it or interact with it in some way. The opposite would not be very likely: interaction without noticing. Because of this, data can help confirm awareness but not demonstrate lack of awareness. Sometimes it's useful to ask for self-reported metrics about whether the participants noticed or were aware of a specific design element. Measuring noticeability involves pointing out specific elements to the participants and then asking whether they had noticed those elements during the task. Measuring awareness involves asking the participants if they were aware of the feature before the study began. However, data are not always reliable (Albert & Tedesco, 2010). Therefore, we don't recommend that this be your sole measure; you should complement it with other data sources.

Memory is another useful self-reported metric. For example, you can show participants several different elements, only one of which they had actually seen previously, and ask them to choose which one they saw during the task. If they noticed the element, their memory should be better than chance. But perhaps the best way to assess awareness, if you have the technology available, is through the use of behavioral and physiological metrics such as eye-tracking data. Using eye-tracking technology, you can determine the average time spent looking at a certain element, the percentage of participants who looked at it, and even the average time it took to first notice it. Another metric to consider, in the case of websites, is a change in live website data. Looking at how traffic patterns change between different designs will help you determine relative awareness. Simultaneous testing of alternative designs (A/B testing) on live sites is an increasingly common way to measure how small design changes impact user behavior.

3.3.6 Problem Discovery

The goal of problem discovery is to identify major usability issues. In some situations you may not have any preconceived ideas about what the significant usability issues are with a product, but you want to know what annoys users. This is often done for a product that is already built but has not gone through usability evaluation before. A problem discovery study also works well as a periodic checkup to get back in touch with how users are interacting with your product. A discovery study is a little different from other types of usability studies because it is generally open ended. Participants in a problem discovery study may be generating their own tasks, as opposed to being given a list of specific tasks. It's important to strive for realism as much as possible. This might involve

using the live product and their own accounts (if applicable) and performing tasks that are relevant only to them. It might also include evaluating the product in the participants' environments, such as homes or workplaces.

Because they may be performing different tasks and their contexts of use may be different, comparing across participants may be a challenge. Issue-based metrics may be the most appropriate for problem discovery. Assuming you capture all the usability issues, it's fairly easy to convert those data into frequency and type. For example, you might discover that 40% of the usability issues pertain to high-level navigation and 20% of the issues to confusing terminology. Even though the exact problems encountered by each participant might be different, you can still generalize to a higher level category of issue. Examining the frequency and severity of specific issues will reveal how many repeat issues are being observed. Is it a one-time occurrence or part of a recurring problem? By cataloging all the issues and assigning severity ratings, you may come away with a quick-hit list of design improvements.

3.3.7 Maximizing Usability for a Critical Product

Although some products strive to be easy to use and efficient, such as a mobile phone or washing machine, critical products *have* to be easy to use and efficient, such as a defibrillator, voting machine, or emergency exit instructions on an airplane. What differentiates a critical product from a noncritical product is that the entire reason for the critical product's existence is for the user to complete a very important task. Not completing that task will have a significant negative outcome.

Measuring usability for any critical product is essential. Just running a few participants through the lab is rarely good enough. It's important that user performance be measured against a target goal. Any critical product that doesn't meet its target usability goal should undergo a redesign. Because of the degree of certainty you want from your data, you may have to run relatively large numbers of participants in the study. One very important metric is user errors. This might include the number of errors or mistakes made while performing a specific task. Errors are not always easy to tabulate, so special attention must be given to how you define an error. It's always best to be very explicit about what constitutes an error and what doesn't.

Task success is also important. We recommend using a binary approach to success in this situation. For example, the true test of a portable defibrillator machine is that someone can use it successfully by himself. In some cases, you may wish to tie task success to more than one metric, such as completing the task successfully within a specific amount of time and with no errors. Other efficiency metrics are also useful. In the example of the defibrillator machine, simply using it correctly is one thing, but doing so in a timely manner is altogether different. Self-reported metrics are relatively less important with critical products. What users think about their use of the product is much less important than their actual success.

3.3.8 Creating an Overall Positive User Experience

Some products strive to create an exceptional user experience. It's simply not enough to be usable. These products need to be engaging, thought-provoking, entertaining, and maybe even slightly addictive. The aesthetics and visual appeal usually play important roles as well. These are products that you tell a friend about and are not embarrassed to mention at a party. Their popularity usually grows at phenomenal rates. Even though the characteristics of what constitutes a great user experience are subjective, they are still measurable.

Although some performance metrics may be useful, what really matters is what the user thinks, feels, and says with respect to his or her experience. In some ways, this is the opposite perspective of measuring usability of a critical product. If the user struggles a little at first, it may not be the end of the world. What matters is how the user feels at the end of the day. Many self-reported metrics must be considered when measuring the overall user experience.

Satisfaction is perhaps the most common self-reported metric, but it may not always be the best one. Being "satisfied" is usually not enough. One of the most valuable self-reported metrics we've used relates to the participant's expectation. The best experiences are those that exceed a participant's expectations. When the participant says something is much easier, more efficient, or more entertaining than expected, you know you are onto something.

Another set of self-reported metrics relates to future use. For example, you might ask questions related to likelihood to purchase, recommend to a friend, or use in the future. The Net Promoter Score is a widely used metric to measure likelihood of future use. Another interesting set of metrics relates to subconscious reactions that users may be having. For example, if you want to make sure your product is engaging, you can look at physiological metrics. Changes in pupil diameter can be used to gauge the level of arousal or, if you're trying to eliminate stress as much as possible, you can measure heart rate or skin conductance.

3.3.9 Evaluating the Impact of Subtle Changes

Not all design changes have an obvious impact on user behavior. Some design changes are much more subtle, and their impact on user behavior is less clear. Small trends, given enough users, can have huge implications for a large population of users. The subtle changes may involve different aspects of the visual design, such as font choice and size, placement, visual contrast, color, and image choice. Nonvisual design elements, such as subtle changes to content or terminology, can also have an impact on the user experience.

Perhaps the best way to measure the impact of subtle design changes is through live-site metrics from A/B tests. A/B testing involves comparing a control design against an alternative design. For websites, this usually involves diverting a (usually a small) portion of web traffic to an alternative design and comparing metrics such as traffic or purchases to a control design. An online

usability study with a large population can also be very useful. If you don't have access to the technology to run A/B tests or online studies, we recommend using e-mail and online surveys to get feedback from as many representative participants as you can.

3.3.10 Comparing Alternative Designs

One of the most common types of usability studies involves comparing more than one design alternative. Typically, these types of studies take place early in the design process, before any one design has been fully developed. (We often refer to these as "design bakeoffs.") Different design teams put together semi-functional prototypes, and we evaluate each design using a predefined set of metrics. Setting up these studies can be a little tricky. Because the designs are often similar, there is a high likelihood of a learning effect from one design to another. Asking the same participant to perform the same task with all designs usually does not yield reliable results, even when counterbalancing design and task order.

There are two solutions to this problem. You can set up the study as purely between subjects, whereby each participant only uses one design, which provides a clean set of data but requires significantly more participants. Alternatively, you can ask participants to perform the tasks using one primary design (counterbalancing the designs) and then show the other design alternatives and ask for their preference. This way you can get feedback about all the designs from each participant.

The most appropriate metrics to use when comparing multiple designs may be issue-based metrics. Comparing the frequency of high-, medium-, and low-severity issues across different designs will help shed light on which design or designs are more usable. Ideally, one design ends up with fewer issues overall and fewer high-severity issues. Performance metrics such as task success and task times can be useful, but because sample sizes are typically small, these data tend to be of limited value. A couple of self-reported metrics are particularly relevant. One is asking each participant to choose which prototype he would most like to use in the future (as a forced choice comparison). Also, asking each participant to rate each prototype along dimensions, such as ease of use and visual appeal, can be insightful.

3.4 EVALUATION METHODS

One of the great features of collecting UX metrics is that you're not restricted to a certain type of evaluation method (e.g., lab test, online test). Metrics can be collected using almost any kind of evaluation method. This may be surprising because there is a common misperception that metrics can only be collected through large-scale online studies. As you will see, this is simply not the case. Choosing an evaluation method to collect metrics boils down to how many participants are needed and what metrics you're going to use.

3.4.1 Traditional (Moderated) Usability Tests

The most common usability method is a lab test that utilizes a relatively small number of participants (typically 5 to 10). The lab test involves a one-on-one session between a moderator (usability specialist) and a test participant. The moderator asks questions of the participants and gives them a set of tasks to perform on the product in question. The test participant is likely to be thinking aloud as she performs the various tasks. The moderator records the participant's behavior and responses to questions. Lab tests are used most often in formative studies where the goal is to make iterative design improvements. The most important metrics to collect are about issues, including issue frequency, type, and severity. Also, collecting performance data such as task success, errors, and efficiency may also be helpful.

Self-reported metrics can also be collected by having participants answer questions regarding each task or at the conclusion of the study. However, we recommend that you approach performance data and self-reported data very carefully because it's easy to overgeneralize the results to a larger population without an adequate sample size. In fact, we typically only report the frequency of successful tasks or errors. We hesitate even to state the data as a percentage for fear that someone (who is less familiar with usability data or methods) will overgeneralize the data.

Usability tests are not always run with a small number of participants. In some situations, such as comparison tests, you might want to spend some extra time and money by running a larger group of participants (perhaps 10–50 users). The main advantage of running a test with more participants is that as your sample size increases, so does your confidence in your data. Also, this will afford you the ability to collect a wider range of data. In fact, all performance, self-reported, and physiological metrics are fair game. But there are a few metrics that you should be cautious about. For example, inferring website traffic patterns from usability-lab data is probably not very reliable, nor is looking at how subtle design changes might impact the user experience. In these cases, it is better to test with hundreds or even thousands of participants in an online study.

FOCUS GROUPS VERSUS USABILITY TESTS

When some people first hear about usability testing, they believe it is the same as a focus group. But in our experience, the similarity between the two methods begins and ends with the fact that they both involve representative participants. In a focus group, participants commonly watch someone demonstrate or describe a potential product and then react to it. In a usability test, participants actually try to use some version of the product themselves. We've seen many cases where a prototype got rave reviews from focus groups and then failed miserably in a usability test.

3.4.2 Online (Unmoderated) Usability Tests

Online studies involve testing with many participants at the same time. It's an excellent way to collect a lot of usability data in a relatively short amount of time from users who are dispersed geographically. Online studies are usually set up similarly to a lab test in that there are some background or screener questions, tasks, and follow-up questions. Participants go through a predefined script of questions and tasks, and all their data are collected automatically. You can collect a wide range of data, including many performance metrics and self-reported metrics. It may be difficult to collect issues-based data because you're not observing participants directly. But the performance and self-reported data can point to issues, and verbatim comments can help infer their causes. Albert, Tullis, and Tedesco (2010) go into detail about how to plan, design, launch, and analyze an online usability study.

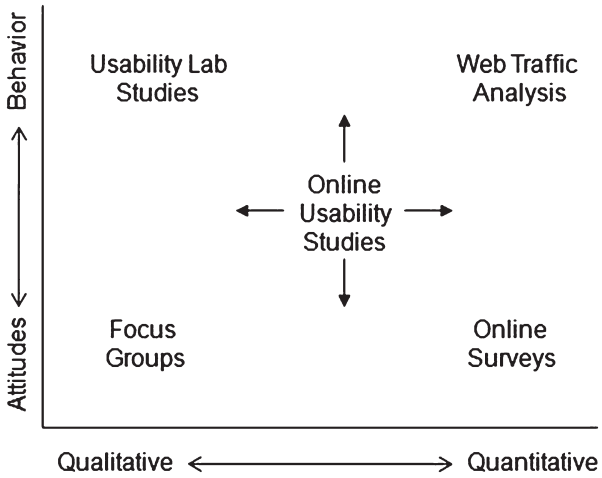


Figure 3.1 How online usability testing tools fit with other common user research methods.

Unlike other methods, online usability studies provide the researcher a tremendous amount of flexibility in the amount and type of data they collect. Online usability studies can be used to collect both qualitative and quantitative data and can focus on either user attitudes or behaviors (see Figure 3.1). The focus of an online study depends largely on the project goals and is rarely limited by the type or amount of data collected. While online studies are an excellent way to collect data, it is less ideal when the UX researcher is trying to gainer a deeper insight into the users' behaviors and motivations.

Online usability tools come in many different flavors; however, there are a few different types of tools that each specialize in a different aspect of the user experience. Figure 3.2 shows the breakdown of different types of online testing tools. These tools are changing constantly, with new ones becoming available every day, with many new features and functionality.

Quantitative-based tools focus on data collection. They typically are set up to collect data from 100+ participants and provide some very nice analytical and reporting functions.

- Full-service tools such as Keynote's WebEffective, Imperium, and Webnographer provide a complete range of features and functionality for carrying out any type of online study, along with support from a team of experts to design an online study and perhaps help with the analysis.
- Self-service tools include Loop11, UserZoom, and UTE. These tools provide a tremendous amount of functionality to the researcher, with minimal support from the vendor. These tools are increasingly becoming more powerful and easy to use, with low-cost options.

- Card-sorting/IA tools help the researcher collect data about how users think about and organize information. Tools such as OptimalSort, TreeJack, and WebSort are very useful, easy to set up, and affordable.
- Surveys are increasingly become useful to the UX researcher. Tools such as Qualtrics, SurveyGizmo, and SurveyMonkey let the researcher embed images into the survey and collect a wide variety of self-reported metrics, along with other useful click metrics.
- Click/Mouse tools such as Chalkmark, Usabilla, ClickTale, and FiveSecondTest let the researcher click data about where users click on a web page or how they move their mouse around. These tools are useful for testing the awareness of key features, intuitiveness of the navigation, or what just grab users' attention the most.

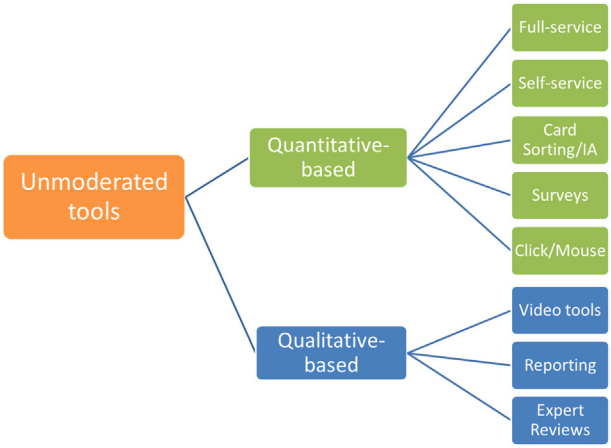


Figure 3.2 A breakdown of the different types of online (unmoderated) testing tools.

Qualitative-based online tools are designed to collect data from a small number of participants who are interacting with a product. These tools are extremely helpful for gaining insight into the nature of the problems that users encounter, as well as provide direction on possible design solutions. There are different types of qualitative-based tools.

- Video tools such as UserTesting.com, Userlytics, and OpenHallway allow you to collect a rich set of qualitative data about the users' experience in using a product in the form of a video file. Observing these videos lets the researcher collect performance metrics, and possibly self-reported metrics, depending on the capabilities of the tool.
- Reporting tools provide the user with an actual report that is typically a list of verbatim comments from users about their experience in using the product. The metrics may be limited, but it is certainly possible to do text analysis of the feedback, looking for common trends or patterns in data.
- Expert review tools such as Concept Feedback provide the user researcher with feedback from a group of "experts" about a product's design and usability. While the feedback is typically qualitative in nature, the researcher might also collect self-reported metrics from each reviewer.

WHICH ONE GOES FIRST? LAB OR ONLINE TEST?

We often get questions about which should go first, a traditional lab study, followed by an online study, or vice versa. There are some pretty strong arguments for both sides.

Lab first, then online	Online first, then lab
Identify/fix "low-hanging fruit"and then focus on remaining tasks with large sample size	Identify the most significant issues online through metrics and then use lab study to gather deeper qualitative understanding of those issues
Generate new concepts, ideas, or questions through lab testing and then test/validate online	Collect video clips or more quotes of users to help bring metrics to life
Validate attitudes/preferences observed in lab testing	Gather all the metrics to validate design—if it tests well, then no need to bring users into the lab

3.4.3 Online Surveys

Many UX researchers think of online surveys strictly for collecting data about preferences and attitudes, and firmly in the camp of market researchers. This is no longer the case. For example, many online survey tools allow you to include images, such as a prototype design, within the body of the survey. Including images within a survey will allow you to collect feedback on visual appeal, page layout, perceived ease of use, and likelihood to use, to name just a few metrics. We have found online surveys to be a quick and easy way to compare different types of visual designs, measure satisfaction with different web pages, and even preferences for various types of navigation schemes. As long as you don't require your participants to interact with the product directly, an online survey may suit your needs.

The main drawback of online surveys is that the data received from each participant are somewhat limited, but that may be offset by the larger number of participants. So, depending on your goals, an online survey tool may be a viable option.

INTERACTING WITH DESIGNS IN AN ONLINE SURVEY

Some online survey tools let participants have some level of interaction with images. This is exciting because it means you can ask participants to click on different areas of a design that are most (or least) useful or where they would go to perform certain tasks. Figure 3.3 is an example of a click map generated from an online survey. It shows different places where participants clicked to begin a task. In addition to collecting data on images, you can also control the time images are displayed. This is very helpful in gathering first impressions of a design or testing whether they see certain visual elements (sometime referred to as a "blink test").

Figure 3.3 An example of a click heat map created with the Qualtrics survey tool.

3.5 OTHER STUDY DETAILS

Many other details must be considered when planning a usability study. Several important issues to consider are budget/timelines, participants, data collection, and data cleanup.

3.5.1 Budgets and Timelines

The cost and time of running a usability study with metrics depend on the evaluation method, metrics chosen, participants, and available tools. It's impossible for us to give even approximate costs or time estimates for any particular type of usability study. The best we can do is to provide a few general rules of thumb for estimating costs and time for some common types of studies. When making these estimates, we recommend that you consider carefully all the variables that go into any usability study and communicate those estimates to business sponsors (or whoever is funding the study) as early as possible. Also, it's wise to add at least a 10% buffer for both costs and time, knowing that there may be some unforeseen costs and delays.

If you are running a formative study with a small number of participants (10 or fewer), collecting metrics should have little, if any, impact on the overall timeline or budget. Collecting and analyzing basic metrics on issue frequency and severity should at most add a few hours to any study. Just allow yourself a little extra time to analyze data once the study is complete. If you're not yet very familiar with collecting these metrics, give yourself some extra time to set up tasks and agree on a method for making severity ratings prior to starting the test. Because

it is a formative study, you should make every attempt to get the findings back to the stakeholders as quickly as possible to influence the next design iteration and not slow down the project.

In the case of running a lab test with a larger number of participants (usually more than a dozen), including metrics may have more of an impact on the budget and timeline. The most significant cost impact may be any additional costs for recruiting and compensating the participants. These costs depend on who they are (e.g., internal to your company versus external), how participants are recruited, and whether the test will be in a local lab or conducted with remote sessions. The most significant impact on the timeline is likely to be the additional time required to run the larger number of participants. Depending on your billing or cost-recovery model, there may also be additional costs because of the increased time for the usability specialists. Keep in mind that you will also need extra time to clean up and analyze the data.

Running an online (unmoderated) study is quite different in terms of costs and time. Typically, about half of the time is usually spent setting up the study, from identifying and validating tasks, creating questions and scales, evaluating the prototypes or designs, identifying and/or recruiting participants, and developing the online script. Unlike traditional lab tests where a lot of time is spent collecting the data, running an online study requires little, if any, time on the part of the usability specialist for data collection. With most online usability testing technologies you simply flip the switch and then monitor the data as they pour in.

The other half of the time is spent cleaning up and analyzing the data. It's very common to underestimate the time required for this. Data are often not in a format that readily allows analysis. For example, you will need to filter out extreme values (particularly when collecting time data), check for data inconsistencies, and code new variables based on the raw data (such as creating top-2-box variables for self-reported data). We have found that we can run an online study in about 100 to 200 person-hours. This includes everything from the planning phase through data collection, analysis, and presentation. The estimate can vary by up to 50% in either direction based on the scope of the study. Many of these details are covered in the book "Beyond the Usability Lab: Conducting Large-scale Online User Experience Studies" (Albert, Tullis, & Tedesco, 2010).

3.5.2 Participants

The specific participants in any usability study have a major impact on its findings. It's critical that you carefully plan how to include the most representative participants as possible in your study. The steps you will go through in recruiting participants are essentially the same whether you're collecting metrics or not.

The first step is to identify recruiting criteria that will be used to determine whether a specific person is eligible to participate in the study. Criteria should be as specific as possible to reduce the possibility of recruiting someone who does not fit the profile(s). We often recruit participants based on many characteristics, including their experience with the web, years away from retirement, or experience

with various financial transactions. As part of identifying criteria, you may segment participant types. For example, you may recruit a certain number of new participants as well as ones who have experience with the existing product.

After deciding on the types of participants you want, you need to figure out how many you need. As you saw in Section 2.1.2, the number of participants needed for a usability test is one of the most hotly debated issues in the field. Many factors enter into the decision, including the diversity of the user population, the complexity of the product, and the specific goals of the study. As a general rule of thumb, however, testing with about six to eight participants for each iteration in a formative study works well. The most significant usability findings will be observed with the first six or so participants. If there are distinct groups of users, it's helpful to have at least four from each group.

For summative usability studies, we recommend having data from 50 to 100 representative users for each distinct user group. If you're in a crunch, you can go as low as 30 participants, but the variance in the data will be quite high, making it difficult to generalize the findings to a broader population. In the case of studies where you are testing the impact of potentially subtle design changes, having at least 100 participants for each distinct user group is advisable.

After determining the sample size, you will need to plan the recruiting strategy. This is essentially how you are actually going to get people to participate in the study. You might generate a list of possible participants from customer data and then write a screener that a recruiter uses when contacting potential participants. You might send out requests to participate via e-mail distribution lists. You can screen or segment participants through a series of background questions or you might decide to use a third party to handle all of the recruiting. Some of these companies have quite extensive user panels to draw on. Other options exist, such as posting an announcement on the web or e-mailing a specific group of potential participants. Different strategies work for different organizations.

DOES GEOGRAPHY MATTER?

One of the most common questions we get from our clients is whether we need to recruit participants from different cities, regions, and countries. The answer is usually no—geography doesn't matter when collecting usability data. It's very unlikely that participants in New York are going to have a different set of issues than participants in Chicago, London, or even Walla Walla, Washington. But there are some exceptions. If the product you are evaluating has a large corporate presence in one location, it may bias responses. For example, if you want to test Walmart.com in their hometown of Benton, Arkansas, you might find it hard to get a neutral, unbiased set of results. Also, location can have an impact on user goals for some products. For example, if you are evaluating an e-commerce clothing website, you might collect different data from participants in urban or rural settings, or participants in different countries, where the needs and preferences can vary quite a bit. Even when it doesn't really make sense to test in different locations, some clients still choose to test products in different regions, simply to prevent senior management from questioning the validity of the results.

3.5.3 Data Collection

It's important to think about how the data are going to be collected. You should plan well in advance how you are going to capture all the data that you need for your study. The decisions you make may have a significant impact on how much work you have to do further down the road when you begin analysis.

In the case of a lab test with a fairly small number of participants, Excel probably works as well as anything for collecting data. Make sure you have a template in place for quickly capturing the data during the test. Ideally, this is not done by the moderator but by a note taker or someone behind the scenes who can enter data quickly and easily. We recommend that data be entered in numeric format as much as possible. For example, if you are coding task success, it is best to code it as a "1" (success) and "0" (failure). Data entered in a text format will eventually have to be converted, with the exception of verbatim comments.

The most important thing when capturing data is for everyone on the usability team to know the coding scheme extremely well. If anyone starts flipping scales (confusing the high and low values) or does not understand what to enter for certain variables, you will have to either recode or throw data out. We strongly recommend that you offer training to others who will be helping you collect data. Just think of it as inexpensive insurance to make sure you end up with clean data.

For studies involving larger numbers of participants, consider using a data-capture tool. If you are running an online study, data are typically collected automatically. You should also have the option of downloading the raw data into Excel or various statistical programs such as SAS and SPSS.

3.5.4 Data Cleanup

Data rarely come out in a format that is instantly ready to analyze. Some sort of cleanup is usually needed to get your data in a format that allows for quick and easy analysis. Data cleanup may include the following.

- **Filtering data**. You should check for extreme values in the data set. The most likely culprit will be task completion times (in the case of online studies). Some participants may have gone out to lunch in the middle of the study so their task times will be unusually large. Also, some participants may have taken an impossibly short amount of time to complete the task. This is likely an indicator that they were not truly engaged in the study. Some general rules for how to filter time data are included in Section 4.2. You should also consider filtering out data for participants who do not reflect your target audience or where outside factors impacted the results. We've had more than a few usability testing sessions interrupted by a fire drill!
- **Creating new variables**. Building on the raw data set is very useful. For example, you might want to create a top-2-box variable for self-reported rating scales by counting the number of participants who gave one of the

two highest ratings. Perhaps you want to aggregate all success data into one overall success average representing all tasks. Or you might want to combine several metrics using a *z*-score transformation (described in Section 8.1.3) to create an overall usability score.

- **Verifying responses**. In some situations, particularly for online studies, participant responses may need to be verified. For example, if you notice that a large percentage of participants are all giving the same wrong answer, this should be investigated.
- **Checking consistency**. It's important to make sure that data are captured properly. A consistency check might include comparing task completion times and successes to self-reported metrics. If many participants completed a task in a relatively short period of time and were successful but gave the task a very low rating, there may be a problem with either how the data were captured or participants confusing the scales of the question. This is quite common with scales involving self-reported ease of use.
- **Transferring data**. It's common to capture and clean up data using Excel, then use another program such as SPSS to run some statistics (although all the basic statistics can be done with Excel), and then move back to Excel to create the charts and graphs.

Data cleanup can take anywhere from an hour to a couple of weeks. For simple usability studies, with just a couple of metrics, cleanup should be very quick. Obviously, the more metrics you are dealing with, the more time it will take. Also, online studies can take longer because more checks are being done. You want to make sure that the technology is coding all the data correctly.

3.6 SUMMARY

Running a usability study including metrics requires some planning. The following are some key points to remember.

- The first decision you must make is whether you are going to take a formative or summative approach. A formative approach involves collecting data to help improve the design before it is launched or released. It is most appropriate when you have an opportunity to impact the design of the product positively. A summative approach is taken when you want to measure the extent to which certain target goals were achieved. Summative testing is also sometimes used in competitive usability studies.
- When deciding on the most appropriate metrics, two main aspects of the user experience to consider are performance and satisfaction. Performance metrics characterize what the user does and include measures such as task success, task time, and the amount of effort required to achieve a desired outcome. Satisfaction metrics relate to what users think or feel about their experience.
- Budgets and timelines need to be planned out well in advance when running any usability studies involving metrics. If you are running a formative

study with a relatively small number of participants, collecting metrics should have little, if any, impact on the overall timeline or budget. Otherwise, special attention must be paid to estimating and communicating costs and time for larger scale studies.

- Three general types of evaluation methods are used in collecting usability data. Lab tests with small numbers of participants are best in formative testing. These studies typically focus on issues-based metrics. Lab tests with large numbers of participants (more than a dozen) are best to capture a combination of qualitative and quantitative data. These studies usually measure different aspects of performance, such as success, completion time, and errors. Online studies with very large numbers of participants (more than 100) are best to examine subtle design changes and preference.

CHAPTER 4

Performance Metrics

CONTENTS

Anyone who uses technology has to interact with some type of interface to accomplish their goals. For example, a user of a website clicks on different links, a user of a word-processing application enters information via a keyboard, or a user of a video game system pushes buttons on a remote control or waves a controller in the air. No matter the technology, users are behaving or interacting with a product in some way. These behaviors form the cornerstone of performance metrics.

Every type of user behavior is measurable in some way. Behaviors that achieve a goal for a user are especially important to the user experience. For example,

Measuring the User Experience. DOI: http://dx.doi.org/10.1016/B978-0-12-415781-1.00004-2

you can measure whether users clicking through a website (behavior) found what they were looking for (goal). You can measure how long it took users to enter and format a page of text properly in a word-processing application or how many buttons users pressed in trying to cook a frozen dinner in a microwave. All performance metrics are calculated based on specific user behaviors.

Performance metrics not only rely on user behaviors, but also on the use of scenarios or tasks. For example, if you want to measure success, the user needs to have specific tasks or goals in mind. The task may be to find the price of a sweater or submit an expense report. Without tasks, performance metrics aren't possible. You can't measure success if the user is only browsing a website aimlessly or playing with a piece of software. How do you know if he or she was successful? But this doesn't mean that the tasks must be something arbitrary given to the users. They could be whatever the users came to a live website to do or something that the participants in a usability study generate themselves. Often we focus studies on key or basic tasks.

Performance metrics are among the most valuable tools for any usability professional. They're the best way to evaluate the effectiveness and efficiency of many different products. If users are making many errors, you know there are opportunities for improvement. If users are taking four times longer to complete a task than what was expected, efficiency can be improved greatly. Performance metrics are the best way of knowing how well users are actually using a product.

Performance metrics are also useful in estimating the *magnitude* of a specific usability issue. Many times it's not enough to know that a particular issue exists. You probably want to know *how many* people are likely to encounter the same issue after the product is released. For example, by calculating a success rate that includes a confidence interval, you can derive a reasonable estimate of how big a usability issue really is. By measuring task completion times, you can determine what percentage of your target audience will be able to complete a task within a specified amount of time. If only 20% of the target users are successful at a particular task, it should be fairly obvious that the task has a usability problem.

Senior managers and other key stakeholders on a project usually sit up and pay attention to performance metrics, especially when they are presented effectively. Managers will want to know how many users are able to complete a core set of tasks successfully using a product. They see these performance metrics as a strong indicator of overall usability and a potential predictor of cost savings or increases in revenue.

Performance metrics are not the magical elixir for every situation. Similar to other metrics, an adequate sample size is required. Although the statistics will work whether you have 2 or 100 users, your confidence level will change dramatically depending on the sample size. If you're only concerned about identifying the lowest of the low-hanging fruit, such as identifying only the most severe problems with a product, performance metrics are probably not a good use of time or money. But if you need a more fine-grained evaluation and have the time to collect data from 10 or more users, you should be able to derive meaningful performance metrics with reasonable confidence levels.

Avoid overrelying on performance metrics when your goal is simply to uncover basic usability problems. When reporting task success or completion time, it can be easy to lose sight of the underlying issues behind the data. Performance metrics tell the *what* very effectively but not the *why*. Performance data can point to tasks or parts of an interface that were particularly problematic for users, but they don't identify the causes of the problems. You will usually want to supplement it with other data, such as observational or self-reported data, to better understand why they were problems and how they might be fixed.

Five basic types of performance metrics are covered in this chapter.

1. *Task success* is perhaps the most widely used performance metric. It measures how effectively users are able to complete a given set of tasks. Two different types of task success are reviewed: binary success and levels of success. Of course you can also measure task failure.
2. *Time on task* is a common performance metric that measures how much time is required to complete a task.
3. *Errors* reflect the mistakes made during a task. Errors can be useful in pointing out particularly confusing or misleading parts of an interface.
4. *Efficiency* can be assessed by examining the amount of effort a user expends to complete a task, such as the number of clicks in a website or the number of button presses on a mobile phone.
5. *Learnability* is a way to measure how performance improves or fails to improve over time.

4.1 TASK SUCCESS

The most common usability metric is task success, which can be calculated for practically any usability study that includes tasks. It's almost a universal metric because it can be calculated for such a wide variety of *things* being tested—from websites to kitchen appliances. As long as the user has a reasonably well-defined task, you can measure success.

Task success is something that almost anyone can relate to. It doesn't require elaborate explanations of measurement techniques or statistics to get the point across. If your users can't complete their tasks, then you know something is wrong. Seeing users fail to complete a simple task can be pretty compelling evidence that something needs to be fixed.

To measure task success, each task that users are asked to perform must have a clear end state or goal, such as purchasing a product, finding the answer to a specific question, or completing an online application form. To measure success, you need to know what constitutes success, so you should define success criteria for each task prior to data collection. If you don't predefine criteria, you run the risk of constructing a poorly worded task and not collecting clean success data. Here are examples of two tasks with clear and not-so-clear end states:

- Find the 5-year gain or loss for IBM stock (clear end state)
- Research ways to save for your retirement (not a clear end state)

Although the second task may be perfectly appropriate in certain types of usability studies, it's not appropriate for measuring task success.

The most common way of measuring success in a lab-based usability test is to have the user articulate the answer verbally after completing the task. This is natural for the user, but sometimes it results in answers that are difficult to interpret. Users might give extra or arbitrary information that makes it difficult to interpret the answer. In these situations, you may need to probe the users to make sure they actually completed the task successfully.

Another way to collect success data is by having users provide their answers in a more structured way, such as using an online tool or paper form. Each task might have a set of multiple-choice responses. Users might choose the correct answer from a list of four to five distracters. It's important to make the distracters as realistic as possible. Try to avoid write-in answers if possible. It's much more time-consuming to analyze each write-in answer, and it may involve judgment calls, thereby adding more noise to the data.

In some cases the correct solution to a task may not be verifiable because it depends on the user's specific situation, and testing is not being performed in person. For example, if you ask users to find the balance in their savings account, there's no way to know what that amount really is unless we're sitting next to them while they do it. So in this case, you might use a proxy measure of success. For example, you could ask users to identify the title of the page that shows their balance. This works well as long as the title of the page is unique and obvious and you're confident that they are able to actually see the balance if they reached this page.

4.1.1 Binary Success

Binary success is the simplest and most common way of measuring task success. Users either completed a task successfully or they didn't. It's kind of like a "pass/fail" course in college. Binary success is appropriate to use when the success of the product depends on users completing a task or set of tasks. Getting close doesn't count. The only thing that matters is that they accomplish a goal with their tasks. For example, when evaluating the usability of a defibrillator device (to resuscitate people during a heart attack), the only thing that matters is being able to use it correctly without making any mistakes within a certain amount of time. Anything less would be a major problem, especially for the recipient! A less dramatic example might be a task that involves a goal of purchasing a book on a website. Although it may be helpful to know where in the process someone failed, if your company's revenue depends on selling those books, that's what really matters.

Each time users perform a task, they should be given a "success" or "failure" score. Typically, these scores are in the form of 1's (for success) and 0's (for failure). (Analysis is easier if you assign a numeric score rather than a text value of "success" or "failure.") By using a numeric score, you can easily calculate the percent correct as well as other statistics you might need. Simply calculate the average of the 1's and 0's to determine the percent correct. Assuming you have more

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Average
Participant 1	1	1	1	0	1	1	1	1	0	1	80%
Participant 2	1	0	1	0	1	0	1	0	0	1	50%
Participant 3	1	1	0	0	0	0	1	0	0	0	30%
Participant 4	1	0	0	0	1	0	1	1	0	0	40%
Participant 5	0	0	1	0	0	1	0	0	0	0	20%
Participant 6	1	1	1	1	1	0	1	1	1	1	90%
Participant 7	0	1	1	0	0	1	1	1	0	1	60%
Participant 8	0	0	0	0	1	0	0	0	0	1	20%
Participant 9	1	0	0	0	0	1	1	1	0	1	50%
Participant 10	1	1	0	1	1	1	1	1	0	1	80%
Average	70%	50%	50%	20%	60%	50%	80%	60%	10%	70%	52.0%

Table 4.1 Task success data for 10 participants and 10 tasks.

than one participant and more than one task, there are always two ways you can calculate task success:

- By looking at the average success rate for each *task* across the participants
- By looking at the average success rate for each *participant* across the tasks

As an example, consider the data in Table 4.1. Averages across the bottom represent the task success rates for each *task*. Averages along the right represent the success rates for each *participant*. As long as there are no missing data, the averages of those two sets of averages will always be the same.

DOES TASK SUCCESS ALWAYS MEAN *FACTUAL* SUCCESS?

The usual definition of task success is achieving some factually correct or clearly defined state. For example, if you're using the NASA site to find who the Commander of Apollo 12 was, there's a single factually correct answer (Charles "Pete" Conrad, Jr.). Or if you're using an e-commerce site to purchase a copy of "Pride and Prejudice," then purchasing that book would indicate success. But in some cases, perhaps what's important is not so much reaching a factual answer or achieving a specific goal, but rather users *being satisfied* they have achieved a certain state. For example, just before the 2008 U.S. presidential election, we conducted an online study comparing the websites of the two primary candidates, Barack Obama and John McCain. The tasks included things such as finding the candidate's position on Social Security. Task success was measured by self-report only (Yes I Found It, No I Didn't Find, or I'm Not Sure), as for this kind of site the important thing is whether users *believe* they found the information they were looking for. Sometimes it can be very interesting to look at the correlation between *perceived* success and *factual* success.

The most common way to analyze and present binary success rates is by task. This involves simply presenting the percentage of participants who completed each task successfully. Figure 4.1 shows the task success rates for the data in Table 4.1. This approach is most useful when you want to compare success rates for each task. You can then do a more detailed analysis of each task by looking at the specific problems to determine what changes may be needed to address them. For example, Figure 4.1 shows that the "Find Category" and "Checkout" tasks appear to be problematic.

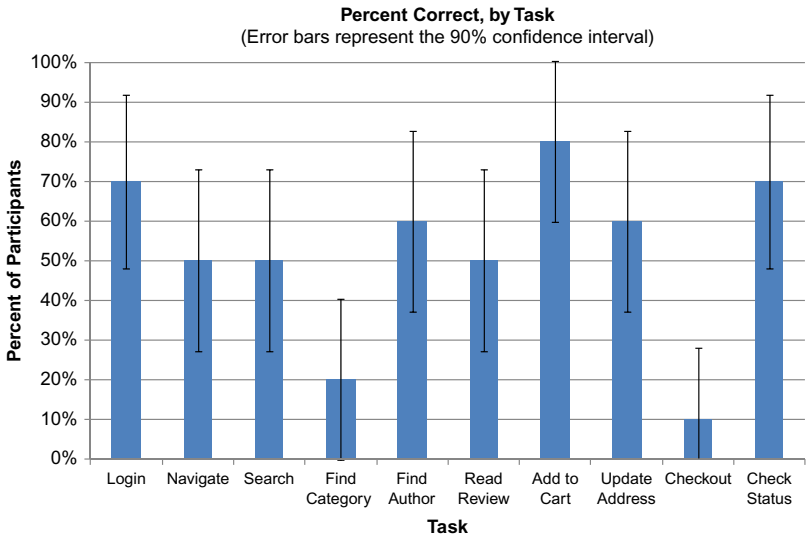


Figure 4.1 Task success rates for the data in Table 4.1, including a 90% confidence interval for each task.

TYPES OF TASK FAILURE

There are many different ways in which a participant might fail a task, but they tend to fall into a few categories:

- **Giving up**—Participants indicate that they would not continue with the task if they were doing this on their own.
- **Moderator "calls" it**—The study moderator stops the task because it's clear that the participant is not making any progress or is becoming especially frustrated.
- **Too long**—The participant completed the task but not within a predefined time period. (Certain tasks are only considered successful if they can be accomplished within a given time period.)
- **Wrong**—Participants thought that they completed the task successfully, but they actually did not (e.g., concluding that Neil Armstrong was the Commander of Apollo 12 instead of Pete Conrad). In many cases, these are the most serious kinds of task failures because the participants don't realize they are failures. In the real world, the consequences of these failures may not become clear until much later (e.g., you intended to order a copy of "Pride and Prejudice" but are rather surprised when "Pride and Prejudice and Zombies" shows up in the mail several days later!)

Another common way of looking at binary success is by user or type of user. As always in reporting usability data, you should be careful to maintain the anonymity of users in the study using numbers or other nonidentifiable descriptors. The main value of looking at success data from a user perspective is that you can identify different groups of users who perform differently or encounter different sets of problems. Here are some of the common ways to segment different users:

- Frequency of use (infrequent users versus frequent users)
- Previous experience using the product
- Domain expertise (low-domain knowledge versus high-domain knowledge)
- Age group

Task success for different groups of participants is also used when each group is given a different design to work with. For example, participants in a usability study might be assigned randomly to use either Version A or Version B of a prototype website. A key comparison will be the average task success rate for participants using Version A vs those using Version B.

If you have a relatively large number of users in a usability study, it may be helpful to present binary success data as a frequency distribution (Figure 4.2). This is a convenient way to visually represent the variability in binary task success data. For example, in Figure 4.2, six users in the evaluation of the original website completed 61 to 70% of the tasks successfully, one completed fewer than 50%, and only two completed as many as 81 to 90%. In a revised design, six users had a success rate of 91% or greater, and no user had a success rate below 61%. Illustrating that the two distributions of task success barely overlap is a much more dramatic way of showing the improvement across the iterations than simply reporting the two means.

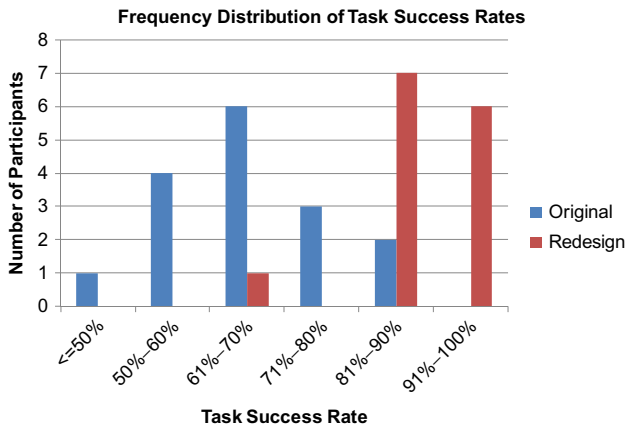


Figure 4.2 Frequency distributions of binary success rates from usability tests of the original version of a website and the redesigned version (data from LeDoux, Connor, & Tullis, 2005).

CALCULATING CONFIDENCE INTERVALS FOR BINARY SUCCESS

One of the most important aspects of analyzing and presenting binary success is including confidence intervals. Confidence intervals are essential because they reflect your trust or confidence in the data. In most usability studies, binary success data are based on relatively small samples (e.g., 5 to 20 users). Consequently, the binary success metric may not be as reliable as we would like it to be. For example, if 4 out of 5 users completed a task successfully, how confident can we be that 80% of the larger population of users will be able to complete that task successfully? Obviously, we would be more confident if 16 out of 20 users completed the task successfully and even more confident if 80 out of 100 did.

Fortunately, there is a way to take this into account. Binary success rates are essentially proportions: the proportion of users who completed a given task successfully. For example, if 5 of the 10 participants completed a task, the success rate is 5/10 = 0.5. The appropriate way to calculate a confidence interval for a proportion like this is to use a binomial confidence interval. Several methods are available for calculating binomial confidence intervals, such as the Wald Method and the Exact Method. But as Sauro and Lewis (2005) have shown, many of those methods are too conservative or too liberal in their calculation of the confidence interval when dealing with the small sample sizes we commonly have in usability tests. They found that a modified version of the Wald Method, called the Adjusted Wald, yielded the best results when calculating a confidence interval for task success data.

CONFIDENCE INTERVAL CALCULATOR

Jeff Sauro has provided a very useful calculator for determining confidence intervals for binary success on his website http://www.measuringusability.com/wald. By entering the total number of people who attempted a given task and how many of them completed it successfully, this tool will perform the Wald, Adjusted Wald, Exact, and Score calculations of the confidence interval for the mean task completion rate automatically. You can choose to calculate a 99, 95, or 90% confidence interval. If you really want to calculate confidence intervals for binary success data yourself, the details are included on our website.

If 4 out of 5 users completed a given task successfully, the Adjusted Wald Method yields a 95% confidence interval for that task completion rate ranging from 36 to 98%—a rather large range! However, if 16 out of 20 users completed the task successfully (the same proportion), the Adjusted Wald Method yields a 95% confidence interval of 58 to 93%. If you *really* got carried away and ran a usability test with 100 participants, of whom 80 completed the task successfully, the 95% confidence interval would be 71 to 87%. As is almost always the case with confidence intervals, larger sample sizes yield smaller (or more accurate) intervals.

4.1.2 Levels of Success

Identifying levels of success is useful when there are reasonable shades of gray associated with task success. The user receives some value from completing a task partially. Think of it as partial credit on a homework assignment if you showed your work, even though you got the wrong answer. For example, assume that a user's task is to find the least expensive digital camera with at least 10 megapixel resolution, at least 12× optical zoom, and weighing no more than 3 pounds. What if the user found a camera that met most of those criteria but had a 10×

optical zoom instead of 12×? According to a strict binary success approach, that would be a failure. But you're losing some important information by doing that. The user actually came very close to completing the task successfully. In some cases, this might be acceptable to a user. For some types of products, coming close to fully completing a task may provide value to the user. Also, it may be helpful for you to know why some users failed a task or with which particular tasks users needed help.

SHOULD YOU INCLUDE TASKS THAT CAN'T BE DONE?

An interesting question is whether a usability study should include tasks that can't be done using the product being testing. For example, assume you're testing an online bookstore that only carries mystery novels. Would it be appropriate to include a task that involves trying to find a book that the store doesn't carry, such as a science-fiction novel? If one of the goals of the study is to determine how well users can determine what the store does *not* carry, we think it could make sense. In the real world, when you come to a new website, you don't automatically know everything that can or can't be done using the site. A well-designed site not only makes it clear what *is* available on the site, but also what's *not* available. However, when tasks are presented in a usability study, there's probably an implicit understanding that they *can* be done. So we think if you do include tasks that can't be done, you should make it clear up front that some of the tasks may not be possible.

HOW TO COLLECT AND MEASURE LEVELS OF SUCCESS

Collecting and measuring levels of success data is very similar to binary success data except that you must define the various levels. There are a couple of approaches to levels of success:

- Based on the user's experience in completing a task. Some users might struggle or require assistance, while others complete their tasks without any difficulty.
- Based on the users accomplishing the task in different ways. Some users might accomplish the task in an optimal way, while others might accomplish it in ways that are less than optimal.

Levels of success based on the degree to which users complete a task typically have between three and six levels. A common approach is to use three levels: complete success, partial success, and complete failure.

Levels of success data are almost as easy to collect and measure as binary success data. It just means defining what you mean by "complete success" and by "complete failure." Anything in between is considered a partial success. A more

granular approach is to break out each level according to whether assistance was given or not. Below is an example of six different levels of completion:

- Complete success
    - With assistance
    - Without assistance
- Partial success
    - With assistance
    - Without assistance
- Failure
    - User thought it was complete, but it wasn't
    - User gave up

If you do decide to use levels of success, it's important to clearly define the levels beforehand. Also, consider having multiple observers independently assess the levels for each task and then reach a consensus.

A common issue when measuring levels of success is deciding what constitutes "giving assistance" to the participant. Here are some examples of situations we define as giving assistance:

- Moderator takes the participant back to a home page or resets to an initial (pretask) state. This form of assistance may reorient the participant and help avoid certain behaviors that initially resulted in confusion.
- Moderator asks the participant probing questions or restates the task. This may cause the user to think about her behavior or choices in a different way.
- Moderator answers a question or provides information that helps the participant complete the task.
- Participant seeks help from an outside source. For example, the participant calls a phone representative, uses some other website, consults a user manual, or accesses an online help system.

Level of success can also be examined in terms of the user experience. We commonly find that some tasks are completed without any difficulty, whereas others are completed with minor or major problems along the way. It's important to distinguish between these different experiences. A four-point scoring method can be used for each task:

1 = No problem. The user completed the task successfully without any difficulty or inefficiency.
2 = Minor problem. The user completed the task successfully but took a slight detour. He made one or two small mistakes but recovered quickly and was successful.
3 = Major problem. The user completed the task successfully but had major problems. She struggled and took a major detour in her eventual successful completion of the task.
4 = Failure/gave up. The user provided the wrong answer or gave up before completing the task or the moderator moved on to the next task before successful completion.

When using this scoring system, it's important to remember that these data are ordinal (see Chapter 2). Therefore, you should not report an average score. Rather, present the data as frequencies for each level of completion. This scoring system is relatively easy to use, and we usually see agreement on the various levels by different usability specialists observing the same interactions. Also, you can aggregate the data into a binary success rate if you need to. Finally, this scoring system is usually easy to explain to your audience. It's also helpful to focus on the 3's and 4's as part of design improvements; there's usually no need to worry about the 1's and 2's.

HOW TO ANALYZE AND PRESENT LEVELS OF SUCCESS

In analyzing levels of success, the first thing you should do is create a stacked bar chart. This will show the percentage of users who fall into each category or level, including failures. Make sure that the bars add up to 100%. Figure 4.3 is an example of a common way to present levels of success.

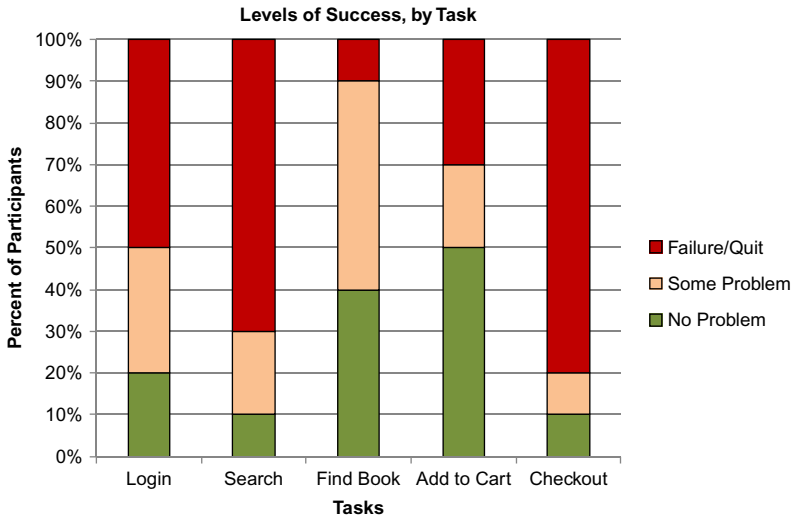


Figure 4.3 Stacked bar chart showing different levels of success based on task completion.

4.1.3 Issues in Measuring Success

Obviously, an important issue in measuring task success is simply how you define whether a task was successful. The key is to clearly define beforehand what criteria are for completing each task *successfully*. Try to think through the various situations that might arise for each task and decide whether or not they constitute success. For example, is a task successful if the user finds the right answer but reports it in the wrong format? Also, what happens if he reports the right answer but then restates his answer incorrectly? When unexpected situations arise during the test, make note of them and try to reach a consensus among the observers afterward about those cases.

One issue that commonly arises during a usability evaluation is how or when to end a task if the user is not successful. In essence, this is the "stopping rule" for unsuccessful tasks. Here are some of the common approaches to ending an unsuccessful task:

1. Tell the users at the beginning of the session that they should continue to work on each task until they either complete it or reach the point at which, in the real world, they would give up or seek assistance (from technical support, a colleague, etc.).
2. Apply a "three strikes and you're out" rule. This means that the users get three attempts (or whatever number you decide) to complete a task before you stop them. The main difficulty with this approach is defining what is meant by an "attempt." It could be three different strategies, three wrong answers, or three different "detours" in finding specific information. However you define it, there will be a considerable amount of discretion on behalf of the moderator or scorer.
3. "Call" the task after a predefined amount of time has passed. Set a time limit, such as 5 minutes. After the time has expired, move on to the next task. In most cases, it is better not to tell the user that you are timing them. By doing so, you create a more stressful, "test-like" environment.

Of course, you always have to be sensitive to the user's state in any usability test and potentially end a task (or even the session) if you see that the user is becoming particularly frustrated or agitated.

4.2 TIME ON TASK

Time on task (sometimes referred to as task completion time or simply task time) is a good way to measure the efficiency of a product. In most situations, the faster a user can complete a task, the better the experience. In fact, it would be pretty unusual for a user to complain that a task took less time than expected. But there are some exceptions to the assumption that faster is better. One could be a game, where the user doesn't want to finish it too quickly. The main purpose of most games is the experience itself rather than quick completion of a task. Another exception may be e-learning. For example, if you're putting together an online training course, slower may be better. Users may retain more if they spend more time completing the tasks rather than rushing through the course.

TIME ON TASK VS WEB SESSION DURATION

Our assertion that faster task times are generally better seems at odds with the view from web analytics that you want longer page view or session durations. From a web-analytics perspective, longer page-view durations (the amount of time each user is viewing each page) and longer session durations (the amount of time each user is spending on the site) are generally considered good things. The argument is that they represent greater

"engagement" with the site, or the site is considered "stickier". Part of the reason that our assertion seems at odds with that perspective is that we don't agree with it. Session and page-view duration are examples of metrics that are from the perspective of the site owner rather than the user. We would still argue that users generally want to be spending *less* time on the site, not *more*. But there is a way in which the two viewpoints might be reconciled. Perhaps a goal of a site might be to get users to perform more in-depth or complex tasks rather than just superficial ones (e.g., rebalancing their financial portfolio instead of just checking their balances). More complex tasks will generally yield longer times on the site *and* longer task times than superficial tasks.

4.2.1 Importance of Measuring Time on Task

Time on task is particularly important for products where tasks are performed repeatedly by the user. For example, if you're designing an application for use by customer service representatives of an airline, the time it takes to complete a phone reservation would be an important measure of efficiency. The faster the airline agent can complete a reservation, presumably the more calls that can be handled and, ultimately, the more money can be saved. The more often a task is performed by the same user, the more important efficiency becomes. One of the side benefits of measuring time on task is that it can be relatively straightforward to calculate cost savings due to an increase in efficiency and then derive an actual return on investment (ROI) . Calculating ROI is discussed in more detail in Chapter 9.

4.2.2 How to Collect and Measure Time on Task

Time on task is simply the time elapsed between the start of a task and the end of a task, usually expressed in minutes and seconds. Logistically, time on task can be measured in many different ways. The moderator or note taker can use a stopwatch or any other time-keeping device that can measure at the minute and second levels. Using a digital watch or application on a smartphone, you could simply record the start and end times. When video recording a usability session, we find it's helpful to use the time-stamp feature of most recorders to display the time and then to mark those times as the task start and stop times. If you choose to record time on task manually, it's important to be very diligent about when to start and stop the clock and/or record the start and stop times. It may also be helpful to have two people record the times and to be unobtrusive in recording the times.

AUTOMATED TOOLS FOR MEASURING TIME ON TASK

A much easier and less error-prone way of recording task times is using an automated tool. Some tools that can assist in logging of task times include the following:

- Usability Activity Log from Bit Debris Solutions (http://www.bitdebris.com/category/Usability-Activity-Log.aspx)

- The Observer XT from Noldus Information Technology (http://www.noldus.com/human-behavior-research/products/the-observer-xt)
- Ovo Logger from Ovo Studios (http://www.ovostudios.com/ovologger.asp)
- Morae from TechSmith (http://www.techsmith.com/morae.html)
- Usability Testing Environment (UTE) from Mind Design Systems (http://utetool.com/)
- Usability Test Data Logger from UserFocus (http://www.userfocus.co.uk/resources/datalogger.html)

Our website, MeasuringUX.com, also includes a simple macro for use in Microsoft Word for logging start and finish times. An automated method of logging has several advantages. Not only is it less error-prone but it's also much less obtrusive. The last thing you want is a participant in a usability test to feel nervous from watching you press the start and stop button on your stopwatch or smartphone.

TURNING ON AND OFF THE CLOCK

Not only do you need a way to measure time, but you also need some rules about *how* to measure time. Perhaps the most important rule is when to turn the clock on and off. Turning on the clock is fairly straightforward: If you have the participants read the task aloud, you start the clock as soon as they finish reading the task.

Turning off the clock is a more complicated issue. Automated time-keeping tools typically have an "answer" button. Users are required to hit the "answer" button, at which point the timing ends, and they are asked to provide an answer and perhaps a few additional questions. If you are not using an automated method, you can have users report the answer verbally or perhaps even write it down. However, there are many situations in which you may not be sure if they have found the answer. In this situation, it's important for participants to indicate their answer as quickly as possible. In any case, you want to stop timing when the participant states the answer or otherwise indicates that she has completed the task.

TABULATING TIME DATA

The first thing you need to do is arrange the data in a table, as shown in Table 4.2. Typically, you will want a list of all the participants in the first column, followed by the time data for each task in the remaining columns (expressed in seconds, or minutes if the tasks are long). Table 4.2 also shows summary data, including the average, median, geometric mean, and confidence intervals for each task.

	Task 1	Task 2	Task 3	Task 4	Task 5
Participant 1	259	112	135	58	8
Participant 2	253	64	278	160	22
Participant 3	42	51	60	57	26
Participant 4	38	108	115	146	26
Participant 5	33	142	66	47	38
Participant 6	33	54	261	26	42
Participant 7	36	152	53	22	44
Participant 8	112	65	171	133	46
Participant 9	29	92	147	56	56
Participant 10	158	113	136	83	64
Participant 11	24	69	119	25	68
Participant 12	108	50	145	15	75
Participant 13	110	128	97	97	78
Participant 14	37	66	105	83	80
Participant 15	116	78	40	163	100
Participant 16	129	152	67	168	109
Participant 17	31	51	51	119	116
Participant 18	33	97	44	81	127
Participant 19	75	124	286	103	236
Participant 20	76	62	108	185	245
Mean	86.6	91.5	124.2	91.4	80.3
Median	58.5	85.0	111.5	83.0	66.0
Geometric mean	65.2	85.2	105.0	73.2	60.3
90% confidence interval	31.1	15.4	33.1	23.6	28.0
Lower bound	55.5	76.1	91.1	67.7	52.3
Upper bound	117.7	106.9	157.3	115.0	108.3

Table 4.2 Time-on-task data for 20 participants and five tasks (all data are expressed in seconds).

WORKING WITH TIME DATA IN EXCEL

If you use Excel to log data during a usability test, it's often convenient to use times that are formatted as hours, minutes, and (sometimes) seconds (hh:mm:ss). Excel provides a variety of formats for time data. This makes it easy to enter times, but it complicates matters slightly when you need to calculate an elapsed time. For example, assume that a task started at 12:46 PM and ended at 1:04 PM. Although you can look at those times

and determine that the elapsed time was 18 minutes, how to get Excel to calculate that isn't so obvious. Internally, Excel stores all times as a number reflecting the number of seconds elapsed since midnight. So to convert an Excel time to minutes, multiply it by 60 (the number of minutes in an hour) and then by 24 (the number of hours in a day). To convert to seconds, multiply by another 60 (the number of seconds in a minute). Here's what it looks like in Excel, including the formula:

D2 | fx | =C2*60*24

	A	B	C	D	E
1	Start Time	Finish Time	Elapsed	Minutes	Seconds
2	12:46:00 PM	1:04:00 PM	0.0125	18	1080

4.2.3 Analyzing and Presenting Time-on-Task Data

You can analyze and present time-on-task data in many different ways. Perhaps the most common way is to look at the average amount of time spent on any particular task or set of tasks by averaging all the times for each user by task (Figure 4.4). This is a straightforward and intuitive way to report time-on-task data. One downside is the potential variability across users. For example, if you have several users who took an exceedingly long time to complete a task, it may increase the average considerably. Therefore, you should always report a confidence interval to show the variability in the time data. This will not only show the variability within the same task but also help visualize the difference across tasks to determine whether there is a statistically significant difference between tasks.

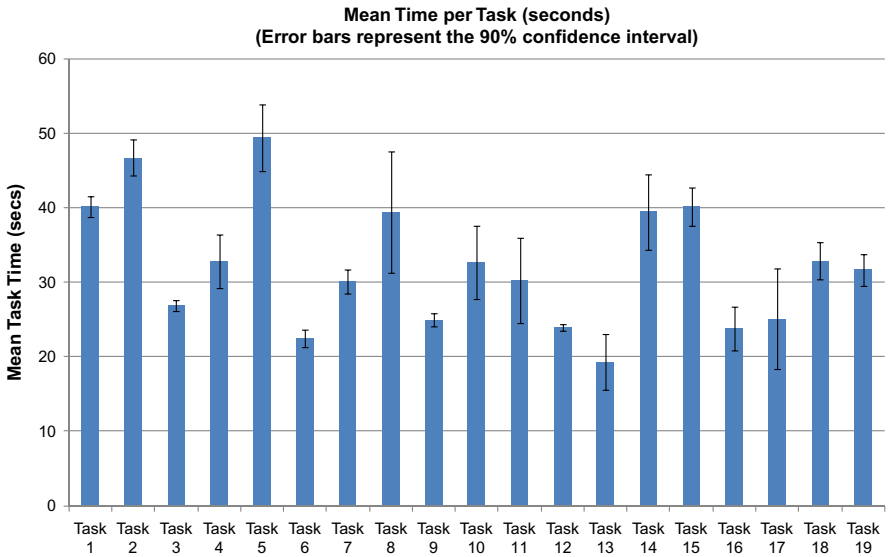


Figure 4.4 Mean time on task, in seconds, for 19 tasks. Error bars represent a 90% confidence interval. These data are from an online study of a prototype website.

WHAT'S THE RIGHT PRECISION FOR TIME DATA?

How accurate do you need to be with your time data? Of course, it depends on what you're measuring, but the majority of the times we deal with it in the user experience world are either in seconds or minutes. It's very rare that we need to record subsecond times. Similarly, if you're dealing with times that are more than an hour, it's probably not necessary to be more accurate than whole minutes.

Sometimes it's more appropriate to summarize time-on-task data using the median rather than the mean. The median is the middle point in an ordered list of all the times: Half of the times are below the median and half are above the median. Similarly, the geometric mean is potentially less biased than the mean. Time data are typically skewed, in which case the median or geometric mean may be more appropriate. In practice, we find that using these other methods of summarizing time data may change the overall level of the times, but the kinds of patterns you're interested in (e.g., comparisons across tasks) usually stay the same; the same tasks still took the longest or shortest times overall.

EXCEL TIP

The median can be calculated in Excel using the = MEDIAN function. The geometric mean can be calculated using the = GEOMEAN function.

WHAT'S A GEOMETRIC MEAN?

While the mean (or arithmetic average) is based on the *sum* of a set of numbers, the geometric mean is based on their *product*. For example, the mean of 2 and 8 is (2 + 8)/2, or 10/2, which is 5. The geometric mean of 2 and 8 is sqrt(2*8), or sqrt(16), which is 4. The geometric mean will usually be smaller than the mean.

RANGES

A variation on calculating average completion time by task is to create ranges, or discrete time intervals, and report the frequency of users who fall into each time interval. This is a useful way to visualize the spread of completion times by all users. In addition, this might be a helpful approach to look for any patterns in the type of users who fall within certain segments. For example, you may want to focus on those users who had particularly long completion times to see if they share any common characteristics.

THRESHOLDS

Another useful way to analyze task time data is by using a threshold. In many situations, the only thing that matters is whether users can complete certain tasks

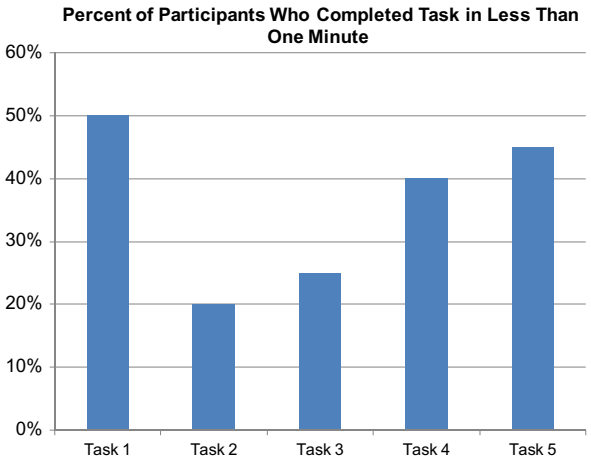


Figure 4.5 An example showing the percentage of users who completed each task in less than 1 minute.

within an acceptable amount of time. In many ways, the average is unimportant. The main goal is to minimize the number of users who need an excessive amount of time to complete a task. The main issue is determining what the threshold should be for any given task. One way is to perform the task yourself, keeping track of the time, and then double or triple that number. Alternatively, you could work with the product team to come up with a threshold for each task based on competitive data or even a best guess. Once you have set your threshold, simply calculate the percentage of users above or below the threshold and plot as illustrated in Figure 4.5.

DISTRIBUTIONS AND OUTLIERS

Whenever analyzing time data, it's critical to look at the distribution. This is particularly true for time-on-task data collected via automated tools (when the moderator is not present). Participants might take a phone call or even go out to lunch in the middle of a task. The last thing you want is to include a task time of 2 hours among other times of only 15 to 20 seconds when calculating an average! It's perfectly acceptable to exclude outliers from your analysis, and many statistical techniques for identifying them are available. Sometimes we exclude any times that are more than two or three standard deviations above the mean. Alternatively, we sometimes set up thresholds, knowing that it should never take a user more than *x* seconds to complete a task. You should have some rationale for using an arbitrary threshold for excluding outliers.

The opposite problem—participants apparently completing a task in unusually short amounts of time—is also common in online studies. Some participants may be in such a hurry or only care about the compensation that they simply fly through the study as fast as they can. In most cases, it's very easy to identify these individuals through their time data. For each task, determine the fastest possible time. This would be the time it would take someone with perfect knowledge and optimal efficiency to complete the task. For example, if there is no way you, as an expert user of the product, can finish the task in less than 8 seconds, then it is highly unlikely that a typical user could complete the task any faster. Once you have established this minimum acceptable time, you should identify the tasks that have times less than that minimum. These are candidates for removal—not just of the time but of the entire task (including any other data for the task such as success or subjective rating). Unless you can find evidence suggesting otherwise, the time indicates that the participant did not make a reasonable attempt at the task. If a participant did this for multiple tasks, you should consider dropping that participant. You can expect anywhere

from 5 to 10% of the participants in an online study to be in it only for the compensation.

4.2.4 Issues to Consider When Using Time Data

Some of the issues to think about when analyzing time data is whether to look at all tasks or just successful tasks, what the impact of using a think-aloud protocol might be, and whether to tell test participants that time is being measured.

ONLY SUCCESSFUL TASKS OR ALL TASKS?

Perhaps the first issue to consider is whether you should include times for only successful tasks or all tasks in the analysis. The main advantage of only including successful tasks is that it is a cleaner measure of efficiency. For example, time data for unsuccessful tasks are often very difficult to estimate. Some users will keep on trying until you practically unplug the computer. Any task that ends with the participant giving up or the moderator "pulling the plug" is going to result in highly variable time data.

The main advantage of analyzing time data for *all* tasks, successful or not, is that it is a more accurate reflection of the overall user experience. For example, if only a small percentage of users were successful, but that particular group was very efficient, the overall time on task is going to be low. Therefore, it is easy to misinterpret time-on-task data when only analyzing successful tasks. Another advantage of analyzing time data for all tasks is that it is an independent measure in relation to task success data. If you only analyze time data for successful tasks, you're introducing a dependency between the two sets of data.

A good rule is that if the participant always determined when to give up on unsuccessful tasks, you should include all times in the analyses. If the moderator sometimes decided when to end an unsuccessful task, then use only the times for the successful tasks.

USING A CONCURRENT THINK-ALOUD PROTOCOL

Another important issue to consider is whether to use a concurrent think-aloud protocol when collecting time data (i.e., asking participants to think aloud while they are going through the tasks). Most usability specialists rely heavily on a concurrent think-aloud protocol to gain important insight into the user experience. But sometimes a think-aloud protocol leads to a tangential topic or a lengthy interaction with the moderator. The last thing you want to do is measure time on task while a participant is giving a 10-minute diatribe on the importance of fast-loading web pages. When you want to capture time on task but also use a concurrent think-aloud protocol, a good solution is to ask participants to "hold" any longer comments for the time between tasks. Then you can have a dialog with the participant about the just-completed task after the "clock is stopped."

RETROSPECTIVE THINK ALOUD (RTA)

A technique that's gaining in popularity among many usability professionals is retrospective think aloud (e.g., Birns, Joffre, Leclerc, & Paulsen, 2002; Guan, Lee, Cuddihy, & Ramey, 2006; Petrie & Precious, 2010). With this technique, participants typically remain silent while they are interacting with the product being tested. Then, after all the tasks, they are shown some kind of "reminder" of what they did during the session and are asked to describe what they were thinking or doing at various points in the interaction. The reminder can take several different forms, including a video replay of screen activity, perhaps with a camera view of the user, or an eye-tracking replay showing what the user was looking at. This technique probably yields the most accurate task time data. There's also some evidence that the additional cognitive load of concurrent think aloud causes participants to be less successful with their tasks. For example, van den Haak, de Jong, and Schellens (2004) found that participants in a usability study of a library website were successful with only 37% of their tasks when using concurrent think aloud, but they were successful with 47% when using RTA. But keep in mind that it will take almost twice as long to run sessions using RTA.

SHOULD YOU TELL PARTICIPANTS ABOUT THE TIME MEASUREMENT?

An important question to consider is whether to tell participants you are recording their time. It's possible that if you don't, participants won't behave in an efficient manner. It's not uncommon for participants to explore different parts of a website when they are in the middle of a task. On the flip side, if you tell them they are being timed, they may become nervous and feel they are the ones being tested and not the product. A good compromise is asking the participants to perform the tasks as quickly and accurately as possible, without volunteering that they are being explicitly timed. If the participant happens to ask (which they rarely do), then simply state that you are noting the start and finish time for each task.

4.3 ERRORS

Some user experience professionals believe errors and usability issues are essentially the same thing. Although they are certainly related, they are actually quite different. A usability issue is the underlying *cause* of a problem, whereas one or more errors are a possible *outcome* of an issue. For example, if users are experiencing a problem in completing a purchase on an e-commerce website, the issue (or cause) may be confusing labeling of the products. The error, or the result of the issue, may be the act of choosing the wrong options for the product they want to buy. Essentially, errors are incorrect actions that may lead to task failure.

4.3.1 When to Measure Errors

In some situations it's helpful to identify and classify errors rather than just document usability issues. Measuring errors is useful when you want to understand the

specific action or set of actions that may result in task failure. For example, a user may make the wrong selection on a web page and sell a stock instead of buying more. A user may push the wrong button on a medical device and deliver the wrong medication to a patient. In both cases, it's important to know what errors were made and how different design elements may increase or decrease the frequency of errors.

Errors are a useful way of evaluating user performance. While being able to complete a task successfully within a reasonable amount of time is important, the number of errors made during the interaction is also very revealing. Errors can tell you how many mistakes were made, where they were made while interacting with the product, how various designs produce different frequencies and types of errors, and generally how usable something really is.

Measuring errors is not right for every situation. We've found that there are three general situations where measuring errors might be useful:

1. When an error will result in a significant loss in efficiency—for example, when an error results in a loss of data, requires the user to reenter information, or slows the user significantly in completing a task.
2. When an error will result in significant costs to your organization or the end user—for example, if an error will result in increased call volumes to customer support or in increased product returns.
3. When an error will result in task failure—for example, if an error will cause a patient to receive the wrong medication, a voter to vote for the wrong candidate accidentally, or a web user to buy the wrong product.

4.3.2 What Constitutes an Error?

Surprisingly, there is no widely accepted definition of what constitutes an error. Obviously, it's some type of incorrect action on the part of the user. Generally an error is an action that causes the user to stray from the path to successful completion. Sometimes failing to take an action can be an error. Errors can be based on many different types of actions by the user, such as the following:

- Entering incorrect data into a form field (such as typing the wrong password during a login attempt)
- Making the wrong choice in a menu or drop-down list (such as selecting "Delete" instead of "Modify")
- Taking an incorrect sequence of actions (such as reformatting their home media server when all they were trying to do was play a recorded TV show)
- Failing to take a key action (such as clicking on a key link on a web page)

Obviously, the range of possible actions will depend on the product you are studying (website, cell phone, DVD player, etc.). When you're trying to determine what constitutes an error, first make a list of all the possible actions a user can take on your product. Some of those actions are errors. Once you have a universe of possible actions, you can then start to define many of the different types of errors that can be made using the product.

4.3.3 Collecting and Measuring Errors

Measuring errors is not always easy. Similar to other performance metrics, you need to know what the correct action should be or, in some cases, the correct set of actions. For example, if you're studying a password reset form, you need to know what is considered the correct set of actions to reset the password successfully and what is not. The better you can define the universe of correct and incorrect actions, the easier it will be to measure errors.

An important consideration is whether a given task presents only a single error opportunity or multiple error opportunities. An error opportunity is basically a chance to make a mistake. For example, if you're measuring the usability of a typical login screen, at least two error opportunities are possible: making an error when entering the user name and making an error when entering the password. If you're measuring the usability of an online form, there could be as many error opportunities as there are fields on the form.

In some cases there might be multiple error opportunities for a task but you only care about one of them. For example, you might be interested only in whether users click on a specific link that you know will be critical to completing their task. Even though errors could be made on other places on the page, you're narrowing your scope of interest to that single link. If users don't click on the link, it is considered an error.

The most common way of organizing error data is by task. Simply record the number of errors for each task and each user. If there is only a single opportunity for error, the numbers will be 1's and 0's:

0 = No error
1 = One error

If multiple error opportunities are possible, numbers will vary between 0 and the maximum number of error opportunities. The more error opportunities, the harder and more time-consuming it will be to tabulate the data. You can count errors while observing users during a lab study, by reviewing videos after the sessions are over, or by collecting the data using an automated or online tool.

If you can clearly define all the possible error opportunities, another approach could be to identify the presence (1) or absence (0) of each error opportunity for each user and task. The average of these for a task would then reflect the incidence of those errors.

4.3.4 Analyzing and Presenting Errors

The analysis and presentation of error data differ slightly depending on whether a task has only one error opportunity or multiple error opportunities. If each task has only one error opportunity, then the data are binary for each task (the user made an error or didn't), which means that the analyses are basically all the same as they are for binary task success. You could, for example, look at average error rates per task or per participant. Figure 4.6 is an example of presenting

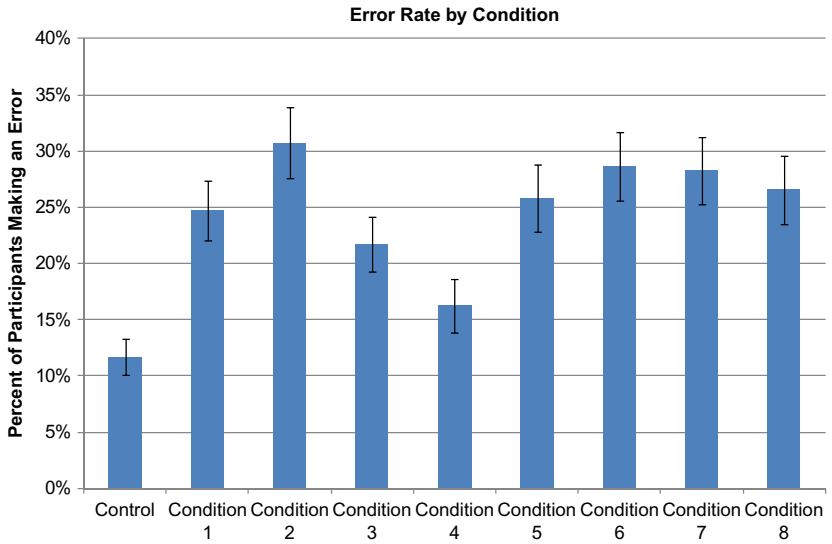


Figure 4.6 An example showing how to present data for single error opportunities. In this study, only one error opportunity per task (entering a password incorrectly) was possible, and the graph shows the percentage of participants who made an error for each condition.

errors based on a single opportunity per task. In this example, they were interested in the percentage of participants who experienced an error when using different types of on-screen keyboards (Tullis, Mangan, & Rosenbaum, 2007). The control condition was the current QWERTY keyboard layout.

In many situations, there are multiple opportunities for errors per task (e.g., multiple input fields in a "new account" application). Here are some of the common ways to analyze data from tasks with multiple error opportunities:

- A good place to start is to look at the frequency of errors for each task. You will be able to see which tasks are resulting in the most errors. But this may be misleading if each task has a different number of error opportunities. In that case, it might be better to divide the total number of errors for the task by the total number of error opportunities. This creates an error rate that takes into account the number of opportunities.
- You could calculate the average number of errors made by each participant for each task. This will also tell you which tasks are producing the most errors. However, it may be more meaningful because it suggests that a typical user might experience *x* number of errors on a particular task when using the product. Another advantage is that it takes into account extremes. If you are simply looking at the frequency of errors for each task, some users may be the source of most of the errors, whereas many others are performing the task error-free. By taking an average number of errors for each user, this bias is reduced.

- In some situations it might be interesting to know which tasks fall above or below a threshold. For example, for some tasks, an error rate above 20% is unacceptable, whereas for others, an error rate above 5% is unacceptable. The most straightforward analysis is to first establish an acceptable threshold for each task or each participant. Next, calculate whether that specific task's error rate or user error count was above or below the threshold.
- Sometimes you want to take into account that not all errors are created equal. Some errors are much more serious than others. You could assign a severity level to each error, such as high, medium, or low, and then calculate the frequency of each error type. This could help the project team focus on the issues that seem to be associated with the most serious errors.

4.3.5 Issues to Consider When Using Error Metrics

Several important issues must be considered when looking at errors. First, make sure you are not double counting errors. Double counting happens when you assign more than one error to the same event. For example, assume you are counting errors in a password field. If a user typed an extra character in the password, you could count that as an "extra character" error, but you shouldn't also count it as an "incorrect character" error.

Sometimes you need to know more than just an error rate; you need to know *why* different errors are occurring. The best way to do this is by looking at each type of error. Basically, you want to try to code each error by type of error. Coding should be based on the various types of errors that occurred. For example, continuing with the password example, the types of errors might include "missing character," "transposed characters," "extra character," and so on. At a higher level, you might have "navigation error," "selection error," "interpretation error," and so on. Once you have coded each error, you can run frequencies on the error type for each task to better understand exactly where the problems lie. This will also help improve the efficiency with which you collect error data.

In some cases, an error is the same as failing to complete a task—for example, with a login page that allows only one chance at logging in. If no errors occur while logging in, it is the same as task success. If an error occurs, it is the same as task failure. In this case, it might be easier to report errors as task failure. It's not so much a data issue as it is a presentation issue. It's important to make sure your audience understands your metrics clearly.

Another enlightening metric can be the incidence of repeated errors—namely the case where a participant makes essentially the same mistake more than once, such as repeatedly clicking on the same link that looks like it might be the right one but isn't.

4.4 EFFICIENCY

Time on task is often used as a measure of efficiency, but another way to measure efficiency is to look at the amount of effort required to complete a task. This

is typically done by measuring the number of actions or steps that users took in performing each task. An action can take many forms, such as clicking a link on a web page, pressing a button on a microwave oven or a mobile phone, or flipping a switch on an aircraft. Each action a user performs represents a certain amount of effort. The more actions taken by a user, the more effort involved. In most products, the goal is to increase productivity by minimizing the number of discrete actions required to complete a task, thereby minimizing the amount of effort.

What do we mean by effort? There are at least two types of effort: cognitive and physical. Cognitive effort involves finding the right place to perform an action (e.g., finding a link on a web page), deciding what action is necessary (should I click this link?), and interpreting the results of the action. Physical effort involves the physical activity required to take action, such as moving a mouse, inputting text on a keyboard, turning on a switch, and many others.

AN INTERESTING WAY OF MEASURING COGNITIVE EFFORT

One way of measuring cognitive effort is using performance on a task that's peripheral, or secondary, to the participants' primary task. The more cognitive effort the primary task requires, the worse the performance on the secondary task will be. An interesting variation of this was used by Ira Hyman and associates at Western Washington University to measure cell phone distraction (Hyman et al., 2010). They had one of their students ride a unicycle around a popular square on campus while wearing a clown suit (a rather memorable sight!). Then they observed 347 pedestrians walking across the square, some of whom were talking on their cell phones. After crossing the square, they asked the pedestrians if they had seen a unicycling clown. The clown was remembered by 71% of those walking with a friend, 61% of those listening to music, and 51% of those walking alone. But only 25% of those talking on a cell phone remembered the unicycling clown!

Efficiency metrics work well if you are concerned not only with the time it takes to complete a task but also the amount of cognitive and physical effort involved. For example, if you're designing an automobile navigation system, you need to make sure that it does not take much effort to interpret its navigation directions, as the driver's attention must be focused on the road. It would be important to minimize both physical and cognitive effort to use the navigation system.

4.4.1 Collecting and Measuring Efficiency

There are some important points to keep in mind when collecting and measuring efficiency.

- *Identify the action(s) to be measured:* For websites, mouse clicks or page views are common actions. For software, it might be mouse clicks or keystrokes. For appliances or consumer electronics, it could be button

presses. Regardless of the product being evaluated, you should have a clear idea of all the possible actions.

- *Define the start and end of an action:* You need to know when an action begins and ends. Sometimes the action is very quick, such as a press of a button, but other actions can take much longer. An action may be more passive in nature, such as looking at a web page. Some actions have a very clear start and end, whereas other actions are less defined.
- *Count the actions:* You must be able to count the actions. Actions must happen at a pace that can be identified visually or, if they are too fast, by an automated system. Try to avoid having to review hours of video to collect efficiency metrics.
- *Actions must be meaningful:* Each action should represent an incremental increase in cognitive and/or physical effort. The more actions, the more effort. For example, each click of a mouse is almost always an incremental increase in effort.

Once you have identified the actions you want to capture, counting those actions is relatively simple. You can do it manually, such as counting page views or presses of a button. This will work for fairly simple products, but in most cases, it is not practical. Many times a participant is performing these actions at amazing speeds. There may be more than one action every second, so using automated data collection tools is far preferable.

KEYSTROKE-LEVEL MODELING

This discussion of low-level actions such as keystrokes and mouse clicks should sound familiar if you've ever studied theories of human–computer interaction. A framework called GOMS (Goals, Operators, Methods, and Selection rules) dates back to a classic book, "The Psychology of Human–Computer Interaction" (Card, Moran, & Newell, 1983). In it, a user's interaction with a computer is decomposed into its fundamental units, which could be physical, cognitive, or perceptual. Identifying these fundamental units and assigning times to each of them allow you to predict how long a particular interaction will take. A simplified version of GOMS is called the keystroke-level model, which, as its name implies, focuses on keystrokes and mouse clicks (e.g., Sauro, 2009).

4.4.2 Analyzing and Presenting Efficiency Data

The most common way to analyze and present efficiency metrics is by looking at the number of actions each participant takes to complete a task. Simply calculate an average for each task (by participant) to see how many actions are taken. This analysis is helpful in identifying which tasks required the most amount of effort; it works well when each task requires about the same number of actions. However, if some tasks are more complicated than others, it may be misleading. It's also important to represent the confidence intervals (based on a continuous distribution) for this type of chart.

Shaikh, Baker, and Russell (2004) used an efficiency metric based on the number of clicks to accomplish the same task on three different weight-loss sites: Atkins, Jenny Craig, and Weight Watchers. They found that users were significantly more efficient (needed fewer clicks) with the Atkins site than with the Jenny Craig or Weight Watchers sites.

LOSTNESS

Another measure of efficiency sometimes used in studying behavior on the web is called "lostness" (Smith, 1996). Lostness is calculated using three values:

N: The number of *different* web pages visited while performing the task
S: The *total* number of pages visited while performing the task, counting revisits to the same page
R: The *minimum* (optimum) number of pages that must be visited to accomplish the task

Lostness, *L*, is then calculated using the following formula:

$$L = \text{sqrt}[(N/S - 1)^2 + (R/N - 1)^2].$$

Consider the example shown in Figure 4.7. In this case, the user's task is to find something on Product Page C1. Starting on the home page, the minimum number of page visits (*R*) to accomplish this task is three. However, Figure 4.8 illustrates the path a particular user took in getting to that target item. This user started down some incorrect paths before finally getting to the right place, visiting a total of six different pages (*N*), or a total of nine page visits (*S*). So for this example:

$N = 6$
$S = 9$
$R = 3$

$$L = \text{sqrt}[(6/9 - 1)^2 + (3/6 - 1)^2] = 0.60$$

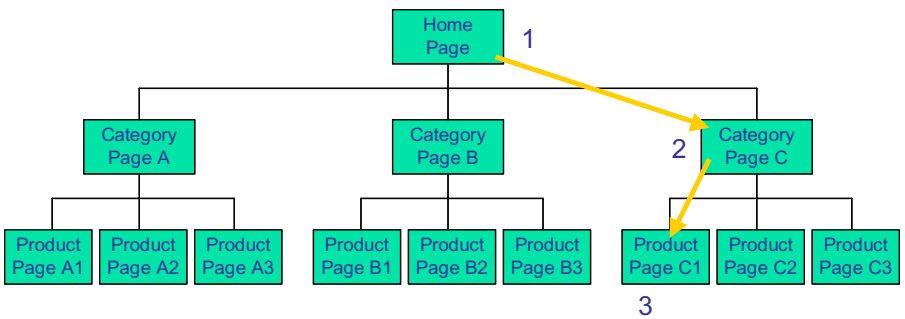


Figure 4.7 Optimum number of steps (three) to accomplish a task that involves finding a target item on Product Page C1 starting from the home page.

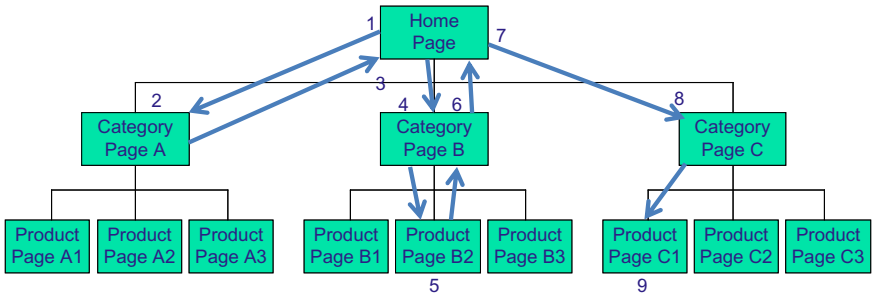


Figure 4.8 Actual number of steps a user took in getting to the target item on Product Page C1. Note that each revisit to the same page is counted, giving a total of nine steps.

A perfect lostness score would be 0. Smith (1996) found that participants with a lostness score less than 0.4 did not exhibit any observable characteristics of being lost. However, she reported that participants with a lostness score greater than 0.5 definitely did appear to be lost. Note that additional measures of lostness have been proposed by Otter & Johnson (2000) and Gwizdka & Spence (2007).

Once you calculate a lostness value, you can easily calculate the average lostness value for each task. The number or percentage of participants who exceed the ideal number of actions can also be indicative of the efficiency of the design. For example, you could show that 25% of the participants exceeded the ideal or minimum number of steps, and you could break it down even further by saying that 50% of the participants completed a task with the minimum number of actions.

BACKTRACKING METRIC

Treejack (http://www.optimalworkshop.com/treejack.htm) is a tool from Optimal Workshop for testing information architectures (IAs). Participants in a Treejack study navigate an information hierarchy to indicate where in the hierarchy they would expect to find a given piece of information or perform some action. Participants can move down the hierarchy or, if they need to, they can move back up it. Several useful metrics come out of a Treejack study, including traditional ones, such as where participants indicated they would expect to find each function. But a particularly interesting metric is a "backtracking" metric that indicates cases where a participant went back *up* the hierarchy. You can then look at the percentage of participants who "backtracked" while performing each task. In our IA studies, we've found this was often the most revealing metric.

4.4.3 Efficiency as a Combination of Task Success and Time

Another view of efficiency is that it's a combination of two of the metrics discussed in this chapter: task success and time on task. The Common Industry

Format for Usability Test Reports (ISO/IEC 25062:2006) specifies that the "core measure of efficiency" is the ratio of the task completion rate to the mean time per task. Basically, it expresses task success per unit time. Most commonly, time per task is expressed in minutes, but seconds could be appropriate if the tasks are very short or even hours if they are unusually long. The unit of time used determines the scale of the results. Your goal is to choose a unit that yields a "reasonable" scale (i.e., one where most of the values fall between 1 and 100%). Table 4.3 shows an example of calculating an efficiency metric based on task completion and task time. Figure 4.9 shows how this efficiency metric looks in a chart.

	Task Completion Rate	Task Time (min)	Efficiency (%)
Task 1	65%	1.5	43
Task 2	67%	1.4	48
Task 3	40%	2.1	19
Task 4	74%	1.7	44
Task 5	85%	1.2	71
Task 6	90%	1.4	64
Task 7	49%	2.1	23
Task 8	33%	1.3	25

Table 4.3 The efficiency measure is simply the ratio of task completion to task time in minutes[a].

[a]Of course, higher values of efficiency are better. In this example, users appear to have been more efficient in performing tasks 5 and 6 than the other tasks.

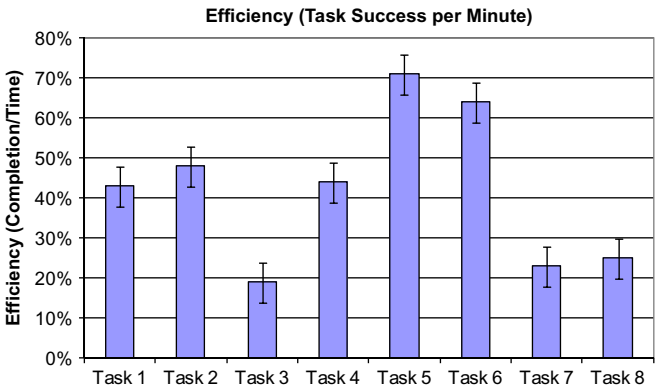


Figure 4.9 An example showing efficiency as a function of completion rate/time.

A slight variation on this approach to calculating efficiency is to count the number of tasks completed successfully by each participant and divide that by the total time spent by the participant on *all* tasks (successful and unsuccessful).

This gives a very straightforward efficiency score for each participant: number of tasks completed successfully per minute (or whatever unit of time you used). If a participant completed 10 tasks successfully in a total time of 10 minutes, then that participant was successfully completing 1 task per minute overall. This works best when all participants attempted the same number of tasks and the tasks are relatively comparable in terms of their level of difficulty.

Figure 4.10 shows data from an online study comparing four different navigation prototypes for a website. This was a between-subjects study, in which each participant used only one of the prototypes, but all participants were asked to perform the same 20 tasks. Over 200 participants used each prototype. We were able to count the number of tasks completed successfully by each participant and divide that by the total time that participant spent. The averages of these (and the 95% confidence intervals) are shown in Figure 4.10.

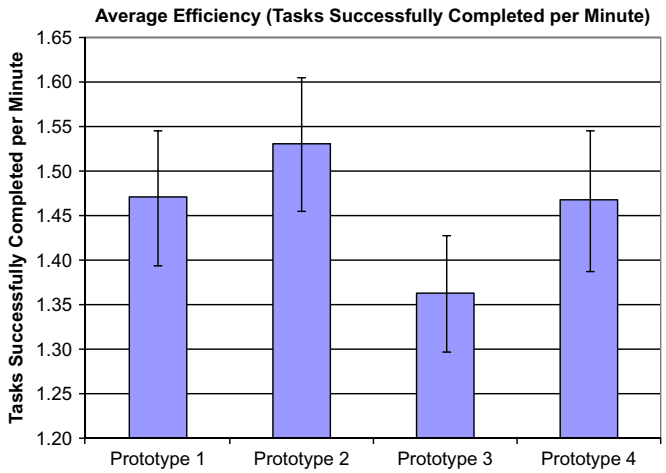


Figure 4.10 Average number of tasks completed successfully per minute in an online study of four different prototypes of navigation for a website. Over 200 participants attempted 20 tasks with each prototype. Participants using Prototype 2 were significantly more efficient (i.e., completed more tasks per minute) than those using Prototype 3.

4.5 LEARNABILITY

Most products, especially new ones, require some amount of learning. Usually, learning does not happen in an instant but occurs over time as experience increases. Experience is based on the amount of time spent using a product and the variety of tasks performed. Learning is sometimes quick and painless, but at other times it is quite arduous and time-consuming. Learnability is the extent to which something can be learned efficiently. It can be measured by looking at how much time and effort are required to become proficient, and ultimately expert in using something. We believe that learnability is an important user

experience metric that doesn't receive as much attention as it should. It's an essential metric if you need to know how someone develops proficiency with a product over time.

Consider the following example. Assume you're a user experience professional who has been asked to evaluate a time-keeping application for employees within their organization. You could go into the lab and test with 10 participants, giving each participant a set of core tasks. You might measure task success, time on task, errors, and even overall satisfaction. Using these metrics will allow you to get some sense of the usability of the application. Although these metrics are useful, they can also be misleading. Since the use of a time-keeping application is not a one-time event, but happens with some degree of frequency, learnability is very important. A key factor is how much time and effort are required to become *proficient* using the time-keeping application. Yes, there may be some initial obstacles when first using the application, but what really matters is "getting up to speed." It's quite common in usability studies to only look at a participant's initial exposure to something, but sometimes it's more important to look at the amount of effort needed to become proficient.

Learning can happen over a short period of time or over longer periods of time. When learning happens over a short period of time, the user tries out different strategies to complete tasks. A short period of time might be several minutes, hours, or days. For example, if users have to submit their timesheets every day using a time-keeping application, they try to quickly develop some type of mental model of how the application works. Memory is not a big factor in learnability; it is more about adapting strategies to maximize efficiency. Within a few hours or days, it is hoped that maximum efficiency is achieved.

LEARNABILITY AND "SELF-SERVICE"

Learnability is much more important today than it was in the early days of computers. The web has facilitated a move toward many more "self-service" applications than we previously had. At the same time, it has fostered an expectation that you should be able to use just about anything on the web without extensive training or practice. In the 1980s, if you wanted to book an airline flight yourself, you called and spoke to a representative who had extensive training in the use of a mainframe-based airline reservation system. Today you go to one of many different websites that let you book an airline flight. How long do you think one of those sites would stay in business if it started by saying, "OK, let's start with a 3-hour training course on the use of our site?" Learnability is a key differentiator in today's self-service economy.

Learning can also happen over a longer time period, such as weeks, months, or years. This is the case where significant gaps exist in time between each use. For example, if you only fill out an expense report every few months, learnability can be a significant challenge because you may have to relearn the application each time you use it. In this situation, memory is very important. The more time there is between experiences with the product, the greater the reliance on memory.

4.5.1 Collecting and Measuring Learnability Data

The process of collecting and measuring learnability data is basically the same as it is for the other performance metrics, but you're collecting the data at multiple times. Each instance of collecting the data is considered a trial. A trial might be every 5 minutes, every day, or once a month. The time between trials, or when you collect the data, is based on expected frequency of use.

The first decision is which type of metrics you want to use. Learnability can be measured using almost any performance metric over time, but the most common ones are those that focus on efficiency, such as time on task, errors, number of steps, or task success per minute. As learning occurs, you expect to see efficiency improve.

After you decide which metrics to use, you need to decide how much time to allow between trials. What do you do when learning occurs over a very long time? What if users interact with a product once every week, month, or even year? The ideal situation would be to bring the same participants into the lab every week, month, or even year. In many cases, this is not very practical. The developers and the business sponsors might not be very pleased if you told them the study will take 3 years to complete. A more realistic approach is to bring in the same participants over a much shorter time span and acknowledge the limitation in the data. Here are a few alternatives:

- *Trials within the same session*. The participant performs the task, or set of tasks, one right after the other, with no breaks in between. This is very easy to administer, but it does not take into account significant memory loss.
- *Trials within the same session but with breaks in between each task*. The break might be a distracter task or anything that might promote forgetting. This is fairly easy to administer, but it tends to make each session relatively long.
- *Trials between sessions*: The participant performs the same tasks over multiple sessions, with at least 1 day in between. This may be the least practical, but most realistic, if the product is used sporadically over an extended period of time.

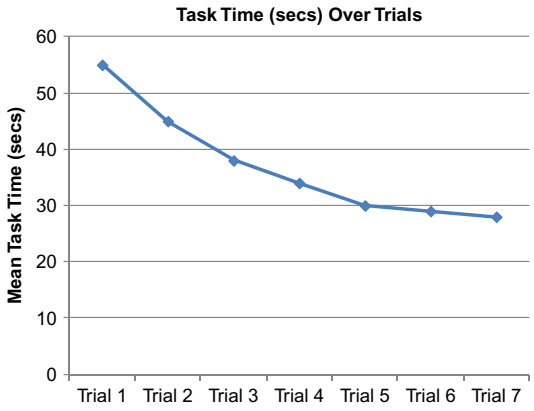


Figure 4.11 An example of how to present learnability data based on time on task.

4.5.2 Analyzing and Presenting Learnability Data

The most common way to analyze and present learnability data is by examining a specific performance metric (such as time on task, number of steps, or number of errors) by trial for each task or aggregated across all tasks. This will show you how that performance metric changes as a function of experience, as illustrated in Figure 4.11. You could aggregate all the tasks together and represent them as a single line of data or you could look at each task as separate lines of data. This can help determine how the

learnability of different tasks compare, but it also can also make the chart harder to interpret.

The first aspect of the chart you should notice is the slope of the line(s). Ideally, the slope (sometimes called the learning curve) is fairly flat and low on the *y* axis (in the case of errors, time on task, number of steps, or any other metric where a smaller number is better). If you want to determine whether a statistically significant difference between the learning curves (or slopes) exists, you need to perform an analysis of variance and see if there is a main effect of trial. (See Chapter 2 for a discussion of analysis of variance.)

You should also notice the point of asymptote, or essentially where the line starts to flatten out. This is the point at which users have learned as much as they can, and there is very little room for improvement. Project team members are always interested in how long it will take someone to reach maximum performance.

Finally, you should look at the difference between the highest and the lowest values on the *y* axis. This will tell you how much learning must occur to reach maximum performance. If the gap is small, users will be able to learn the product quickly. If the gap is large, users may take quite some time to become proficient with the product. One easy way to analyze the gap between highest and lowest scores is by looking at the ratio of the two. Here is an example:

- If the average time on the first trial is 80 seconds and on the last trial is 60 seconds, the ratio shows that users are initially taking 1.3 times longer.
- If the average number of errors on the first trial is 2.1 and on the last trial is 0.3, the ratio shows a 7 times improvement from the first trial to the last trial.

It may be helpful to look at how many trials are needed to reach maximum performance. This is a good way to characterize the amount of learning required to become proficient in using the product.

In some cases you might want to compare learnability across different conditions, as shown in Figure 4.12. In this study (Tullis, Mangan, & Rosenbaum, 2007), they were interested in how speed (efficiency) of entering a password changed over time using different types of on-screen keyboards. As you can see from the data, there is an improvement from the first trial to the second trial, but then the times flatten out

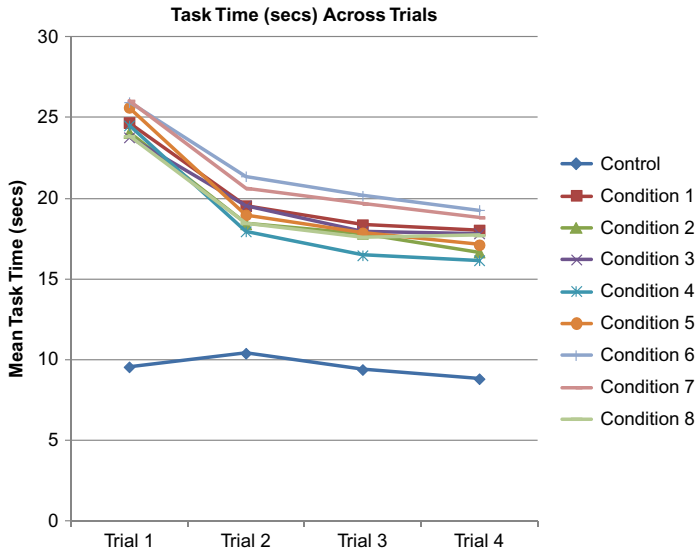


Figure 4.12 Looking at the learnability of different types of on-screen keyboards.

pretty quickly. Also, all the on-screen keyboards were significantly slower than the control condition, which was a real keyboard.

4.5.3 Issues to Consider When Measuring Learnability

Two of the key issues to address when measuring learnability are (1) what should be considered a trial and (2) how many trials to include.

WHAT IS A TRIAL?

In some situations learning is continuous. This means that the user is interacting with the product fairly continuously without any significant breaks in time. Memory is much less a factor in this situation. Learning is more about developing and modifying different strategies to complete a set of tasks. The whole concept of trials does not make much sense for continuous learning. What do you do in this situation? One approach is to take your measurements at specified time intervals. For example, you may need to take measurements every 5 minutes, 15 minutes, or every hour. In one usability study we conducted, we wanted to evaluate the learnability of a new suite of applications that would be used many times every day. We started by bringing the participants into the lab for their first exposure to the applications and their initial tasks. They then went back to their regular jobs and began using the applications to do their normal work. We brought them back into the lab 1 month later and had them perform basically the same tasks again (with minor changes in details) while we took the same performance measures. Finally, we brought them back one more time after another month and repeated the procedure. In this way, we were able to look at learnability over a 2-month period.

NUMBER OF TRIALS

How many trials should you plan for? Obviously there must be at least two, but in most cases there should be at least three or four. Sometimes it's difficult to predict where in the sequence of trials the most learning will take place or even *if* it will take place. In this situation, you should err on the side of more trials than you think you might need to reach stable performance.

4.6 SUMMARY

Performance metrics are powerful tools used to evaluate the usability of any product. They are the cornerstone of usability and can inform key decisions, such as whether a new product is ready to launch. Performance metrics are always based on user behavior rather than what they say. There are five general types of performance metrics.

1. *Task success* metrics are used when you are interested in whether users are able to complete tasks using the product. Sometimes you might only be interested in whether a user is successful or not based on a strict set of criteria (binary success). Other times you might be interested in defining different levels of success based on the degree of completion,

the user's experience in finding an answer, or the quality of the answer given.

2. *Time on task* is helpful when you are concerned about how quickly users can perform tasks with the product. You might look at the time it takes to complete a task for all users, a subset of users, or the proportion of users who can complete a task within a desired time limit.
3. *Errors* are a useful measure based on the number of mistakes users make while attempting to complete a task. A task might have a single error opportunity or multiple error opportunities, and some types of errors may be more important than others.
4. *Efficiency* is a way of evaluating the amount of effort (cognitive and physical) required to complete a task. Efficiency is often measured by the number of steps or actions required to complete a task or by the ratio of the task success rate to the average time per task.
5. *Learnability* involves looking at how any efficiency metric changes over time. Learnability is useful if you want to examine how and when users reach proficiency in using a product.

This page intentionally left blank

CHAPTER 5

Issue-Based Metrics

CONTENTS

Most user experience professionals probably consider identifying usability issues and providing design recommendations the most important parts of their job. A usability issue might involve confusion around a particular term or piece of content, method of navigation, or just not noticing something that should be noticed. These types of issues, and many others, are typically identified as part of an iterative process in which designs are being evaluated and improved throughout the design and development process. This process provides tremendous value to product design and is the cornerstone of the UX profession.

Measuring the User Experience. DOI: http://dx.doi.org/10.1016/B978-0-12-415781-1.00005-4

Usability issues are generally thought of as purely qualitative. They typically include the identification and description of a problem one or more participants experienced and, in many cases, an assessment of the underlying cause of the problem. Most UX professionals also include specific recommendations for remedying the problem and many report positive findings as well (i.e., something that worked particularly well).

Most UX professionals don't strongly associate metrics with usability issues. This may be because of the gray areas in identifying issues or because identifying issues is part of an iterative design process, and metrics are perceived as adding little value. However, not only is it possible to measure usability issues, but doing so also adds value in product design while not slowing down the iterative design process.

This chapter reviews some simple metrics around usability issues. It also discusses different ways of identifying usability issues, prioritizing the importance of different types of issues, and factors you need to think about when measuring usability issues.

5.1 WHAT IS A USABILITY ISSUE?

What do we mean by usability issues? Usability issues are based on behavior in using a product. As a UX professional you interpret the cause of these issues, such as confusing terminology or hidden navigation. Examples of the more common types of usability issues include:

- Behaviors that prevent task completion
- Behaviors that takes someone "off course"
- An expression of frustration by the participant
- Not seeing something that should be noticed
- A participant says a task is complete when it is not
- Performing an action that leads away from task success
- Misinterpreting some piece of content
- Choosing the wrong link to navigate through web pages

A key point to consider in defining usability issues is how they will be addressed. The most common use is in an iterative design process focused on improving the product. In that context, the most useful issues are those that point to possible improvements in the product. In other words, it helps if issues are reasonably actionable. If they don't point directly to a part of the interface that was causing a problem, they should at least give you some hint of where to begin looking. For example, we once saw an issue in a usability test report that said, "The mental model of the application does not match the user's mental model." Note that no behavior was mentioned. And that was it. Although this may be an interesting interpretation of some behavior in a theoretical sense, it does very little to guide designers and developers in addressing the issue.

However, consider an issue like this: "Many participants were confused by the top-level navigation menu (which is the interpretation of the behavior),

often jumping around from one section to another trying to find what they were looking for (the behavior)." Particularly if this issue is followed by a variety of detailed examples describing what happened, it could be very helpful. It tells you where to start looking (the top-level navigation), and the more detailed examples of additional behaviors may help focus on some possible solutions. Molich, Jeffries, and Dumas (2007) conducted an interesting study of usability recommendations and ways to make them more useful and usable. They suggest that all usability recommendations improve the overall user experience of the application, take into account business and technical constraints, and are specific and clear.

Of course, not all usability issues are things to be avoided. Some usability issues are positive. These are sometimes called usability "findings," as the term *issues* often has negative connotations. Here are some examples of positive usability issues:

- All participants were able to log into the application
- There were no errors in completing the search task.
- Participants were faster at creating a report

The main reason for reporting positive findings, in addition to providing some positive reinforcement for the project team, is to make sure that these aspects of the interface don't get "broken" in future design iterations.

5.1.1 Real Issues versus False Issues

One of the most difficult parts of any usability professional's job is determining which usability issues are real and which are merely an aberration. Obvious issues are those that most, if not all, participants encounter. For example, it may be obvious when participants select the wrong option from a poorly worded menu, get taken down the wrong path, and then spend a significant amount of time looking for their target in the wrong part of the application. These are behaviors the cause of which are usually a "no brainer" for almost anyone to identify.

Some usability issues are much less obvious, or it's not completely clear whether something is a real issue. For example, what if only 1 out of 10 participants expresses some confusion around a specific piece of content or terminology on a website? Or if only 1 out of 12 participants doesn't notice something she should have? At some point the UX professional must decide whether what he observed is likely to be repeatable with a larger population. In these situations, ask yourself whether the participant's behavior, thought process, perception, or decisions during the task were *logical*. In other words, is there a consistent story or reasoning behind her actions or thoughts? If so, then it may be an issue even if only one participant encountered it. However, no apparent rhyme or reason behind the behavior may be evident. If the participant can't explain why he did what he did, and it only happened once, then it's likely to be idiosyncratic and should probably be ignored.

5.2 HOW TO IDENTIFY AN ISSUE

The most common way to identify usability issues is during a study in which you are interacting with a participant directly. This might be in person or over the phone using remote testing technology. A less common way to identify usability issues is through some automated techniques such as an online study or by observing a video from a participant, similar to what is generated from a site like usertesting.com. This is where you don't have an opportunity to observe participants directly but only have access to their behavioral and self-reported data. Identifying issues through this type of data is more challenging but still quite possible.

Possible usability issues might be predicted beforehand and tracked during test sessions. But be careful that you're really *observing* the issues and not just finding them because you expected to. Your job is certainly easier when you know what to look for, but you might also miss other issues that you never considered. In our testing, we typically have an idea of what to look for, but we also try to keep an open mind to spot the surprise issues. There's no "right" approach; it all depends on the goals of the evaluation. When evaluating products that are in an early conceptual stage, it's more likely that you won't have preset ideas about what the usability issues are. As the product is further refined, you may have a clearer idea of what specific issues you're looking for.

THE ISSUES YOU EXPECT MAY NOT BE THE ONES YOU FIND

One of the earliest sets of guidelines for designing software interfaces was published by Apple (1982). It was called the *Apple IIe Design Guidelines*, and it contained a fascinating story of an early series of usability tests Apple conducted. They were working on the design of a program called *Apple Presents Apple*, which was a demonstration program for customers to use in computer stores. One part of the interface to which the designers paid little attention was asking users whether their monitor was monochrome or color. The initial design of the question was "Are you using a black-and-white monitor?" (They had predicted that users might have trouble with the word *monochrome*.) In the first usability test, they found that a majority of the participants who used a monochrome monitor answered this question incorrectly because their monitor actually displayed text in green, not white!

What followed was a series of hilarious iterations involving questions such as "Does your monitor display multiple colors?" or "Do you see more than one color on the screen?"—all of which kept failing for some participants. In desperation, they were considering including a developer with every computer just to answer this question, but then they finally hit on a question that worked: "Do the words above appear in several different colors?" In short, the issues you expect may not be the issues you find.

5.2.1 In-Person Studies

The best way to facilitate identifying usability issues during an in-person study is using a think-aloud protocol. This involves having participants verbalize their

thoughts as they are working through the tasks. Typically, the participants are reporting what they are doing, what they are trying to accomplish, how confident they are about their decisions, their expectations, and why they performed certain actions. Essentially, it's a stream of consciousness focusing on their interaction with the product. During a think-aloud protocol, you might observe the following:

- Verbal expressions of confusion, frustration, dissatisfaction, pleasure, or surprise
- Verbal expressions of confidence or indecision about a particular action that may be right or wrong
- Participants *not* saying or doing something that they should have done or said
- Nonverbal behaviors such as facial expressions and/or eye movements

In addition to listening to participants, it is very important to observe their behavior. Watching what they are doing, where they struggle, and how they succeed provides a great way to identify usability issues.

5.2.2 Automated Studies

Identifying usability issues through automated studies requires careful data collection. The key is to allow participants to enter verbatim comments at a page or task level. In most automated studies, several data points are collected for each task: success, time, ease-of-use rating, and verbatim comments. Verbatim comments are the best way to understand any possible issues.

One way to collect verbatim comments is to require the participant to provide a comment at the conclusion of each task. This might yield some interesting results, but it doesn't always yield the best results. An alternative that seems to work better is to make the verbatim comment conditional. If the participant provides a low ease-of-use score (e.g., not one of the two highest ratings), then she is asked to provide feedback about why she rated the task that way. Having a more pointed question usually yields more specific, actionable comments. For example, participants might say that they were confused about a particular term or that they couldn't find the link they wanted on a certain page. This type of task-level feedback is usually more valuable than one question after they complete all the tasks (post-study). The only downside of this approach is if the participant adjusts his ratings, after several questions, in order to avoid the open-ended question.

5.3 SEVERITY RATINGS

Not all usability issues are the same: Some are more serious than others. Some usability issues mildly annoy or frustrate users, whereas others cause them to make the wrong decisions or lose data. Obviously, these two different types of usability issues have a very different impact on the user experience, and severity ratings are a useful way to deal with them.

Severity ratings help focus attention on the issues that really matter. There's nothing more frustrating for a developer or business analyst than being handed a list of 82 usability issues that all need to be fixed immediately. By prioritizing usability issues, you're much more likely to have a positive impact on the design, not to mention lessening the likelihood of making enemies with the rest of the design and development team.

The severity of usability issues can be classified in many ways, but most severity rating systems can be boiled down to two different types. In one type of rating system, severity is based purely on the impact on the user experience: The worse the user experience, the higher the severity rating. A second type of severity rating system tries to bring in multiple dimensions or factors, such as business goals and technical implementation costs.

5.3.1 Severity Ratings Based on the User Experience

Many severity ratings are based solely on the impact on the user experience. These rating systems are easy to implement and provide very useful information. They usually have three levels —often something like low, medium, and high severity. Occasionally there is a "catastrophe" level, which is essentially a showstopper (delaying product launch or release—Nielsen, 1993).

When choosing a severity rating system, it's important to look at your organization and the product you are evaluating. Often, a three-level system works well in many situations:

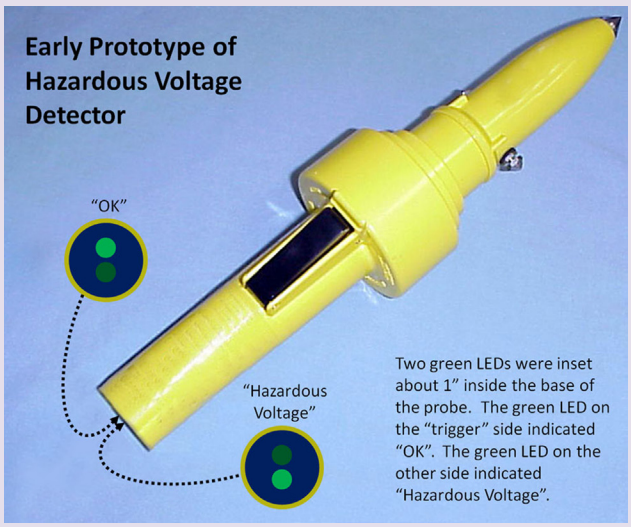
Low: Any issue that annoys or frustrates participants but does not play a role in task failure. These are the types of issues that may lead someone off course, but he still recovers and completes the task. This issue may only reduce efficiency and/ or satisfaction a small amount, if any.
Medium: Any issue that contributes to significant task difficulty but does not cause task failure. Participants often develop workarounds to get to what they need. These issues have an impact on effectiveness and most likely efficiency and satisfaction.
High: Any issue that leads directly to task failure. Basically, there is no way to encounter this issue and still complete the task. This type of issue has a significant impact on effectiveness, efficiency, and satisfaction.

Note that this scheme is a rating of task failure, one of the measures of user experience. In a test in which there are no task failures, there can be no high severity issues.

Another limitation of a three-level scheme from low to high is that user experience professionals often are reluctant to use the "low" category, fearing that those issues may be ignored. That limits the scale to two levels.

AN EXAMPLE OF THE ULTIMATE ISSUE SEVERITY

Tullis (2011) described an example of what we consider the ultimate in issue severity. In the early 1980s he conducted a usability test of a prototype of a handheld device for detecting high voltage on a metallic surface. The device had two indicator lights: one simply indicated that the device is working and the other indicated that there is high voltage present, which could be fatal. Unfortunately, both indicator lights were green. And they were right next to each other. And neither was labeled. After pleading with the designers to change the design, he finally decided to do a quick usability test. He had 10 participants perform 10 simulated tasks with the device. The prototype was rigged to signal the hazardous voltage condition 20% of the time. Out of 100 participant tasks, the indicator lights were interpreted correctly 99 times. But that one error was when it was signaling hazardous voltage. This usability issue could have resulted in serious injury or death to the user. The designers were convinced and the design was changed significantly.



5.3.2 Severity Ratings Based on a Combination of Factors

Severity rating systems that use a combination of factors usually are based on the impact on the user experience coupled with frequency of use and/or impact on the business goals. Nielsen (1993) provides an easy way to combine the impact on the user experience and frequency of use on severity ratings (Figure 5.1). This severity rating system is intuitive and easy to explain.

	Few users experiencing a problem	Many users experiencing a problem
Small impact on the user experience	Low severity	Medium severity
Large impact on the user experience	Medium severity	High severity

Figure 5.1 Severity rating scale taking into account problem frequency and impact on the user experience. Adapted from Nielsen (1993).

Alternatively, it's possible to consider three or even four dimensions, such as impact to the user experience, predicted frequency of occurrence, impact on the business goals, and technical/implementation costs. For example, you might combine four different three-point scales:

- Impact on the user experience (0 = low, 1 = medium, 2 = high)
- Predicted frequency of occurrence (0 = low, 1 = medium, 2 = high)
- Impact on the business goals (0 = low, 1 = medium, 2 = high)
- Technical/implementation costs (0 = high, 1 = medium, 2 = low)

By adding up the four scores, you now have an overall severity rating ranging from 0 to 8. Of course, a certain amount of guesswork is involved in coming up with the levels, but at least all four factors are being taken into consideration. Or, if you really want to get fancy, you can weight each dimension based on some sort of organizational priority.

5.3.3 Using a Severity Rating System

Once you have settled on a severity rating system, you still need to consider a few more things. First, be consistent: Decide on one severity rating system and use it for all your studies. By using the same severity rating system, you will be able to make meaningful comparisons across studies, as well as help train your audience on the differences between the severity levels. The more your audience internalizes the system, the more persuasive you will be in promoting design solutions.

Second, communicate clearly what each level means. Provide examples of each level as much as possible. This is particularly important for other usability specialists on your team who might also be assigning ratings. It's important that developers, designers, and business analysts understand each severity level. The more the "nonusability" audience understands each level, the easier it will be to influence design solutions for the highest priority issues.

Third, try to have more than one usability specialist assign severity ratings to each issue. One approach that works well is to have the usability specialists independently assign severity ratings to each of the issues and then discuss any of the issues where they gave different ratings and try to agree on the appropriate level.

Finally, there's some debate about whether usability issues should be tracked as part of a larger bug-tracking system (Wilson & Coyne, 2001). Wilson argues that it is essential to track usability issues as part of a bug-tracking system because it makes the usability issues more visible, lends more credibility to the usability team, and makes it more likely that the issues will be remedied. Coyne suggests

that usability issues, and the methods to fix them, are much more complex than typical bugs. Therefore, it makes more sense to track usability issues in a separate database. Either way, it's important to track the usability issues and make sure they are addressed, not simply forgotten.

5.3.4 Some Caveats about Rating Systems

Not everyone believes in severity ratings. Kuniavsky (2003) suggests letting your audience provide their own severity ratings. He argues that only those who are deeply familiar with the business model will be able to determine the relative priority of each usability issue.

Bailey (2005) strongly argues against severity rating systems altogether. He cites several studies that show there is very little agreement between usability specialists on the severity rating for any given usability issue (Catani & Biers, 1998; Cockton & Woolrych, 2001; Jacobsen, Hertzum, & John, 1998; Molich & Dumas, 2008). All of these studies generally show that there is very little overlap in what different usability specialists identify as a high-severity issue. Obviously, this is troubling given that many important decisions may be based on severity ratings.

Hertzum et al. (2002) highlight a potentially different problem in assigning severity ratings. In their research they found that when multiple usability specialists are working as part of the same team, each usability specialist rates the issues she personally identifies as more severe than issues identified by the other usability specialists on their own team. This is one aspect known as an evaluator effect, and it poses a significant problem in relying on severity ratings by a single UX professional. As a profession, we don't yet know why severity ratings are not consistent between specialists.

So where does this leave us? We believe that severity ratings are far from perfect, but they still serve a useful purpose. They help direct attention to at least some of the most pressing needs. Without severity ratings, the designers or developers will simply make their own priority list, perhaps based on what's easiest or least expensive to implement. Even though there is subjectivity involved in assigning severity ratings, they're better than nothing. We believe that most key stakeholders understand that there is more art than science involved, and they interpret the severity ratings within this broader context.

5.4 ANALYZING AND REPORTING METRICS FOR USABILITY ISSUES

Once you've identified and prioritized the usability issues, it's helpful to do some analyses of the issues themselves. This lets you derive some metrics related to the usability issues. Exactly how you do this will largely depend on the type of usability questions you have in mind. Three general questions can be answered by looking at metrics related to usability issues:

- How is the overall usability of the product? This is helpful if you simply want to get an overall sense of how the product did.

- Is the usability improving with each design iteration? Focus on this question when you need to know how the usability is changing with each new design iteration.
- Where should you focus your efforts to improve the design? The answer to this question is useful when you need to decide where to focus your resources.

All of the analyses we examine can be done with or without severity ratings. Severity ratings simply add a way to filter the issues. Sometimes it's helpful to focus on the high-severity issues. Other times it might make more sense to treat all the usability issues equally.

5.4.1 Frequency of Unique Issues

The simplest way to measure usability issues is to simply count the unique issues. Analyzing the frequency of unique issues is most useful in an iterative design process when you want some high-level data about how the usability is changing with each new design iteration. For example, you might observe that the number of unique issues decreased from 24 to 12 to 4 through the first three design iterations. These data are obviously trending in the right direction, but they're not necessarily iron-clad evidence that the design is significantly better. Perhaps the four remaining issues are so much bigger than all the rest that without addressing them, everything else is unimportant. Therefore, we suggest a thorough analysis and explanation of the issues when presenting this type of data.

Keep in mind that this frequency represents the number of *unique issues*, not the *total number of issues* encountered by all participants. For example, assume Participant A encountered 10 issues, whereas Participant B encountered 14 issues, but 6 of those issues were the same as those from Participant A. If A and B were the only participants, the total number of unique issues would be 18. Figure 5.2 shows an example of how to present the frequency of usability issues when comparing more than one design.

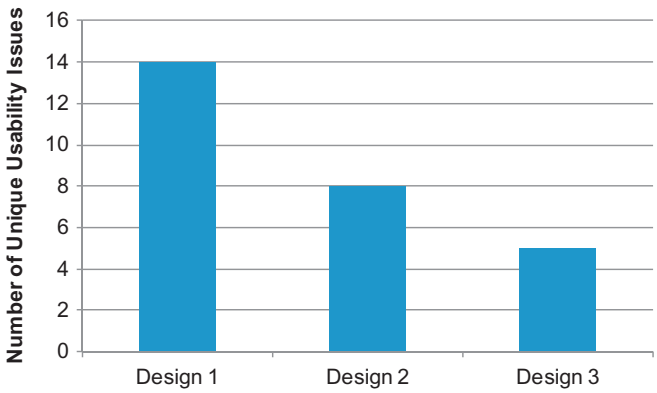


Figure 5.2 Example data showing the number of unique usability issues by design iteration.

The same type of analysis can be performed using usability issues that have been assigned a severity rating. For example, if you have classified your usability issues into three levels (low, medium, and high severity), you can easily look at the number of issues by each type of severity rating. Certainly the most telling data item would be the change in the number of high-priority issues with each design iteration. Looking at the frequency of usability issues by severity rating, as illustrated in Figure 5.3, can be very informative since it is an indicator of

whether the design effort between each iteration is addressing the most important usability issues.

5.4.2 Frequency of Issues Per Participant

It can also be informative to look at the number of (nonunique) issues each participant encountered. Over a series of design iterations, you would expect to see this number decreasing along with the total number of unique issues. For example, Figure 5.4 shows the average number of issues encountered by each participant for three design iterations. Of course, this analysis could also include the average number of issues per participant broken down by severity level. If the average number of issues per participant is steady over a series of iterations, but the total number of unique issues is declining, then you know there is more consistency in the issues that the participants are encountering. This would indicate that the issues encountered by fewer participants are being fixed, whereas those encountered by more participants are not.

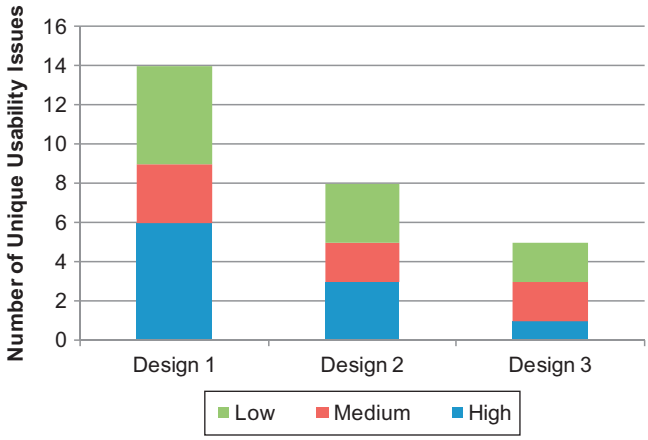


Figure 5.3 Example data showing the number of unique usability issues by design iteration, categorized by severity rating. The change in the number of high-severity issues is probably of key interest.

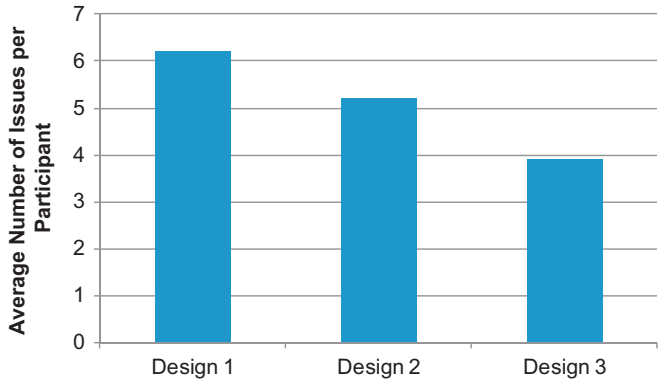


Figure 5.4 Example data showing the average number of usability issues encountered by participants in each of three usability tests.

5.4.3 Frequency of Participants

Another useful way to analyze usability issues is to observe the frequency or percentage of participants who encountered a specific issue. For example, you might be interested in whether participants correctly used some new type of navigation element on your website. You report that half of the participants encountered a specific issue in the first design iteration and only 1 out of 10 encountered the same issue in the second design iteration. This is a useful metric when you need to focus on whether you are improving the usability of specific design elements as opposed to making overall usability improvements.

With this type of analysis, it's important that your criteria for identifying specific issues are consistent between participants and designs. If a description of a specific issue is a bit fuzzy, your data won't mean very much. It's a good idea to explicitly document the issue's exact nature, thereby reducing any interpretation errors across participants or designs. Figure 5.5 shows an example of this type of analysis.

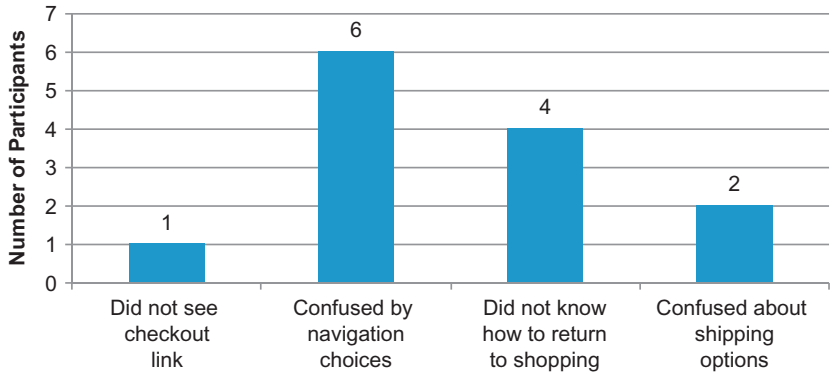


Figure 5.5 Example data showing the frequency of participants who experienced specific usability issues.

The use of severity ratings with this type of analysis is useful in a couple of ways. First, you could use the severity ratings to focus your analysis only on the high-priority issues. For example, you could report that there are five outstanding high-priority usability issues. Furthermore, the percentage of participants who are experiencing these issues is decreasing with each design iteration. Another form of analysis is to aggregate all the high-priority issues to report the percentage of participants who experienced any high-priority issue. This helps you see how overall usability is changing with each design iteration, but it is less helpful in determining whether to address a specific usability problem.

5.4.4 Issues by Category

Sometimes it's helpful to know where to focus design improvements from a tactical perspective. Perhaps you feel that only certain areas of the product are causing the most usability issues, such as navigation, content, terminology, and so forth. In this situation, it can be useful to aggregate usability issues into categories. Simply examine each issue and then categorize it into a type of issue. Then look at the frequencies of issues that fall into each category. Issues can be categorized in many different ways. Just make sure the categorization makes sense to you and your audience, and use a limited number of categories, typically three to eight. If there are too many categories, it won't provide much direction. Figure 5.6 provides an example of usability issues analyzed by category.

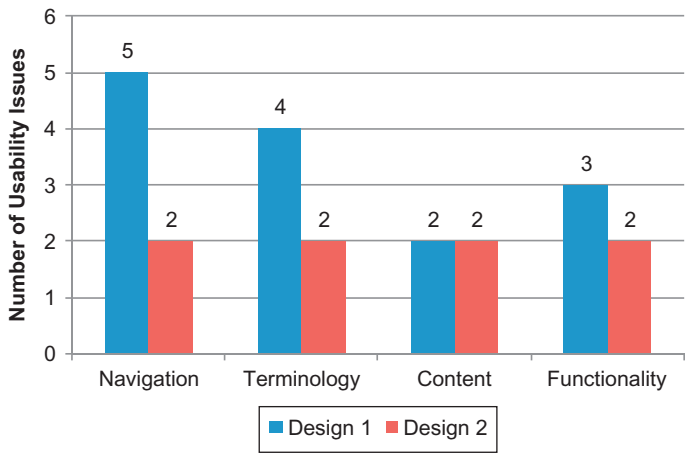


Figure 5.6 Example data showing the frequency of usability issues categorized by type. Note that both navigation and terminology issues were improved from the first to the second design iteration.

5.4.5 Issues by Task

Issues can also be analyzed at a task level. You might be interested in which tasks lead to the most issues, and you can report the number of unique issues that occur for each. This will identify the tasks you should focus on for the next design iteration. Alternatively, you could report the frequency of participants who encounter any issue for each task. This will tell you the pervasiveness of a particular issue. The greater the number of issues for each task, the greater the concern should be.

If you have assigned a severity rating to each issue, it might be useful to analyze the frequency of high-priority issues by task. This is particularly effective if you want to focus on a few of the biggest problems and your design efforts are oriented toward specific tasks. This is also helpful if you are comparing different design iterations using the same tasks.

5.5 CONSISTENCY IN IDENTIFYING USABILITY ISSUES

Much has been written about consistency and bias in identifying and prioritizing usability issues. Unfortunately, the news is not so good. Much of the research shows that there is very little agreement on what a usability issue is or how severe it is.

Perhaps the most exhaustive set of studies, called CUE (Comparative Usability Evaluation) has been coordinated by Rolf Molich. To date, nine separate CUE studies have been conducted, dating back to 1998. Each study was set up in a similar manner. Different teams of usability experts all evaluated the same design. Each team reported their findings, including the identification of the usability issues, along with their design recommendations. The first study, CUE-1 (Molich et al., 1998), showed very little overlap in the issues identified. In fact, only 1 out of the 141 issues was identified by all four teams participating in the study, and 128 out of the 141 issues were identified by single teams. Several years later, in CUE-2, the results were no more encouraging: 75% of all the issues were reported by only 1 of 9 usability teams (Molichet et al., 2004). CUE-4 (Molich & Dumas, 2008) showed similar results: 60% of all the issues were identified by only 1 of the 17 different teams participating in the study. More recently, CUE-8 focused on the consistency of how UX metrics are used and the conclusions that are drawn.

CUE-8—HOW PRACTITIONERS MEASURE WEBSITE USABILITY

by Rolf Molich, Dialog Design

Fifteen experienced professional usability teams simultaneously and independently measured a baseline for the usability of the car rental website Budget.com. This comparative study documented a wide difference in measurement approaches. The 8–10 teams that used similar and well-established approaches reached surprisingly similar results.

Fifteen Teams Measured the Same Website

In May 2009, 15 U.S. and European teams independently and simultaneously carried out usability measurements of the Budget.com car rental website. The goals were to investigate reproducibility of professional usability measurements and how experienced professionals actually carry out usability measurements.

The measurements were based on a common scenario and instructions. The scenario deliberately did not specify in detail which measures the teams were supposed to collect and report, although participants were asked to collect time-on-task, task success, and satisfaction data, as well as any qualitative data they normally would collect. The anonymous reports from the 15 participating teams are available publicly online (http://www.dialogdesign.dk/CUE-8.htm).

All teams were asked to measure the same five tasks in their study, for example, "Rent an intermediate size car at Logan Airport in Boston, Massachusetts, from Thursday 11 June 2009 at 09.00 AM to Monday 15 June at 3.00 PM. If asked for a name, use John Smith, email address john112233@hotmail.com. Do not submit the reservation."

Teams used from 9 to 313 test participants and from 21 to 128 hours to complete the study. Interestingly, the team that tested the most participants also spent the fewest hours on the study. This team used 21 person hours to conduct 313 sessions, which were all unmoderated.

Eight of the 15 teams used the SUS questionnaire for measuring subjective satisfaction. Despite its known shortcomings, SUS seems to be the current industry standard. No other questionnaire was used by more than one team.

Nine teams included qualitative results in addition to the required quantitative results. The general feeling seemed to be that the qualitative results were a highly useful by-product of the measurements.

The study is named CUE-8. It was the eighth in a series of Comparative Usability Evaluation studies (http://www.dialogdesign.dk/CUE.html).

Unmoderated Test Sessions

Six teams used unmoderated, automated measurements. Two of these six teams supplemented unmoderated measurements with moderated measurement sessions. These teams obtained valuable results but some also found that their data from the unattended test sessions were contaminated or invalid. Some participants reported impossible task times, perhaps because they wanted the reward with as little effort as possible.

Examples of contaminated data are 33 seconds to rent a car, which is impossible on the Budget.com website. The presence of obviously contaminated data in the data set raises serious doubts about the validity of all data in the data set. It's easy to spot unrealistic data, but how about a reported time of, for example, 146 seconds to rent a car in a data set that also contains unrealistic data? The 146 seconds look realistic, but how do you know that the unmoderated test participant did not use an unacceptable approach to arrive at the reported time?

Unmoderated measurements are attractive from a resource point of view; however, data contamination is a serious problem and it is not always clear what you are actually

measuring. While both moderated and unmoderated testing have opportunities for things to go wrong, it is more difficult to detect and correct these with unmoderated testing. Further studies of how data contamination can be prevented and how contaminated data can be cleaned efficiently are required.

For unmoderated measurements, the ease of use and intrusiveness of the remote tool influence measurements. Some teams complained about clunky interfaces. We recommend that practitioners demand usable products for usability measurements.

Practitioner's Takeaway from CUE-8

CUE-8 confirmed a number of rules for good measurement practice. Perhaps the most interesting result from CUE-8 is that these rules were not always observed by the participating professional teams.

- Adhere strictly to precisely defined measurement procedures for quantitative tests.
- Report time on task, success/failure rate, and satisfaction for quantitative tests.
- Exclude failed times from average task completion times.
- Understand the inherent variability from samples. Use strict participant screening criteria. Provide confidence intervals around your results if this is possible. Keep in mind that time on task is not distributed normally and therefore confidence intervals as commonly computed on raw scores may be misleading.
- Combine qualitative and quantitative findings in your report. Present what happened (quantitative data) and support it with why it happened (qualitative data). Qualitative data provide considerable insight regarding the serious obstacles that users faced and it is counterproductive not to report this insight.
- Justify the composition and size of your participant samples. This is the only way you have to allow your client to judge how much confidence they should place in your results.
- When using unmoderated methodologies for quantitative tests ensure that you can distinguish between extreme and incorrect results. Although unmoderated testing can exhibit a remarkable productivity in terms of user tasks measured with a limited effort, quantity of data is no substitute for clean data.

Further Information

The 17-page refereed paper "Rent a Car in Just 0, 60, 240 or 1,217 Seconds? Comparative Usability Measurement, CUE-8" describes the results of CUE-8 in detail. The paper is freely available in the November 2010 issue of the *Journal of Usability Studies*.

5.6 BIAS IN IDENTIFYING USABILITY ISSUES

Many different factors can influence how usability issues are identified. Carolyn Snyder (2006) provides a review of many of the ways usability findings might be biased. She concludes that bias cannot be eliminated, but it must be understood. In other words, even though our methods have flaws, they are still useful.

We've distilled the different sources of bias in a usability study into seven general categories:

Participants: Your participants are critical. Every participant brings a certain level of technical expertise, domain knowledge, and motivation. Some participants may be well targeted and others may not. Some participants are comfortable in a lab setting, whereas others are not. All of these factors make a big difference in what usability issues you end up discovering.

Tasks: The tasks you choose have a tremendous impact on what issues are identified. Some tasks might be well defined with a clear end state, others might be open ended, and yet others might be self-generated by each participant. The tasks basically determine what areas of the product are exercised and the ways in which they are exercised. Particularly with a complex product, this can have a major impact on what issues are uncovered.

Method: The method of evaluation is critical. Methods might include traditional lab testing or some type of expert review. Other decisions you make are also important, such as how long each session lasts, whether the participant thinks aloud, or how and when you probe.

Artifact: The nature of the prototype or product you are evaluating has a huge impact on your findings. The type of interaction will vary tremendously whether it is a paper prototype, functional or semifunctional prototype, or production system.

Environment: The physical environment also plays a role. The environment might involve direct interaction with the participant, indirect interaction via a conference call or behind a one-way mirror, or even at someone's home. Other characteristics of the physical environment, such as lighting, seating, observers behind a one-way mirror, and videotaping, can all have an impact on the findings.

Moderators: Different moderators will also influence the issues that are observed. A UX professional's experience, domain knowledge, and motivation all play a key role.

*Expectations:*Norgaard and Hornbaek (2006) found that many usability professionals come into testing with expectations on what are the most problematic areas of the interface. These expectations have a significant impact on what they report, often times missing many other important issues.

An interesting study that sheds some light on these sources of bias was conducted by Lindgaard and Chattratichart (2007). They analyzed the reports from the nine teams in CUE-4 who conducted actual usability tests with real users. They looked at the number of participants in each test, the number of tasks used, and the number of usability issues reported. They found no significant correlation between the number of *participants* in the test and the percentage of usability problems found. However, they did find a significant correlation between the number of *tasks* used and the percentage of usability problems found ($r = 0.82$, $p < 0.01$). When looking at the percentage of *new* problems uncovered, the correlation with the number of tasks was even higher ($r = 0.89$, $p < 0.005$). As Lindgaard and Chattratichart (2007) concluded, these results suggest "that with

careful participant recruitment, investing in wide task coverage is more fruitful than increasing the number of users."

One technique that works well to increase task coverage in a usability test is to define a set of tasks that all participants must complete and another set that is derived for each participant. These additional tasks might be selected based on characteristics of the participant (e.g., an existing customer or a prospect) or might be selected at random. Care must be exercised when making comparisons across participants, as not all participants had the same tasks. In this situation, you may want to limit certain analyses to the core tasks.

THE SPECIAL CASE OF MODERATOR BIAS IN AN EYE-TRACKING STUDY

One of the more difficult aspects of moderating a usability study is controlling where you look during the session. Moderators usually are looking either at the participant or their interaction on a screen, or some other interface. This works well, except in the case of an eye-tracking study. Most eye-tracking studies measure where participants look, and whether participants are noting key elements on the interface. As a moderator, it can be difficult *not* to look at the target when the participant is scanning the interface. Participants can pick up on this easily, begin to notice where *you* are looking, and use that information as a guide to target. It happens very quickly and subtly. While this behavior has not been reported in the user experience literature, we have observed it during our own eye-tracking studies. The best thing to do is to be aware of it, and if you find your eyes starting to wander to the target, simply refocus on the participant, what they are doing, or, if you have to, some other element on the page. Or don't sit in the room with the participant, if that's an option. When you're sitting with the participant in an eye-tracking study, there's also a greater chance that the participant will naturally look at you and away from the screen.

5.7 NUMBER OF PARTICIPANTS

There has been much debate about how many participants are needed in a usability test to reliably identify usability issues. [For a summary of the debate, see Barnum et al. (2003).] Nearly every UX professional seems to have an opinion. Not only are many different opinions floating around out there, but quite a few compelling studies have been conducted on this very topic. From this research, two different camps have emerged: those who believe that five participants are enough to identify most of the usability issues and those who believe that five is nowhere near enough.

5.7.1 Five Participants is Enough

One camp believes that a majority, or about 80%, of usability issues will be observed with the first five participants (Lewis, 1994; Nielsen & Landauer, 1993;

Virzi, 1992). This is known as the "magic number 5." One of the most important ways to figure out how many participants are needed in a usability test is to measure *p*, or the probability of a usability issue being detected by a single test participant. It's important to note that this *p* is different from the *p* value used in tests of significance. The probabilities vary from study to study, but they tend to average around 0.3, or 30%. [For a review of different studies, see Turner, Nielsen, and Lewis (2002).] In a seminal paper, Nielsen and Landauer (1993) found an average probability of 31% based on 11 different studies. This basically means that with each participant, about 31% of the usability problems are being observed.

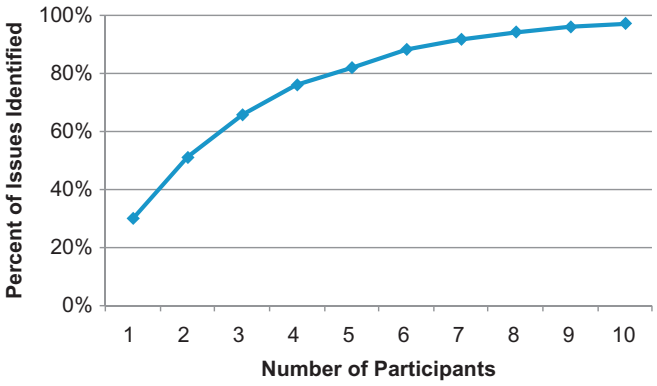


Figure 5.7 Example showing how many users are required to observe the total number of issues in a usability study, given a probability of detection.

Figure 5.7 shows how many issues are observed as a function of the number of participants when the probability of detection is 30%. (Note that this assumes all issues have an equal probability of detection, which may be a big assumption.) As you can see, after the first participant, 30% of the problems are detected; after the third participant, about 66% of the problems are observed; and after the fifth participant, about 83% of the problems have been identified. This claim is backed up not only by this mathematical formula, but by a\necdotal evidence as well. Many UX professionals only test with five or six participants during an iterative design process. In this situation, it is relatively uncommon to test with more than a dozen, with a few exceptions. If the scope of the product is particularly large or if there are distinctly different audiences, then a strong case can be made for testing with more than five participants.

CALCULATING *P*, OR PROBABILITY OF DETECTION

Calculating the probability of detection is fairly straightforward. Simply line up all the usability issues discovered during the test. Then, for each participant, mark how many of the issues were observed with that participant. Add the total number of issues identified with each participant and then divide by the total number of issues. Each test participant will have encountered anywhere from 0 to 100% of the issues. Then, take the average for all the test participants. This is the overall probability rate for the test. Consider the example shown in this table.

Participant	Issue 1	Issue 2	Issue 3	Issue 4	Issue 5	Issue 6	Issue 7	Issue 8	Issue 9	Issue 10	Proportion
P1	X		X		X		X	X	X		0.6
P2	X	X		X		X					0.4
P3			X			X		X	X	X	0.5
P4	X	X			X			X	X	X	0.6
P5				X	X		X				0.3
P6	X					X		X	X		0.4
P7		X	X		X			X	X	X	0.6
P8	X		X	X	X		X	X		X	0.7
P9		X	X	X		X	X		X		0.6
P10		X			X						0.2
Proportion	**0.5**	**0.5**	**0.5**	**0.4**	**0.6**	**0.4**	**0.4**	**0.6**	**0.6**	**0.4**	**0.49**

Once the average proportion has been determined (0.49 in this case), the next step is to calculate how many users are needed to identify a certain percentage of issues. Use the following formula:

$$1 - (1 - p)^n,$$

where n is the number of users.

So if you want to know the proportion of issues that would be identified by a sample of three users:

- $1-(1-0.49)^3$
- $1-(0.51)^3$
- $1-0.133$
- 0.867, or about 87%, of the issues would be identified with a sample of three users from this study

5.7.2 Five Participants is *Not* Enough

Other researchers have challenged this idea of the magic number 5 (Molich et al., 1998; Spool & Schroeder, 2001; Woolrych & Cockton, 2001). Spool and Schroeder (2001) asked participants to purchase various types of products, such as CDs and DVDs, at three different electronics websites. They discovered only 35% of the usability issues after the first five participants—far lower than the

80% predicted by Nielsen (2000). However, in this study the scope of the websites being evaluated was very large, even though the task of buying something was very well defined. Woolrych and Cockton (2001) discount the assertion that five participants are enough, primarily because it does not take into account individual differences.

The analyses by Lindgaard and Chattratichart (2007) of the nine usability tests from CUE-4 also raise doubts about the magic number 5. They compared the results of two teams, A and H, that both did very well, uncovering 42 and 43%, respectively, of the full set of usability problems. Team A used only 6 participants, whereas Team H used 12. At first glance, this might be seen as evidence for the magic number 5, as a team that tested only 6 participants uncovered as many problems as a team that tested 12. But a more detailed analysis reveals a different conclusion. In looking specifically at the overlap of usability issues between just these two reports, they found only 28% in common. More than 70% of the problems were uncovered by only one of the two teams, ruling out the possibility of the five-participant rule applying in this case.

THE EVALUATOR EFFECT

The Evaluator Effect (Hornbaek & Frokjaer, 2008; Jacobson, Hertzum, & John, 1998; Vermeern, van Kesteren, & Bekker, 2003) in usability testing suggests that UX professionals identify a different set of usability issues. In other words, there is little agreement or overlap in the usability issues identified by different UX professionals. The evaluator effect has been observed consistently in the CUE studies led by Rolf Molich (http://www.dialogdesign.dk/CUE.html). Most recently, CUE-9 (Molich, 2011) focused on the Evaluator Effect. Most of the 34 test team leaders in CUE-9 were confident that they found the most significant usability issues. However, there was little overlap in the issues. Furthermore, the test teams felt that running more participants would not help them identify more usability issues.

How do we reconcile this finding in the context of the recommended number of participants? It is easy for a UX professional to say that he found most of the usability issues after testing with 5–10 participants. In fact, they are usually very confident. But how do they really know unless they compare their findings to another UX professional? The fact is, they don't. It is quite possible that additional usability issues, often significant, may be uncovered with an independent assessment by another UX professional.

5.7.3 *Our* Recommendation

We recommend maintaining flexibility regarding sample sizes in usability tests. We feel it may be acceptable to test with 5–10 participants, using one UX team when the following conditions are met:

- It is OK to miss some of the major usability issues. You are more interested in capturing some of the big issues, iterating the design, and retesting. Any improvement is welcome.
- There is only one distinct user group that you believe will think about the design and tasks in a very similar way.
- The scope of the design is limited. There are a small number of screens and/or tasks.

We recommend increasing the number of participants and/or the number of UX teams when the following conditions apply:

- You must capture as many UX issues as possible. In other words, there will be significant negative repercussions if you miss any of the major usability issues.
- There is more than one distinct user group.
- The scope of the design is large. In this case we would recommend a broader set of tasks.

We fully realize that not everyone has access to multiple UX researchers. In this case, try to solicit feedback from any other observers. No one can see everything. Also, you might want to acknowledge that some of the major usability issues might not have been identified.

5.8 SUMMARY

Many usability professionals make their living by identifying usability issues and by providing actionable recommendations for improvement. Providing metrics around usability issues is not commonly done, but it can be incorporated easily into anyone's routine. Measuring usability issues helps you answer some fundamental questions about how good (or bad) the design is, how it is changing with each design iteration, and where to focus resources to remedy the outstanding problems. You should keep the following points in mind when identifying, measuring, and presenting usability issues.

1. The easiest way to identify usability issues is during an in-person lab study, but it can also be done using verbatim comments in an automated study. The more you understand the domain, the easier it will be to spot the issues. Having multiple observers is very helpful in identifying issues.
2. When trying to figure out whether an issue is real, ask yourself whether there is a consistent story behind the user's thought process and behavior. If the story is reasonable, then the issue is likely to be real.
3. The severity of an issue can be determined in several ways. Severity always should take into account the impact on the user experience. Additional factors, such as frequency of use, impact on the business, and persistence, may also be considered. Some severity ratings are based on a simple high/medium/low rating system. Other systems are number based.
4. Some common ways to measure usability issues are measuring the frequency of unique issues, the percentage of participants who experience

a specific issue, and the frequency of issues for different tasks or categories of issue. Additional analysis can be performed on high-severity issues or on how issues change from one design iteration to another.

5. When identifying usability issues, questions about consistency and bias may arise. Bias can come from many sources, and there can be a general lack of agreement on what constitutes an issue. Therefore, it's important to work collaboratively as a team, focusing on high-priority issues, and to understand how different sources of bias impact conclusions. Maximizing task coverage and including an additional UX team may be key.

CHAPTER 6

Self-Reported Metrics

CONTENTS

Measuring the User Experience. DOI: http://dx.doi.org/10.1016/B978-0-12-415781-1.00006-6

Perhaps the most obvious way to learn about the usability of something is to ask the participants to tell you about their experience with it. But exactly how to ask participants so that you get good data is not so obvious. The questions you might ask could take on many forms, including various kinds of rating scales, lists of attributes that the participants choose from, and open-ended questions such as "List the top three things you liked the most about this application." Some of the attributes you might ask about include overall satisfaction, ease of use, effectiveness of navigation, awareness of certain features, clarity of terminology, visual appeal, trust in a company that sponsors a website, enjoyment in playing a game, and many others. But the common feature of all of these is you're asking the participant for information, which is why we think *self-reported* best describes these metrics. And as we will see, one critical type of self-reported data is the verbatim comments made by participants while using a product.

THE EVOLUTION OF USABILITY AND USER EXPERIENCE

One of the historical precedents for the usability field was human factors, or ergonomics, which itself grew primarily out of World War II and a desire to improve airplane cockpits to minimize pilot error. With this ancestry, it's not surprising that much of the early focus of usability was on performance data (e.g., speed and accuracy). But that has been changing, quite significantly we think. Part of the reason for the widespread adoption of the term "user experience," or UX, is the focus that it provides on the entire range of experience that the user has with a product. Even the Usability Professionals Association changed its name in 2012 to the User Experience Professionals Association. All of this reflects the importance of the kind of metrics discussed in this chapter, which try to encompass such states as delight, joy, trust, fun, challenge, anger, frustration, and many more. An interesting analysis was done by Bargas-Avila and Hornbæk (2011) of 66 empirical studies in the UX literature from 2005 to 2009 showing how the studies reflect some of these shifts. They found, for example, that emotions, enjoyment, and aesthetics were the most frequently assessed UX dimensions in the recent studies.

Two other terms sometimes used to describe this kind of data include *subjective data* and *preference data*. *Subjective* is used as a counterpart to *objective*, which is often used to describe performance data from a usability study. But this implies that there's a lack of objectivity to the data you're collecting. Yes, it may be subjective to each participant who's providing the input, but from the perspective of the user experience professional, it is completely objective. Similarly, *preference* is often used as a counterpart to *performance*. Although there's nothing obviously wrong with that, we believe that preference implies a choice of one option over another, which is often not the case in UX studies.

6.1 IMPORTANCE OF SELF-REPORTED DATA

Self-reported data give you the most important information about users' *perception* of the system and their interaction with it. At an emotional level, the data may tell you something about how the users *feel* about the system. In many situations, these kinds of reactions are the main thing that you care about. Even if it takes users forever to perform something with a system, if the experience makes them happy, that may be the only thing that matters.

Your goal is to make the users think of your product first. For example, when deciding what travel-planning website to use for an upcoming vacation, users are more likely to think of the site that they liked the last time they used it. They're much less likely to remember how long the process was or that it took more mouse-clicks than it should have. That's why users' subjective reactions to a website, product, or store may be the best predictor of their likelihood to return or make a purchase in the future.

6.2 RATING SCALES

One of the most common ways to capture self-reported data in a UX study is with some type of rating scales. Two of the classic approaches to rating scales are the Likert scale and the semantic differential scale.

6.2.1 Likert Scales

A typical item in a Likert scale is a statement to which respondents rate their level of agreement. The statement may be positive (e.g., "The terminology used in this interface is clear") or negative (e.g., "I found the navigation options confusing"). Usually a five-point scale of agreement like the following is used:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

In the original version of the scale, Likert (1932) provided "anchor terms" for each point on the scale, such as Agree, and did not use numbers. Some people prefer to use a seven-point scale, but it gets a bit more difficult to come up with descriptive terms for each point as you get to higher numbers. This is one reason many researchers have dropped the intervening labels and just label the two ends (or anchor points) and perhaps the middle, or neutral, point. Many variations on Likert scales are still used today, but most Likert-scale purists would say that the two main characteristics of a Likert scale are (1) it expresses a degree of agreement with a statement and (2) it uses an odd number of response options, thus allowing a neutral response. By convention, the "Strongly Agree" end of a Likert scale is generally shown on the right when presented horizontally.

In designing the *statements* for Likert scales, you need to be careful how you word them. You should avoid adverbs such as *very*, *extremely*, or *absolutely* in the statements and use unmodified versions of adjectives. For example, the statement "This website is beautiful" may yield results that are quite different from "This website is absolutely beautiful," which may decrease the likelihood of strong agreement.

WHO WAS LIKERT?

Many people have heard of Likert scales, but not many know where the name came from or even how to pronounce it! It's pronounced "LICK-ert," not "LIKE-ert." This type of scale is named for Rensis Likert, who created it in 1932.

6.2.2 Semantic Differential Scales

The semantic differential technique involves presenting pairs of bipolar, or opposite, adjectives at either end of a series of scales, such as the following:

Weak	o	o	o	o	o	o	o	Strong
Ugly	o	o	o	o	o	o	o	Beautiful
Cool	o	o	o	o	o	o	o	Warm
Amateur	o	o	o	o	o	o	o	Professional

Like the Likert scale, a five- or seven-point scale is commonly used. The difficult part about the semantic differential technique is coming up with words that are truly opposites. Sometimes a thesaurus can be helpful since it includes antonyms. But you need to be aware of the connotations of different pairings of words. For example, a pairing of "Friendly/Unfriendly" may have a somewhat different connotation and yield different results from "Friendly/Not Friendly" or "Friendly/Hostile."

OSGOOD'S SEMANTIC DIFFERENTIAL

The semantic differential technique was developed by Charles E. Osgood (Osgood et al., 1957), who designed it to measure the connotations of words or concepts. Using factor analysis of large sets of semantic differential data, he found three recurring attitudes that people used in assessing words and phrases: evaluation (such as "good/bad"), potency (such as "strong/weak"), and activity (such as "passive/active").

6.2.3 When to Collect Self-Reported Data

During a usability study, you might collect self-reported data in the form of verbatim comments from a think-aloud protocol while the participants are interacting with the product. Two additional times when you might want to probe more explicitly for self-reported data are immediately after each task (post-task ratings) and at the end of the entire session (poststudy ratings). Poststudy ratings tend to be the more common, but both have advantages. Quick ratings immediately after each task can help pinpoint tasks and parts of the interface that are particularly problematic. More in-depth ratings and open-ended questions at the end of the session can provide an effective overall evaluation after the participant has had a chance to interact with the product more fully.

6.2.4 How to Collect Ratings

Logistically, three techniques can be used to collect self-reported data in a usability test: answer questions or provide ratings orally, record responses on a paper form, or provide responses using some type of online tool. Each technique has its advantages and disadvantages. Having the participant provide responses orally is the easiest method from the participant's perspective, but, of course, it means that an observer needs to record the responses, and may introduce some bias as participants sometimes feel uncomfortable verbally stating poor ratings. This works best for a single, quick rating after each task.

Paper forms and online forms are suitable both for quick ratings and for longer surveys. Paper forms may be easier to create than online, but they involve manual entry of data, including the potential for errors in interpreting handwriting. Online forms are getting easier to create, as evidenced by the number of web-based questionnaire tools available, and participants are getting more accustomed to using them. One technique that works well is to have a laptop computer or perhaps tablet computer with the online questionnaire next to the participant's computer in the usability lab. The participant can then refer to the application or website easily while completing the online survey.

ONLINE SURVEY TOOLS

Many tools are available for creating and administering surveys via the web. Doing a search on "online survey tools" turns up a pretty extensive list. Some of them include Google Docs' Forms, Qualtrics.com, SnapSurveys.com, SurveyGizmo.com, SurveyMonkey.com, SurveyShare.com, and Zoomerang.com. Most of these tools support a variety of question types, including rating scales, check boxes, drop-down lists, grids, and open-ended questions. These tools generally have some type of free trial or other limited-functionality subscription that lets you try out the service for free.

6.2.5 Biases in Collecting Self-Reported Data

Some studies have shown that people who are asked directly for self-reported data, either in person or over the phone, provide more positive feedback than when asked through an anonymous web survey (e.g., Dillman et al., 2008). This is called the social desirability bias (Nancarrow & Brace, 2000), in which respondents tend to give answers they believe will make them look better in the eyes of others. For example, people who are called on the phone and asked to evaluate their satisfaction with a product typically report higher satisfaction than if they reported their satisfaction levels in a more anonymous way. Telephone respondents or participants in a usability lab essentially want to tell us what they think we want to hear, and that is usually positive feedback about our product.

Therefore, we suggest collecting post-test data in such a way that the moderator or facilitator does not see the user's responses until after the participant has left. This might mean either turning away or leaving the room when the user fills out the automated or paper survey. Making the survey itself anonymous may also elicit more honest reactions. Some UX researchers have suggested asking participants in a usability test to complete a post-test survey after they get back to their office or home. This can be done by giving them a paper survey and a postage-paid envelope to mail it back or by e-mailing a pointer to an online survey. The main drawback of this approach is that you will typically have some drop-off in terms of who completes the survey. Another drawback is that it increases the amount of time between users' interaction with the product and their evaluation via the survey, which could have unpredictable results.

6.2.6 General Guidelines for Rating Scales

Crafting good rating scales and questions is difficult; it's both an art and a science. So before you go off on your own, look at existing sets of questions, such as those in this chapter, to see if you can't use those instead. But if you decide that you need to create your own, here are some general points to consider:

- **Multiple scales to help "triangulate."** When creating scales to assess a specific attribute such as visual appeal, credibility, or responsiveness, the main thing to remember is that you will probably get more reliable data if you can think of a few different ways to ask participants to assess the attribute. In analyzing the results, you would average those responses together to arrive at the participant's overall reaction for that attribute. Likewise, the success of questionnaires that include both positive and negative statements to which participants respond would suggest the value of including both types of statements.
- **Odd or even number of values?** The number of values to use in rating scales can be a source of heated debate among UX professionals. Many of the arguments center on the use of an even or odd number of points on the scale. An odd number of points has a center, or neutral, point, whereas an even number does not, thus forcing the user slightly toward one end or the other on the scale. We believe that in most real-world

situations a neutral reaction is a perfectly valid reaction and should be allowed on a rating scale. So in most cases we use rating scales with an odd number of points. However, there's some indication that not including a midpoint may minimize the effect of the social desirability bias in a face-to-face administration of rating scales (e.g., Garland, 1991).

- **Total number of points.** The other issue, of course, is the actual number of points to use on the rating scales. Some people seem to believe "more is always better," but we don't really agree with that. The survey literature suggests that any more than nine points rarely provides useful additional information (e.g., Cox, 1980; Friedman & Friedman, 1986). In practice, we almost always use five or seven points.

IS FIVE POINTS ENOUGH FOR A RATING SCALE?

Kraig Finstad (2010) did an interesting study comparing five- and seven-point versions of the same set of rating scales [the System Usability Scale (SUS), discussed later in this chapter]. The ratings were administered orally. He counted the number of times that the participant answered with an "interpolation," such as 3.5, 3½, or "between 3 and 4." In other words, the participant wanted to pick a value *between* two of the values given on the scale. He found that participants using the five-point version of the scale were significantly more likely to use interpolations than those using the seven-point version. In fact, about 3% of the individual ratings on the five-point scale were interpolations, while *none* of the ratings on the seven-point scale were. This would suggest that verbal (and perhaps paper-based) rating scales, where the participant could be tempted to use interpolations, might yield more accurate results with seven-point scales.

SHOULD YOU NUMBER SCALE VALUES?

One of the issues that comes up in designing rating scales is whether to show the user a numeric value for each scale position. Our sense is that with scales of no more than five or seven values, adding numbers for each position is not necessary. But as you increase the number of scale values, numbers might become more useful in helping the user keep track of where she or he is on the scale. But don't use something like −3, −2, −1, 0, +1, +2, +3. Studies have shown that people tend to avoid using zero or negative values (e.g., Sangster et al., 2001; Schwartz et al., 1991).

6.2.7 Analyzing Rating-Scale Data

The most common technique for analyzing data from rating scales is to assign a numeric value to each of the scale positions and then compute the averages. For example, in the case of a five-point Likert scale, you might assign a value of 1 to the "Strongly Disagree" end of the scale and a value of "5" to the "Strongly Agree" end. These averages can then be compared across different

tasks, studies, user groups, and so on. This is common practice among most UX professionals as well as market researchers. Even though rating-scale data are not technically interval data, many professionals treat it as interval. For example, we assume the distance between a 1 and a 2 on a Likert scale is the same as the distance between a 2 and a 3 on the same scale. This assumption is called *degrees of intervalness*. We also assume that a value *between* any two of the scale positions has meaning. The bottom line is that it is close enough to interval data that we can treat it as such.

When analyzing data from rating scales, it's always important to look at the actual frequency distribution of the responses. Because of the relatively small number of response options (e.g., 5–9) for each rating scale, it's even more important to look at the distribution than it is for truly continuous data such as task times. You might see important information in the distribution of responses that you would totally miss if you just looked at the average. For example, let's assume you asked 20 users to rate their agreement with the statement "This website is easy to use" on a 1 to 7 scale, and the resulting average rating was 4 (right in the middle). You might conclude that the users were basically just lukewarm about the site's ease of use. But then you look at the distribution of the ratings and you see that 10 users rated it a "1" and 10 rated it a "7". So, in fact, no one was lukewarm. They either thought it was great or they hated it. You might then want to do some segmentation analysis to see if the people who hated it have anything in common (e.g., they had never used the site before) vs the people who loved it (e.g., long-time users of the site).

WHAT NUMBER SHOULD RATING SCALES START WITH?

Regardless of whether you show numbers for each scale value to the *user*, you will normally use numbers *internally* for analysis. But what number should the scales start with, zero or one? It generally doesn't matter, as long as you report what the scale is whenever showing mean ratings (e.g., a mean of 3.2 on a scale of 1 to 5). But there are some cases where it's convenient to start the scale at zero, particularly if you want to express the ratings as percentages of the best possible rating. On a scale of 1 to 5, a rating of 5 would correspond to 100%, but a rating of 1 does not correspond to 20%, as some might think (e.g., calculating the percentage by multiplying the rating by 20, which is wrong). On a scale of 1 to 5, 1 is the lowest possible rating, so it should correspond to 0%. Consequently, we find it's easier to keep our sanity by internally numbering rating scales starting at zero, so that a rating of 0 corresponds to 0%.

Another way to analyze rating-scale data is by looking at top-box or top-2-box scores. Assume you're using a rating scale of 1 to 5, with 5 meaning "Strongly Agree." The sample data in Figure 6.1 illustrate the calculation of top-box and top-2-box scores. A top-box score would be the percentage of participants who gave a rating of 5. Similarly, a top-2-box score would be the

percentage of participants who gave a rating of 4 or 5. (Top-2-box scores are used more commonly with larger scales, such as 7 or 9 points.) The theory behind this method of analysis is that it lets you focus on how many participants gave very positive ratings. (Note that the analysis can also be done as a bottom-box or bottom-2-box analysis, focusing on the other extreme.) Keep in mind that when you convert to a top-box or top-2-box score, the data can no longer be considered interval. Therefore, you should just report the data as frequencies (e.g., the percentage of users who gave a top-box rating). Also keep in mind that you lose information by calculating a top-box or top-2-box score. Lower ratings are ignored by this analysis.

C2 f_x =IF(B2>4,1,0)

	A	B	C	D
1	**Participant**	**Rating (1-5)**	**Top Box?**	**Top 2 Box?**
2	P1	4	0	1
3	P2	5	1	1
4	P3	3	0	0
5	P4	4	0	1
6	P5	2	0	0
7	P6	3	0	0
8	P7	5	1	1
9	P8	4	0	1
10	P9	3	0	0
11	P10	5	1	1
12	**Averages**	**3.8**	**30%**	**60%**

Figure 6.1 Example of the calculation of top-box and top-2-box scores from ratings in Excel. The "=IF" function in Excel is used to check whether an individual rating is greater than 4 (for Top Box) or greater than 3 (for Top 2 Box). If it is, a value of "1" is given. If not, a value of "0" is given. Averaging these 1's and 0's together gives you the percentage of Top-Box or Top-2-Box scores.

From a practical standpoint, what difference does it make when you analyze rating scales using means vs top-box or top-2-box scores? To illustrate the difference, we looked at data from an online study conducted on the eve of the 2008 U.S. presidential election (Tullis, 2008). There were two leading candidates, Barack Obama and John McCain, both of whom had websites about their candidacy. Participants were asked to perform the same four tasks on one of the sites (which they were assigned to randomly). After each task, they were asked to rate how easy it was on a scale of 1 to 5, with 1 = Very Difficult and 5 = Very Easy. A total of 25 participants performed tasks on the Obama site and 19 on the McCain site. We then analyzed the task ease ratings by calculating the means, top-box scores, and top-2-box scores. The results are shown in Figure 6.2.

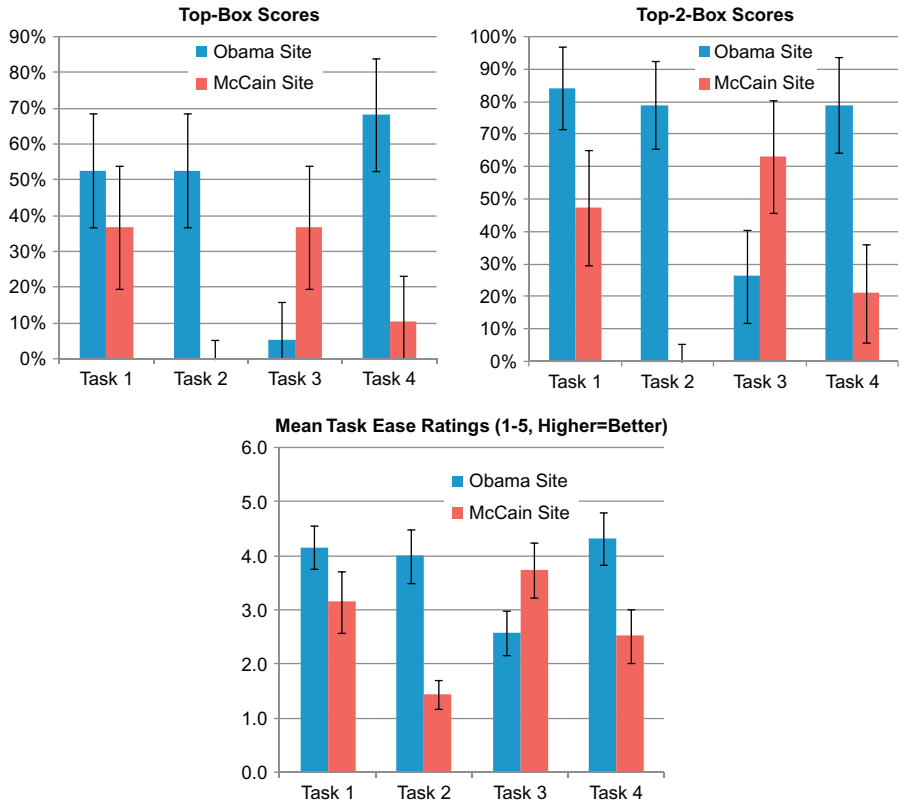


Figure 6.2 Three different analyses of task ease ratings from a study of the Obama and McCain websites (Tullis, 2008): mean ratings, top-2-box scores, and top-box scores. Note how similar patterns are revealed in all three analysis methods, but the apparent disparity between the two sites differs. In each chart, error bars represent a 90% confidence interval.

All three charts seem to indicate that the Obama site got a higher rating than the McCain site for three tasks (Tasks 1, 2, and 4), while the McCain site got a higher rating than the Obama site for one task (Task 3). However, the apparent disparity between the two sites differs depending on the analysis method. There tends to be a greater difference between the two sites with the top-box and top-2-box scores compared to the means. (And no, that's not an error in the top-box and top-2-box charts for Task 2. *None* of the participants gave that task a top-box or top-2-box rating for the McCain site.) But also note that the error bars tend to be larger with the top-box and top-2-box scores compared to the means.

Should you analyze rating scales using means or top-box scores? In practice, we generally use means because they take all data into account (not ignoring some ratings as in top-box or top-2-box analyses). But because some companies or senior executives are more familiar with top-box scores (often from market

research), we use top-box scores in some situations. (It's always important to understand who you're presenting your results to.)

HOW DO YOU CALCULATE CONFIDENCE INTERVALS FOR TOP-BOX SCORES?

If you're calculating means of ratings, then you can calculate confidence intervals in the same way you do for any other continuous data: using the "=CONFIDENCE" function in Excel. But if you're calculating top-box or top-2-box scores, it's not so simple. When you calculate a top-box or top-2-box value for each rating, you're turning it into binary data: each rating is either a top-box value (or top-2-box value) or it's not. This is obvious from Figure 6.1, where each of the top-box (or top-2-box) values is either a "0" or a "1". This should ring some mental bells: it's like the task success data that we examined in Chapter 4. When dealing with binary data, confidence intervals need to be calculated using the Adjusted Wald Method. See Chapter 4 for details.

6.3 POST-TASK RATINGS

The main goal of ratings associated with each task is to give you some insight into which tasks the participants thought were the most difficult. This can then point you toward parts of the system or aspects of the product that need improvement. One way to capture this information is to ask the participant to rate each task on one or more scales. The next few sections examine some of the specific techniques that have been used. For example, the data shown in Figure 6.2 show that users of the Obama site rated Task 3 as the most difficult, while users of the McCain site rated Task 2 as the most difficult.

6.3.1 Ease of Use

Probably the most common rating scale involves simply asking users to rate how easy or how difficult each task was. This typically involves asking them to rate the task using a five- or seven-point scale. Some UX professionals prefer to use a traditional Likert scale, such as "This task was easy to complete" (1 = Strongly Disagree, 3 = Neither Agree nor Disagree, 5 = Strongly Agree). Others prefer to use a semantic differential technique with anchor terms such as "Easy/Difficult." Either technique will provide you with a measure of perceived usability on a task level. Sauro and Dumas (2009) tested a single seven-point rating scale, which they coined the "Single Ease Question":

Overall, this task was?

Very Difficult o o o o o o o Very Easy

They compared it to several other post-task ratings and found it to be among the most effective.

6.3.2 After-Scenario Questionnaire (ASQ)

Jim Lewis (1991) developed a set of three rating scales—the After-Scenario Questionnaire—designed to be used after the user completes a set of related tasks or a scenario:

1. "I am satisfied with the ease of completing the tasks in this scenario."
2. "I am satisfied with the amount of time it took to complete the tasks in this scenario."
3. "I am satisfied with the support information (online help, messages, documentation) when completing the tasks."

Each of these statements is accompanied by a seven-point rating scale of "Strongly Disagree" to "Strongly Agree." Note that these questions in the ASQ touch upon three fundamental areas of usability: effectiveness (question 1), efficiency (question 2), and satisfaction (all three).

6.3.3 Expectation Measure

Albert and Dixon (2003) proposed a different approach to assessing users' subjective reactions to each task. Specifically, they argued that the most important thing about each task is how easy or difficult it was *in comparison to* how easy or difficult the user *thought* it was going to be. So before the users actually did any of the tasks, they asked them to rate how easy/difficult they *expect* each of the tasks to be, based simply on their understanding of the tasks and the type of product. Users expect some tasks to be easier than others. For example, getting the current quote on a stock should be easier than rebalancing an entire portfolio. Then, after performing each task, the users were asked to rate how easy/difficult the task *actually was*. The "before" rating is called the *expectation* rating, and the "after" rating is called the *experience* rating. They used the same seven-point rating scales (1 = Very Difficult, 7 = Very Easy) for both ratings. For each task you can then calculate an average *expectation rating* and an average *experience rating*. You can then visualize these two scores for each task as a scatterplot, as shown in Figure 6.3.

Figure 6.3 Comparison of average expectation ratings and average experience ratings for a set of tasks in a usability test. Which quadrants the tasks fall into can help you prioritize which tasks to focus on improving. Adapted from Albert and Dixon (2003); used with permission.

The four quadrants of the scatterplot provide some interesting insight into the tasks and where you should focus your attention when making improvements:

1. In the lower right are the tasks that the users thought would be *easy* but actually turned out to be *difficult*. These probably represent the tasks that are the biggest dissatisfiers for the users—those that were the biggest disappointment. These are the tasks you should focus on first, which is why this is called the "Fix It Fast" quadrant.
2. In the upper right are the tasks that the users thought would be *easy* and actually *were* easy. These are working just fine. You don't want to "break" them by making changes that would have a negative impact. That's why this is called the "Don't Touch It" quadrant.
3. In the upper left are the tasks that the users thought would be *difficult* and actually were *easy*. These are pleasant surprises, both for the users and the designers of the system! These could represent features of your site or system that may help distinguish you from the competition, which is why this is called the "Promote It" quadrant.
4. In the lower left are the tasks that the users thought would be *difficult* and actually *were* difficult. There are no big surprises here, but there might be some important opportunities to make improvements. That's why this is called the "Big Opportunities" quadrant.

6.3.4 A Comparison of Post-task Self-Reported Metrics

Tedesco and Tullis (2006) compared a variety of task-based self-reported metrics in an online usability study. Specifically, they tested the following five different methods for eliciting self-reported ratings after each task.

- *Condition 1:* "Overall, this task was: Very Difficult ….. Very Easy." This was a very simple post-task rating scale that many usability teams commonly use.
- *Condition 2:* "Please rate the usability of the site for this task: Very Difficult to Use ….. Very Easy to Use." Obviously, this is very similar to Condition 1 but with an emphasis on the usability of *the site* for the task. Perhaps only usability geeks detect the difference, but we wanted to find out!
- *Condition 3:* "Overall, I am satisfied with the ease of completing this task: Strongly Disagree ….. Strongly Agree"and "Overall, I'm satisfied with the amount of time it took to complete this task: Strongly Disagree ….. Strongly Agree." These are two of the three questions used in Lewis's (1991) ASQ. The third question in the ASQ asks about support information, such as online help, which was not relevant in this study, so it was not used.
- *Condition 4:* (Before doing all tasks): "How difficult or easy do you expect this task to be? Very Difficult ….. Very Easy" (After doing each task): "How difficult or easy did you find this task to be? Very Difficult ….. Very Easy." This is the expectation measure by Albert and Dixon (2003).
- *Condition 5:* "Please assign a number between 1 and 100 to represent how well the website supported you for this task. Remember: 1 would mean that the site was not at all supportive and completely unusable.

A score of 100 would mean that the site was perfect and would require absolutely no improvement." This condition was loosely based on a method called Usability Magnitude Estimation (McGee, 2004) in which test participants are asked to create their own "usability scale."

These techniques were compared in an online study. The participants performed six tasks on a live application used to look up information about employees (phone number, location, manager, etc.). Each participant used only one of the five self-report techniques. A total of 1131 people participated in the online study, with at least 210 participants using each self-report technique.

The main goal of this study was to see if these rating techniques are sensitive to detecting differences in perceived difficulty of the tasks. But we also wanted to see how the perceived difficulty of the tasks corresponded to the task performance data. We collected task time and binary success data (i.e., whether users found the correct answer for each task and how long that took). As shown in Figure 6.4, there were significant differences in the performance data across the tasks. Task 2 appears to have been the most challenging, whereas Task 4 was the easiest.

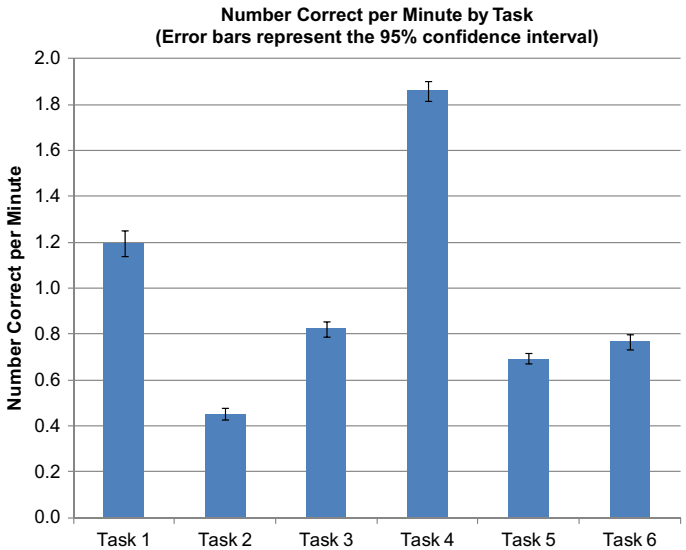


Figure 6.4 Performance data showing that users had the most difficulty with Task 2 and the least difficulty with Task 4. Adapted from Tedesco and Tullis (2006); used with permission.

As shown in Figure 6.5, a somewhat similar pattern of the tasks was reflected by the task ratings (averaged across all five techniques). In comparing task performance with task ratings, correlations were significant for all five conditions ($p < 0.01$). Overall, Spearman rank correlation comparing performance data and task ratings for the six tasks was significant: $Rs = 0.83$.

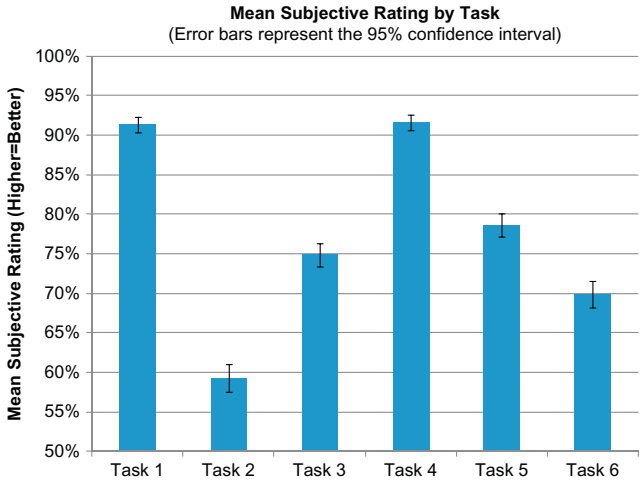


Figure 6.5 Average subjective ratings across all techniques. Ratings are all expressed as a percentage of the maximum possible rating. Similar to the performance data, Task 2 yielded the worst ratings, whereas Task 4 yielded among the best. Adapted from Tedesco and Tullis (2006); used with permission.

Figure 6.6 shows the averages of task ratings for each of the tasks, split out by condition. The key finding is that the pattern of the results was very similar regardless of which technique was used. This is not surprising, given the very large sample (total *N* of 1131). In other words, at large sample sizes, all five of the techniques can effectively distinguish between the tasks.

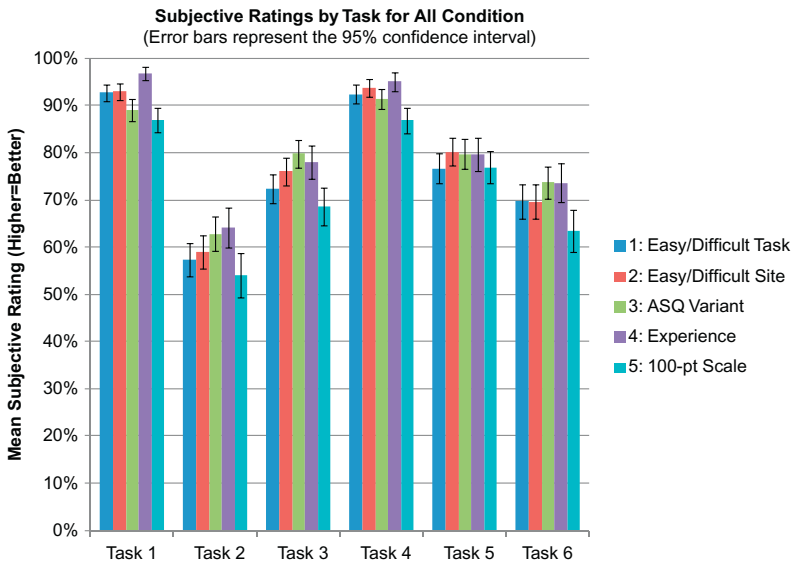


Figure 6.6 Average subjective ratings split by task and condition. All five conditions (self-report techniques) yielded essentially the same pattern of results for the six tasks. Adapted from Tedesco and Tullis (2006); used with permission.

But what about the smaller sample sizes more typical of usability tests? To answer that question, we did a subsampling analysis looking at large numbers of random samples of different sizes taken from the full data set. The results of this are shown in Figure 6.7, where the correlation between the data from the subsamples and the full data set is shown for each subsample size.

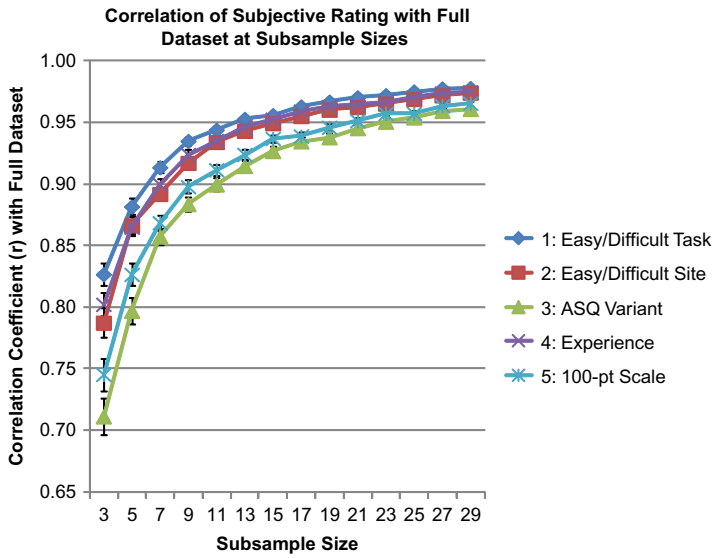


Figure 6.7 Results of a subsampling analysis showing average correlations between ratings for the six tasks from subsamples of various sizes and the full data set for each condition. Error bars represent the 95% confidence interval for the mean. Adapted from Tedesco and Tullis (2006); used with permission.

The key finding was that one of the five conditions, Condition 1 resulted in better correlations starting at the smallest sample sizes and continuing. Even at a sample size of only seven, which is typical of many usability tests, its correlation with the full data set averaged 0.91, which was significantly higher than any of the other conditions. So Condition 1, which was the simplest rating scale ("Overall, this task was Very Difficult…Very Easy"), was also the most reliable at smaller sample sizes.

RATINGS DURING A TASK?

At least one study (Teague et al., 2001) indicated that you might get a more accurate measure of the user's experience with a task by asking for ratings *during* the conduct of the task. They found that participants' ratings of ease of use were significantly higher after the task was competed than during the task. It could be that task success changes participants' perception of how difficult the task was to complete.

6.4 POSTSESSION RATINGS

One of the most common uses of self-reported metrics is as an overall measure of perceived usability that participants are asked to give after having completed their interactions with the product. These can be used as an overall "barometer" of the usability of the product, particularly if you establish a track record with the same measurement technique over time. Similarly, these kinds of ratings can be used to compare multiple design alternatives in a single usability study or to compare your product, application, or website to the competition. Let's look at some of the postsession rating techniques that have been used.

6.4.1 Aggregating Individual Task Ratings

Perhaps the simplest way to look at overall perceived usability is to take an average of the individual task-based ratings. Of course, this assumes that you did in fact collect ratings (e.g., ease of use) after each task. If you did, then simply take an average of them. Or, if some tasks are more important than others, take a weighted average. Keep in mind that these data are different from one snapshot at the end of the session. By looking at self-reported data across all tasks, you're really taking an average perception as it changes over time. Alternatively, when you collect self-reported data just once at the end of the session, you are really measuring the participant's last impression of the experience.

This last impression is the perception they will leave with, which will likely influence any future decisions they make about your product. So if you want to measure perceived ease of use for the product based on individual task performance, then aggregate self-reported data from multiple tasks. However, if you're interested in knowing the lasting usability perception, then we recommend using one of the following techniques that takes a single snapshot at the end of the session.

6.4.2 System Usability Scale

One of the most widely used tools for assessing the perceived usability of a system or product is the System Usability Scale. It was originally developed by John Brooke in 1986 while he was working at Digital Equipment Corporation (Brooke, 1996). As shown in Figure 6.8, it consists of 10 statements to which users rate their level of agreement. Half the statements are worded positively and half are worded negatively. A five-point scale of agreement is used for each. A technique for combining the 10 ratings into an overall score (on a scale of 0 to 100) is also given. It's convenient to think of SUS scores as percentages, as they are on a scale of 0 to 100, with 100 representing a perfect score.

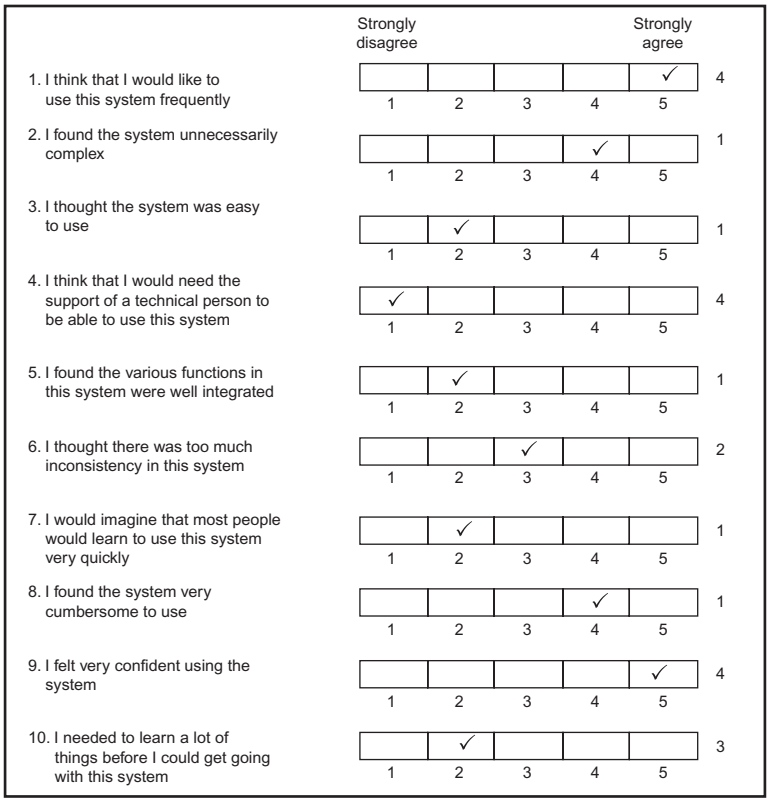


Figure 6.8 The System Usability Scale, developed by John Brooke at Digital Equipment Corporation and an example of scoring it.

CALCULATING A SUS SCORE

To calculate a SUS score, first sum the score contributions from each item. Each item's score contribution will range from 0 to 4. For items 1, 3, 5, 7, and 9, the score contribution is the scale position minus 1. For items 2, 4, 6, 8, and 10, the contribution is 5 minus the scale position. Multiply the sum of the scores by 2.5 to obtain the overall SUS score. Consider the sample data in Figure 6.8. The sum of the values, using these rules, is 22. Multiply that by 2.5 to get the overall SUS score of 55 or, better yet, download our spreadsheet for calculating SUS scores from www.MeasuringUX.com.

The SUS has been made freely available for use in usability studies, both for research purposes and for industry use. The only prerequisite for its use is that any published report should acknowledge the source of the measure. Because it has been so widely used, quite a few studies in the usability literature have

reported SUS scores for many different products and systems, including desktop applications, websites, voice-response systems, and various consumer products. Tullis (2008) and Bangor, Kortum, and Miller (2009) both reported analyses of SUS scores from a wide variety of studies. Tullis (2008a) reported data from 129 different uses of SUS, while Bangor and colleagues (2009) reported data from 206. Frequency distributions of the two sets of data are remarkably similar, as shown in Figure 6.9, with a median study score of 69 for Tullis data and 71 for Bangor et al. data. Bangor and colleagues suggested the following interpretation of SUS scores based on their data:

- <50: Not acceptable
- 50–70: Marginal
- >70: Acceptable

FACTORS IN SUS

Although SUS was originally designed to assess perceived usability as a single attribute, Lewis and Sauro (2009) found that there are actually two factors in SUS. Eight of the questions reflect a usability factor and two reflect a learnability factor. It's easy to compute both from raw SUS ratings.

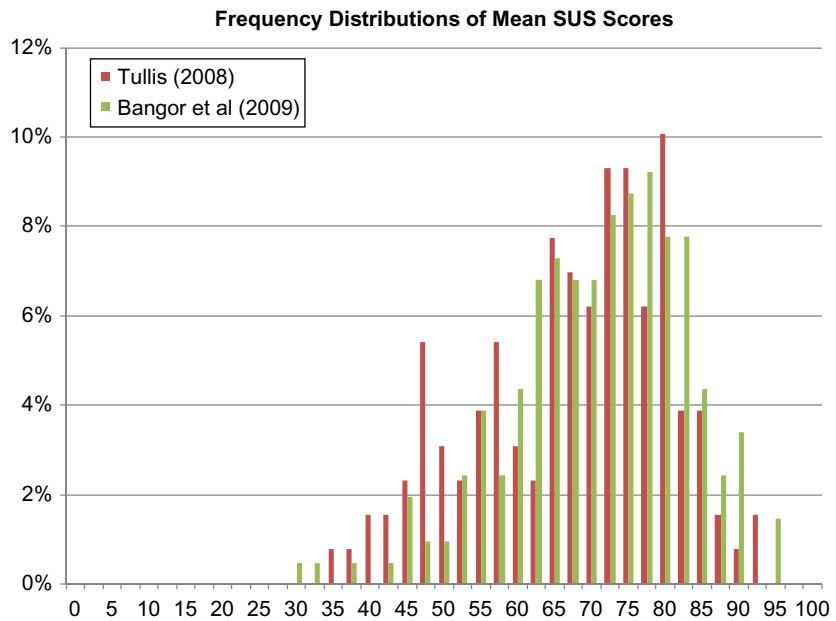


Figure 6.9 Frequency distributions of mean SUS scores reported by Tullis (2008a) and by Bangor et al. (2009). Tullis data are based on a total of 129 study conditions, and Bangor et al. data are based on 206.

DO YOU NEED BOTH POSITIVE AND NEGATIVE STATEMENTS IN SUS?

As shown in Figure 6.8, half of the statements in SUS are positive and half are negative. While some argue that this approach keeps participants "on their toes," others have argued that it also seems to confuse some participants, perhaps causing erroneous responses. Sauro and Lewis (2011) conducted a study in which they compared the traditional version of SUS to an all-positive version. They found no significant difference between mean SUS scores for traditional and all-positive versions. But in a review of 27 SUS data sets, they found evidence that 11% of the studies had some miscoding of SUS data and 13% of the individual SUS questionnaires contained mistakes by users. They suggest using an all-positive version of SUS to avoid some of those possible errors. If you want to use the all-positive version, see Sauro and Lewis (2011) for an example.

6.4.3 Computer System Usability Questionnaire

Jim Lewis (1995), who developed the ASQ technique for post-task ratings, also developed the Computer System Usability Questionnaire (CSUQ) to do an overall assessment of a system at the end of a usability study. The CSUQ is very similar to Lewis's Post-Study System Usability Questionnaire (PSSUQ), with only minor changes in wording. The PSSUQ was originally designed to be administered in person, whereas CSUQ was designed to be administered by mail or online. CSUQ consists of the following 19 statements to which the user rates agreement on a seven-point scale of "Strongly Disagree" to "Strongly Agree," plus N/A:

1. Overall, I am satisfied with how easy it is to use this system.
2. It was simple to use this system.
3. I could effectively complete the tasks and scenarios using this system.
4. I was able to complete the tasks and scenarios quickly using this system.
5. I was able to efficiently complete the tasks and scenarios using this system.
6. I felt comfortable using this system.
7. It was easy to learn to use this system.
8. I believe I could become productive quickly using this system.
9. The system gave error messages that clearly told me how to fix problems.
10. Whenever I made a mistake using the system, I could recover easily and quickly.
11. The information (such as online help, on-screen messages, and other documentation) provided with this system was clear.
12. It was easy to find the information I needed.
13. The information provided for the system was easy to understand.
14. The information was effective in helping me complete the tasks and scenarios.
15. The organization of information on the system screens was clear.
16. The interface of this system was pleasant.
17. I liked using the interface of this system.

18. This system has all the functions and capabilities I expect it to have.
19. Overall, I am satisfied with this system.

Unlike SUS, all of the statements in CSUQ are worded positively. Factor analyses of a large number of CSUQ and PSSUQ responses have shown that the results may be viewed in four main categories: System Usefulness, Information Quality, Interface Quality, and Overall Satisfaction.

6.4.4 Questionnaire for User Interface Satisfaction

The Questionnaire for User Interface Satisfaction (QUIS) was developed by a team in the Human–Computer Interaction Laboratory (HCIL) at the University of Maryland (Chin, Diehl, & Norman, 1988). As shown in Figure 6.10, QUIS

OVERALL REACTION TO THE SOFTWARE		0	1	2	3	4	5	6	7	8	9		NA
1.	terrible	○	○	○	○	○	○	○	○	○	○	wonderful	○
2.	difficult	○	○	○	○	○	○	○	○	○	○	easy	○
3.	frustrating	○	○	○	○	○	○	○	○	○	○	satisfying	○
4.	inadequate power	○	○	○	○	○	○	○	○	○	○	adequate power	○
5.	dull	○	○	○	○	○	○	○	○	○	○	stimulating	○
6.	rigid	○	○	○	○	○	○	○	○	○	○	flexible	○
SCREEN		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**		**NA**
7. Reading characters on the screen	hard	○	○	○	○	○	○	○	○	○	○	easy	○
8. Highlighting simplifies task	not at all	○	○	○	○	○	○	○	○	○	○	very much	○
9. Organization of information	confusing	○	○	○	○	○	○	○	○	○	○	very clear	○
10. Sequence of screens	confusing	○	○	○	○	○	○	○	○	○	○	very clear	○
TERMINOLOGY AND SYSTEM INFORMATION		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**		**NA**
11. Use of terms throughout system	inconsistent	○	○	○	○	○	○	○	○	○	○	consistent	○
12. Terminology related to task	never	○	○	○	○	○	○	○	○	○	○	always	○
13. Position of messages on screen	inconsistent	○	○	○	○	○	○	○	○	○	○	consistent	○
14. Prompts for input	confusing	○	○	○	○	○	○	○	○	○	○	clear	○
15. Computer informs about its progress	never	○	○	○	○	○	○	○	○	○	○	always	○
16. Error messages	unhelpful	○	○	○	○	○	○	○	○	○	○	helpful	○
LEARNING		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**		**NA**
17. Learning to operate the system	difficult	○	○	○	○	○	○	○	○	○	○	easy	○
18. Exploring new features by trial and error	difficult	○	○	○	○	○	○	○	○	○	○	easy	○
19. Remembering names and use of commands	difficult	○	○	○	○	○	○	○	○	○	○	easy	○
20. Performing tasks is straightforward	never	○	○	○	○	○	○	○	○	○	○	always	○
21. Help messages on the screen	unhelpful	○	○	○	○	○	○	○	○	○	○	helpful	○
22. Supplemental reference materials	confusing	○	○	○	○	○	○	○	○	○	○	clear	○
SYSTEM CAPABILITIES		**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**		**NA**
23. System speed	too slow	○	○	○	○	○	○	○	○	○	○	fast enough	○
24. System reliability	unreliable	○	○	○	○	○	○	○	○	○	○	reliable	○
25. System tends to be	noisy	○	○	○	○	○	○	○	○	○	○	quiet	○
26. Correcting your mistakes	difficult	○	○	○	○	○	○	○	○	○	○	easy	○
27. Designed for all levels of users	never	○	○	○	○	○	○	○	○	○	○	always	○
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**		**NA**

Figure 6.10 QUIS, developed by the HCIL at the University of Maryland. Commercial use requires a license from the Office of Technology Commercialization at the University of Maryland.

consists of 27 rating scales divided into five categories: Overall Reaction, Screen, Terminology/System Information, Learning, and System Capabilities. The ratings are on 10-point scales whose anchors change depending on the statement. The first 6 scales (assessing Overall Reaction) are polar opposites with no statements (e.g., Terrible/Wonderful, Difficult/Easy, Frustrating/Satisfying). QUIS can be licensed from the University of Maryland's Office of Technology Commercialization (http://www.lap.umd.edu/QUIS/index.html) and is available in printed and web versions in multiple languages.

GARY PERLMAN'S ONLINE QUESTIONNAIRES

Several of the questionnaires shown in this chapter, as well as a few others, are available for use online through a web interface created by Gary Perlman (http://www.acm.org/perlman/question.html). The questionnaires include QUIS, ASQ, and CSUQ. Options are provided for specifying which questionnaire to use, an e-mail address to submit results, and the name of the system being evaluated. These can be specified as parameters associated with the URL for the online questionnaire. So, for example, to specify the following:

Name of System: MyPage

Questionnaire: CSUQ

Send Results to: *me@gmail.com*

the URL would be http://www.acm.org/perlman/question.cgi?system=MyPage&form=CSUQ&email=me@gmail.com.

By default, all rating scales also provide a mechanism for the user to enter comments. Once the user clicks on the Submit button, data are e-mailed to the address specified, formatted in a name = value format, with one name and value per line.

6.4.5 Usefulness, Satisfaction, and Ease-of-Use Questionnaire

Arnie Lund (2001) proposed the Usefulness, Satisfaction, and Ease of Use (USE) questionnaire, shown in Figure 6.11, which consists of 30 rating scales divided into four categories: Usefulness, Satisfaction, Ease of Use, and Ease of Learning. Each is a positive statement (e.g., "I would recommend it to a friend"), to which the user rates level of agreement on a seven-point Likert scale. In analyzing a large number of responses using this questionnaire, he found that 21 of the 30 scales (identified in Figure 6.11) yielded the highest weights for each of the categories, indicating that they contributed most to the results.

Usefulness
- It helps me be more effective.
- It helps me be more productive.
- It is useful.
- It gives me more control over the activities in my life.
- It makes the things I want to accomplish easier to get done.
- It saves me time when I use it.
- *It meets my needs.*
- It does everything I would expect it to do.

Ease of Use
- It is easy to use.
- It is simple to use.
- It is user friendly.
- It requires the fewest steps possible to accomplish what I want to do with it.
- *It is flexible.*
- *Using it is effortless.*
- *I can use it without written instructions.*
- *I don't notice any inconsistencies as I use it.*
- *Both occasional and regular users would like it.*
- *I can recover from mistakes quickly and easily.*
- *I can use it successfully every time.*

Ease of Learning
- I learned to use it quickly.
- I easily remember how to use it.
- It is easy to learn to use it.
- *I quickly became skillful with it.*

Satisfaction
- I am satisfied with it.
- I would recommend it to a friend.
- It is fun to use.
- It works the way I want it to work.
- It is wonderful.
- I feel I need to have it.
- It is pleasant to use.

Users rate agreement with these statements on a seven-point Likert scale, ranging from strongly disagree to strongly agree. Statements in *italics* were found to weight less heavily than the others.

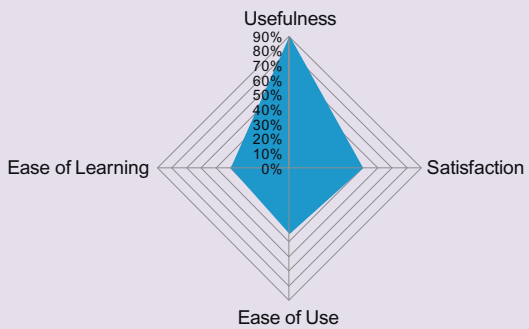
Figure 6.11 The USE questionnaire. From the work of Lund (2001); used with permission.

VISUALIZING DATA USING RADAR CHARTS

Some of the techniques for capturing self-reported data yield values on several dimensions. For example, the USE questionnaire can yield values for Usefulness, Satisfaction, Ease of Use, and Ease of Learning. Similarly, CSUQ can yield values for System Usefulness, Information Quality, Interface Quality, and Overall Satisfaction. One technique that can be useful for visualizing the results in a situation like this is a radar chart. Assume you had the following summary values from a study with the USE questionnaire:

- Usefulness = 90%
- Satisfaction = 50%
- Ease of Use = 45%
- Ease of Learning = 40%

Plotting these values as a radar chart would give you the chart shown here.



To create these charts, choose "Radar" (under "Other Charts") in Excel. "Filled" radar charts, like the example here, usually work best. The advantage these charts provide is that they help the viewer easily detect patterns as represented by different shapes. For example, a tall, skinny radar chart like the one shown here reflects the fact that users thought the product being evaluated was useful but not particularly easy to use, easy to learn, or satisfying.

6.4.6 Product Reaction Cards

A very different approach to capturing post-test subjective reactions to a product was presented by Joey Benedek and Trish Miner (2002) from Microsoft. As illustrated in Figure 6.12, they presented a set of 118 cards, each containing adjectives (e.g., Fresh, Slow, Sophisticated, Inviting, Entertaining, Incomprehensible). Some of the words are positive and some are negative. The users would then simply choose the cards they felt described the system. After selecting the cards, they were asked to pick the top five cards and explain why they chose each. This technique is intended to be more qualitative in that its main purpose is to elicit commentary from the users. But it can also be used in a quantitative way by counting the number of times each word is chosen by participants. Results can also be visualized using a word cloud (e.g., using Wordle.net). Case study 10.5 provides good examples of word clouds from the product reaction cards.

The complete set of 118 Product Reaction Cards				
Accessible	Creative	Fast	Meaningful	Slow
Advanced	Customizable	Flexible	Motivating	Sophisticated
Annoying	Cutting edge	Fragile	Not Secure	Stable
Appealing	Dated	Fresh	Not Valuable	Sterile
Approachable	Desirable	Friendly	Novel	Stimulating
Attractive	Difficult	Frustrating	Old	Straight Forward
Boring	Disconnected	Fun	Optimistic	Stressful
Business-like	Disruptive	Gets in the way	Ordinary	Time-consuming
Busy	Distracting	Hard to Use	Organized	Time-Saving
Calm	Dull	Helpful	Overbearing	Too Technical
Clean	Easy to use	High quality	Overwhelming	Trustworthy
Clear	Effective	Impersonal	Patronizing	Unapproachable
Collaborative	Efficient	Impressive	Personal	Unattractive
Comfortable	Effortless	Incomprehensible	Poor quality	Uncontrollable
Compatible	Empowering	Inconsistent	Powerful	Unconventional
Compelling	Energetic	Ineffective	Predictable	Understandable
Complex	Engaging	Innovative	Professional	Undesirable
Comprehensive	Entertaining	Inspiring	Relevant	Unpredictable
Confident	Enthusiastic	Integrated	Reliable	Unrefined
Confusing	Essential	Intimidating	Responsive	Usable
Connected	Exceptional	Intuitive	Rigid	Useful
Consistent	Exciting	Inviting	Satisfying	Valuable
Controllable	Expected	Irrelevant	Secure	
Convenient	Familiar	Low Maintenance	Simplistic	

Figure 6.12 Complete set of 118 product reaction cards developed by Joey Benedek and Trish Miner at Microsoft. From Microsoft: "Permission is granted to use this Tool for personal, academic, and commercial purposes. If you wish to use this Tool, or the results obtained from the use of this Tool for personal or academic purposes or in your commercial application, you are required to include the following attribution: Developed by and © 2002 Microsoft Corporation. All rights reserved."

6.4.7 A Comparison of Postsession Self-Reported Metrics

Tullis and Stetson (2004) conducted a study in which we compared a variety of postsession questionnaires for measuring user reactions to websites in an online usability study. We studied the following questionnaires, adapted in the manner indicated for the evaluation of websites.

- *SUS.* It was adapted by replacing the word *system* in every question with *website*.
- *QUIS.* Three of the original rating scales that did not seem to be appropriate to websites were dropped (e.g., "Remembering names and use of commands"). The term *system* was replaced with *website*, and the term *screen* was generally replaced by *web page*.
- *CSUQ.* The term *system* or *computer system* was replaced by *website*.
- *Microsoft's Product Reaction Cards.* Each word was presented with a check box, and the user was asked to choose the words that best describe their interaction with the website. They were free to choose as many or as few words as they wished.
- *Our Questionnaire.* We had been using this questionnaire for several years in usability tests of websites. It was composed of nine positive statements (e.g., "This website is visually appealing"), to which the user responds on a seven-point Likert scale from "Strongly Disagree" to "Strongly Agree."

We used these questionnaires to evaluate two web portals in an online usability study. There were a total of 123 participants in the study, with each participant using one of the questionnaires to evaluate both websites. Participants performed two tasks on each website before completing the questionnaire for that site. When we analyzed the data from all the participants, we found that all five of the questionnaires revealed that Site 1 got significantly better ratings than Site 2. Data were then analyzed to determine what the results would have been at different sample sizes from 6 to 14, as shown in Figure 6.13. At a sample size of 6, only 30 to 40% of the samples would have identified that Site 1 was preferred significantly. But at a sample size of 8, which is relatively common in many lab-based usability tests, we found that SUS would have identified Site 1 as the preferred site 75% of the time—a significantly higher percentage than any of the other questionnaires.

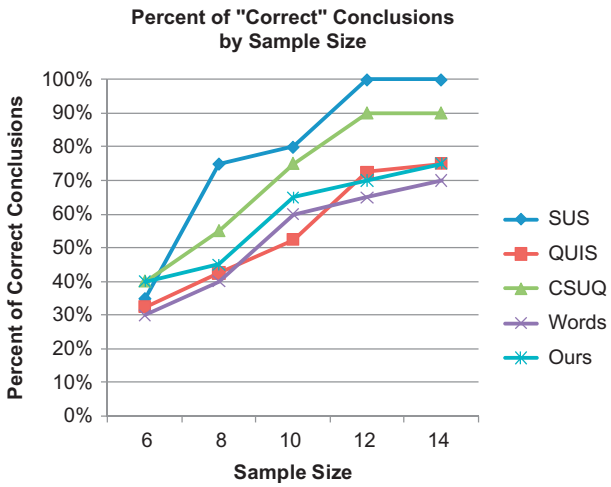


Figure 6.13 Data illustrating the accuracy of results from random subsamples ranging from size 6 to 14. This graph shows what percentage of the random samples yielded the same answer as the full data set at the different sample sizes. Adapted from Tullis and Stetson (2004); used with permission.

It's interesting to speculate why SUS appears to yield more consistent ratings at relatively small sample sizes. One reason may be its use of both positive and negative statements with which users must rate their level of agreement. This may keep participants more alert. Another possible reason

may be that it doesn't try to break down the assessment into more detailed components (e.g., ease of learning, ease of navigation). All 10 of the rating scales in SUS are simply asking for an assessment of the site as a whole, just in slightly different ways.

6.4.8 Net Promoter Score

One self-reported metric that has gained rapidly in popularity, especially among senior executives, is the Net Promoter Score (NPS). It's intended to be a measure of customer loyalty, and was originated by Fred Reichheld in his 2003 article in the *Harvard Business Review*: "One Number You Need to Grow" (Reichheld, 2003). The power of NPS seems to derive from its simplicity, as it uses only one question: "How likely is it that you would recommend [this company, product, website, etc] to a friend or colleague?" The respondent answers using an 11-point scale of 0 (Not at all likely) to 10 (Extremely likely). The respondents are then divided into three categories:

- Detractors: Those who gave ratings of 0–6
- Passives: Those who gave ratings of 7 or 8
- Promoters: Those who gave ratings of 9 or 10

Note that the categorization into Detractors, Passives, and Promoters is nowhere near symmetrical. By design, the bar is set pretty high to be a Promoter, while it's very easy to be a Detractor. To calculate the NPS, you subtract the percentage of Detractors (ratings of 0–6) from the percentage of Promoters (ratings of 9 or 10). Passives are ignored in the calculation. In theory, NPSs can range from −100 to +100.

The NPS is not without its own detractors. One criticism is that the reduction of scores from an 11-point scale to just three categories (Detractors, Passives, Promoters) results in a loss of statistical power and precision. This is similar to the loss of precision when using the "Top Box" or "Top-2-Box" method of analysis discussed earlier in this chapter. But you lose even more precision when you take the *difference* between two percentages (Promoters minus Detractors), which is similar to subtracting "Bottom Box" scores from "Top Box" scores. Each percentage (% Promoters and % Detractors) has its own confidence interval (or margin of error) associated with it. The confidence interval associated with the *difference* between the two percentages is essentially the combination of the two individual confidence intervals. You would typically need a sample size two to four times larger to get an NPS margin of error equivalent to the margin of error for a traditional Top-2-Box score. Case Study 10.1 provides an excellent example of how NPS can be used to improve the user experience.

DOES PERCEIVED USABILITY PREDICT CUSTOMER LOYALTY?

Jeff Sauro (2010) wanted to know whether usability, as measured by SUS, tended to predict customer loyalty, as measured by NPS. He analyzed data from 146 users asked to complete both the SUS questions and the NPS question for a variety of products, including websites and financial applications. The result was a correlation of $r = 0.61$, which is highly significant ($p < 0.001$). He found that Promoters had an average SUS score of 82, while Detractors had an average SUS score of 67.

6.5 USING SUS TO COMPARE DESIGNS

A number of usability studies that involved comparing different designs for accomplishing similar tasks have used the SUS questionnaire as one of the techniques for making the comparison (typically in addition to performance data).

Traci Hart (2004) of the Software Usability Research Laboratory at Wichita State University conducted a usability study comparing three different websites designed for older adults: SeniorNet, SeniorResource, and Seniors-Place. After attempting tasks on each website, participants rated each of them using the SUS questionnaire. The average SUS score for the SeniorResource site was 80%, which was significantly better than the average scores for SeniorNet and Seniors-Place, both of which averaged 63%.

The American Institutes for Research (2001) conducted a usability study comparing Microsoft's Windows ME and Windows XP. They recruited 36 participants whose expertise with Windows ranged from novice to intermediate. They attempted tasks using both versions of Windows and then completed the SUS questionnaire for both. They found that the average SUS score for Windows XP (74%) was significantly higher than the average for Windows ME (56%)($p < 0.0001$).

Sarah Everett, Michael Byrne, and Kristen Greene (2006), from Rice University, conducted a usability study comparing three different types of paper ballots: bubble, arrow, and open response. These ballots were based on actual ballots used in the 2004 U.S. elections. After using each of the ballots in a simulated election, the 42 participants used the SUS questionnaire to rate each one. They found that the bubble ballot received significantly higher SUS ratings than either of the other two ($p < 0.001$).

There's also some evidence that participants who have more experience with a product tend to give it higher SUS ratings than those with less experience. In testing two different applications (one web based and one desktop based), McLellan, Muddimer, and Peres (2012) found that the SUS scores from users who had more extensive experience with a product tended to be about 15% higher compared to users with either no or limited experience with the product.

6.6 ONLINE SERVICES

More and more companies are learning the value of getting feedback from the users of their websites. The currently in-vogue term for this process is listening to the "*Voice of the Customer*," or VoC studies. This is essentially the same process as in postsession self-reported metrics. The main difference is that VoC studies are typically done on live websites. The common approach is that a randomly selected percentage of live-site users get offered a pop-up survey asking for their feedback at a specific point in their interaction with the site—usually on logout, exiting the site, or completing a transaction. Another approach is to provide a standard mechanism for getting this feedback at various places in

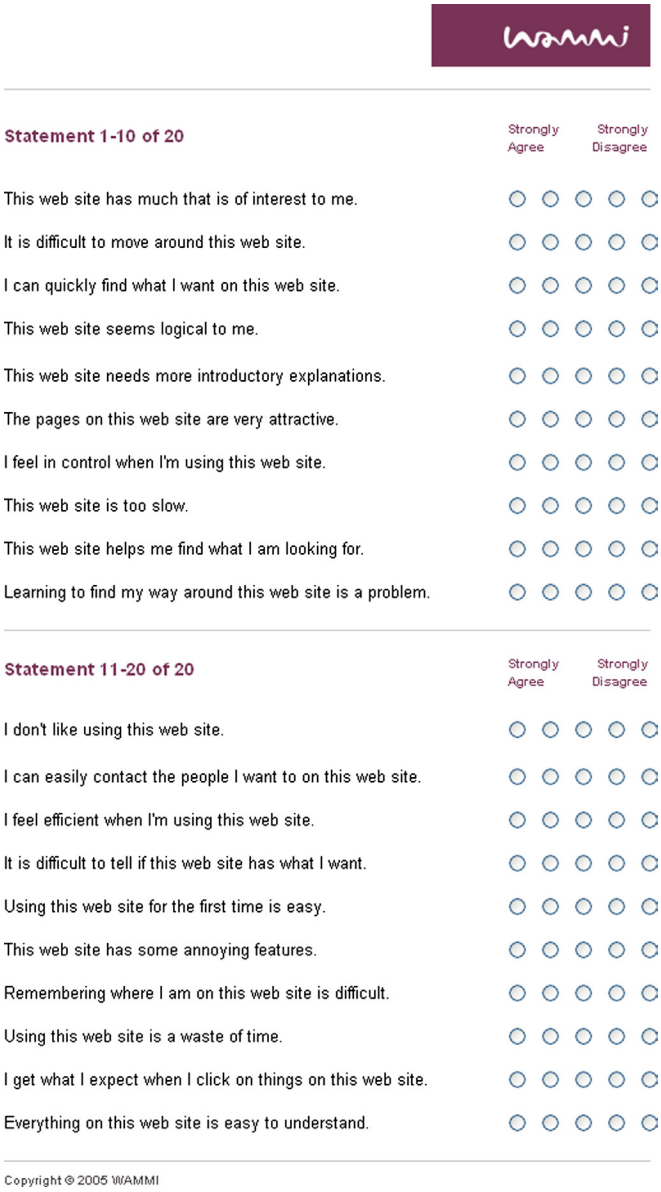


Figure 6.14 The 20 rating scales used by the WAMMI online service.

the site. The following sections present some of these online services. This list is not intended to be exhaustive, but it is at least representative.

6.6.1 Website Analysis and Measurement Inventory

The Website Analysis and Measurement Inventory (WAMMI—www.wammi.com) is an online service that grew out of an earlier tool called Software Usability Measurement Inventory (SUMI), both of which were developed in the Human Factors Research Group of University College Cork in Ireland. Although SUMI is designed for evaluation of software applications, WAMMI is designed for the evaluation of websites.

As shown in Figure 6.14, WAMMI is composed of 20 statements with associated five-point Likert scales of agreement. Like SUS, some of the statements are positive and some are negative. WAMMI is available in most European languages. The primary advantage that a service like WAMMI has over creating your own questionnaire and associated rating scales is that WAMMI has already been used in the evaluation of hundreds of websites worldwide. When used on your site, results are delivered in the form of a comparison against their reference database built from tests of these hundreds of sites.

Results from a WAMMI analysis, as illustrated in Figure 6.15, are divided into five areas: Attractiveness, Controllability, Efficiency, Helpfulness, and Learnability, plus an overall usability score. Each of these scores is standardized (from comparison to their reference database), so a score of 50 is average and 100 is perfect.

6.6.2 American Customer Satisfaction Index

The American Customer Satisfaction Index (ACSI—www.TheACSI.org) was developed at the Stephen M. Ross Business School of the University of Michigan.

It covers a wide range of industries, including retail, automotive, and manufacturing. The analysis of websites using the ACSI methodology is done by Foresee Results (www.ForeseeResults.com). The ACSI has become particularly popular for analyzing U.S. government websites. For example, 100 U.S. government websites were included in their fourth quarter 2012 analyses of e-government websites (ForeSee Results, 2012). Similarly, their annual Top 100 Online Retail Satisfaction Index assesses such popular sites as Amazon, NetFlix, L.L. Bean, J.C. Penney, Avon, and QVC.

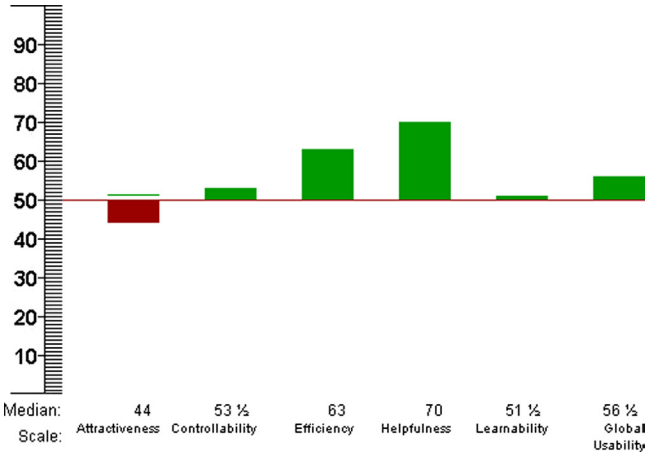


Figure 6.15 Sample data from the WAMMI online service showing average scores in each of five categories, plus an overall usability score.

The ACSI questionnaire for websites is composed of a core set of 14 questions, as shown in Figure 6.16. Each asks for a rating on a 10-point scale of different attributes, such as the quality of information, freshness of content, clarity of site organization, overall satisfaction, and likelihood to return. Specific implementations of the ACSI commonly add additional questions or rating scales.

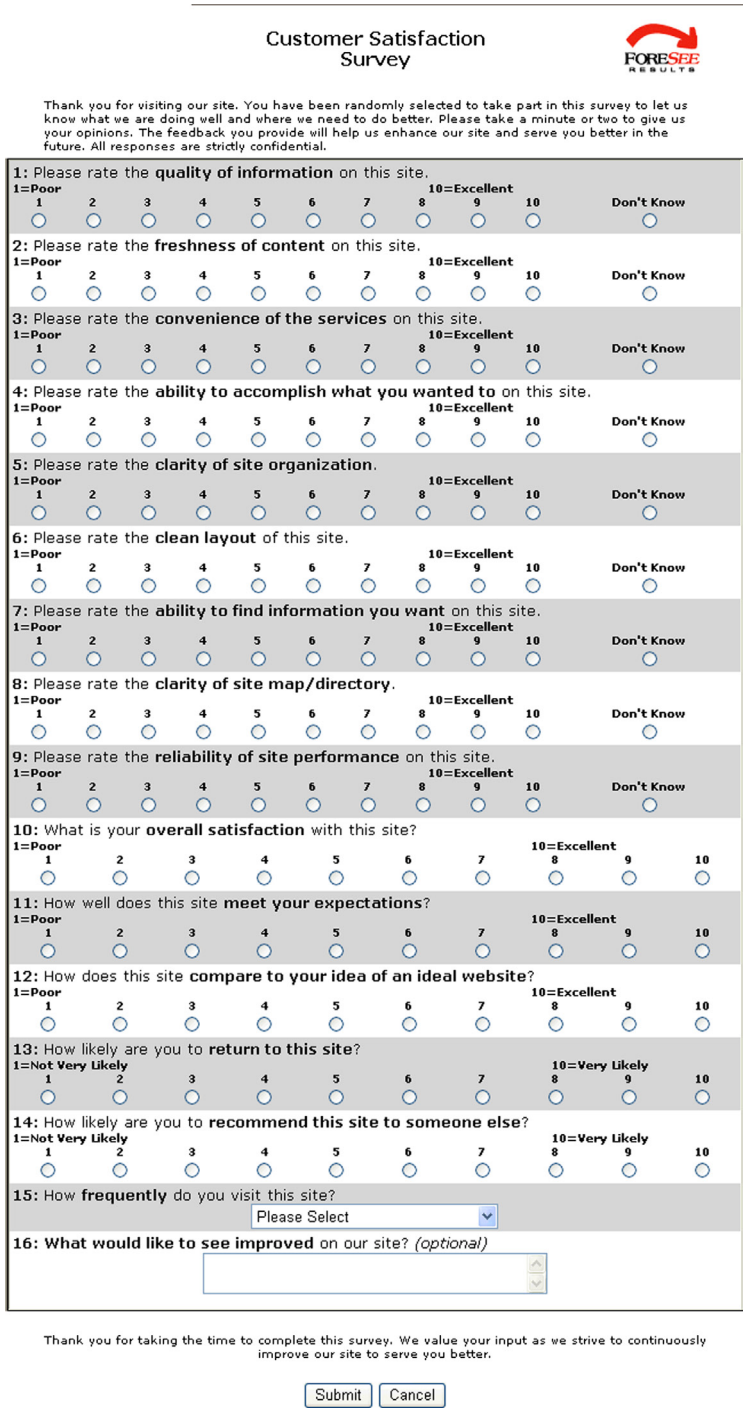
As shown in Figure 6.17, the ACSI results for a website are divided into six quality categories: Content, Functionality, Look & Feel, Navigation, Search, and Site Performance, plus an overall satisfaction score. In addition, they provide average ratings for two "Future Behavior" scores: Likelihood to Return and Recommend to Others. All of the scores are a 100-point scale.

Finally, they also make assessments of the impact that each of the quality scores has on overall satisfaction. This allows you to view the results in four quadrants, as shown in Figure 6.18, plotting the quality scores on the vertical axis and the impact on overall satisfaction on the horizontal axis. Scores in the lower right quadrant (high impact, low score) indicate the areas where you should focus your improvements.

6.6.3 OpinionLab

A somewhat different approach is taken by OpinionLab (www.OpinionLab.com), which provides for page-level feedback from users. In some ways, this can be thought of as a page-level analog of the task-level feedback discussed earlier. As shown in Figure 6.19, a common way for OpinionLab to allow for this page-level feedback is through a floating icon that always stays at the bottom right corner of the page regardless of the scroll position.

Clicking on that icon then leads to one of the methods shown in Figure 6.20 for capturing the feedback. Their scales use five points that are marked simply as − −,−,+−, +, and ++. OpinionLab provides a variety of techniques for



Customer Satisfaction Survey

FORESEE RESULTS

Thank you for visiting our site. You have been randomly selected to take part in this survey to let us know what we are doing well and where we need to do better. Please take a minute or two to give us your opinions. The feedback you provide will help us enhance our site and serve you better in the future. All responses are strictly confidential.

1: Please rate the **quality of information** on this site.
1=Poor 10=Excellent
1 2 3 4 5 6 7 8 9 10 Don't Know

2: Please rate the **freshness of content** on this site.
1=Poor 10=Excellent
1 2 3 4 5 6 7 8 9 10 Don't Know

3: Please rate the **convenience of the services** on this site.
1=Poor 10=Excellent
1 2 3 4 5 6 7 8 9 10 Don't Know

4: Please rate the **ability to accomplish what you wanted to** on this site.
1=Poor 10=Excellent
1 2 3 4 5 6 7 8 9 10 Don't Know

5: Please rate the **clarity of site organization**.
1=Poor 10=Excellent
1 2 3 4 5 6 7 8 9 10 Don't Know

6: Please rate the **clean layout** of this site.
1=Poor 10=Excellent
1 2 3 4 5 6 7 8 9 10 Don't Know

7: Please rate the **ability to find information you want** on this site.
1=Poor 10=Excellent
1 2 3 4 5 6 7 8 9 10 Don't Know

8: Please rate the **clarity of site map/directory**.
1=Poor 10=Excellent
1 2 3 4 5 6 7 8 9 10 Don't Know

9: Please rate the **reliability of site performance** on this site.
1=Poor 10=Excellent
1 2 3 4 5 6 7 8 9 10 Don't Know

10: What is your **overall satisfaction** with this site?
1=Poor 10=Excellent
1 2 3 4 5 6 7 8 9 10

11: How well does this site **meet your expectations**?
1=Poor 10=Excellent
1 2 3 4 5 6 7 8 9 10

12: How does this site **compare to your idea of an ideal website**?
1=Poor 10=Excellent
1 2 3 4 5 6 7 8 9 10

13: How likely are you to **return to this site**?
1=Not Very Likely 10=Very Likely
1 2 3 4 5 6 7 8 9 10

14: How likely are you to **recommend this site to someone else**?
1=Not Very Likely 10=Very Likely
1 2 3 4 5 6 7 8 9 10

15: How **frequently** do you visit this site?
Please Select

16: What would like to see improved on our site? *(optional)*

Thank you for taking the time to complete this survey. We value your input as we strive to continuously improve our site to serve you better.

Submit Cancel

Figure 6.16 Typical questions in an ACSI survey for a website.

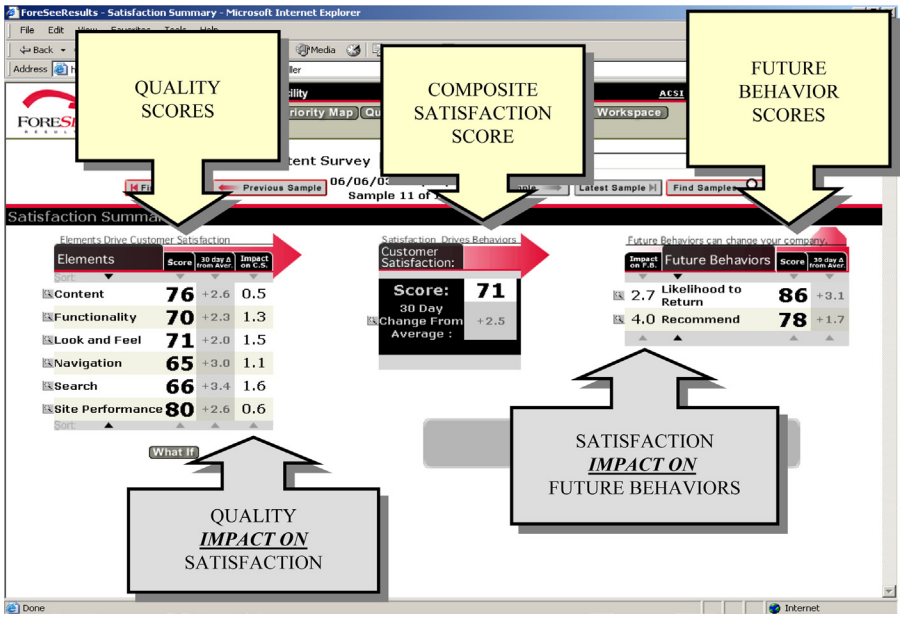


Figure 6.17 Sample results from an ACSI analysis for a website. Scores for six quality areas are shown on the left, along with values estimating the impact that each of those scores has on overall customer satisfaction, which is shown in the center. Scores for two "future behavior" areas are shown on the right, along with values estimating the satisfaction impact on those areas.

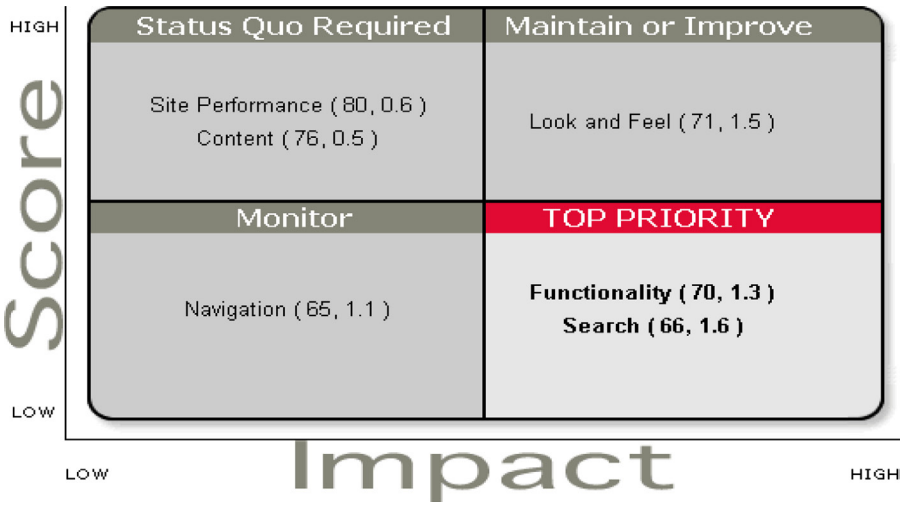


Figure 6.18 Sample results from an ACSI analysis for a website. High and low scores for the six quality areas are represented on the vertical axis, and high and low impact scores are shown on the horizontal axis. The quality areas that fall in the lower right quadrant (Functionality and Search in this example) should be your top priorities for improvement.

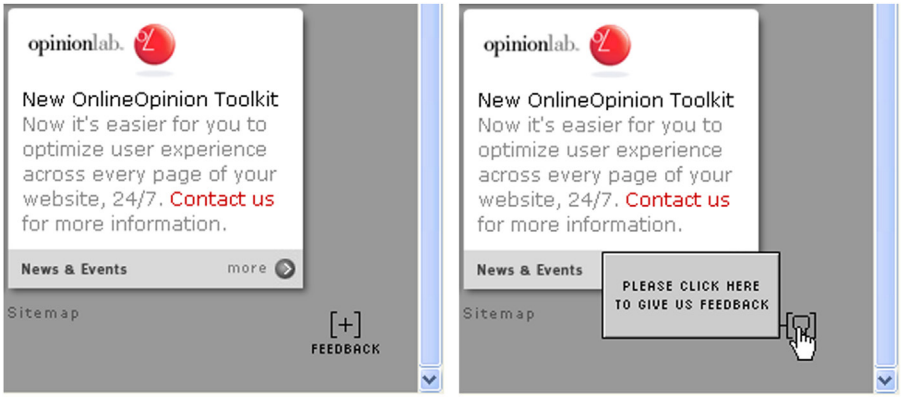


Figure 6.19 An example of a web page containing OpinionLab's feedback mechanism in the lower right corner. This animated icon stays in that position while the user scrolls the page. Moving the mouse over the icon reveals the version shown on the right.

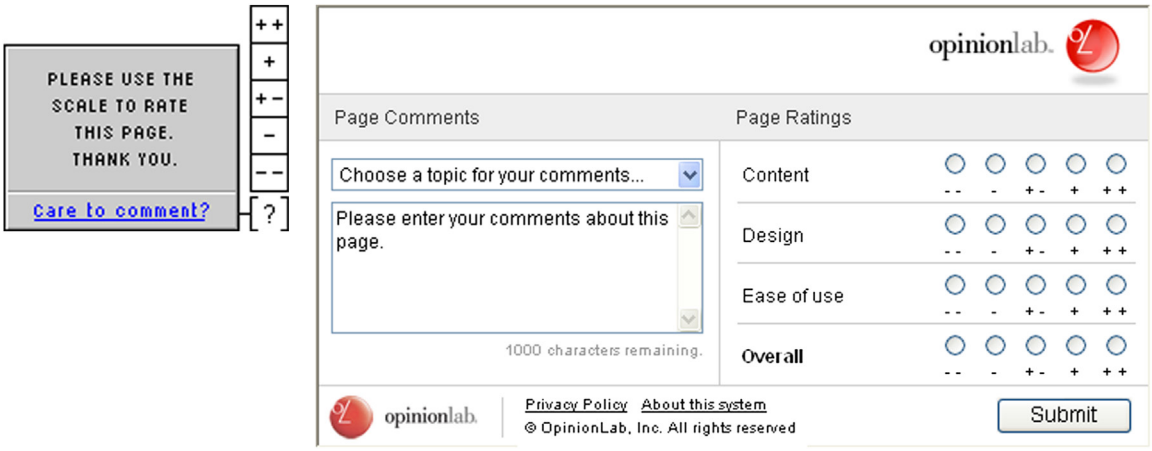


Figure 6.20 Examples of OpinionLab mechanisms for capturing feedback about a web page. The version on the left allows the user to give a quick overall rating of the page. The version on the right allows for more detailed feedback on a few different scales.

visualizing data for a website, including the one shown in Figure 6.21, which allows you to easily spot pages that are getting the most negative feedback and those that are getting the most positive feedback.

6.6.4 Issues with Live-Site Surveys

The following are some of the issues you will need to address when you use live-site surveys.

- *Number of questions*. The fewer questions you have, the higher your response rate is likely to be. That's one reason that companies like

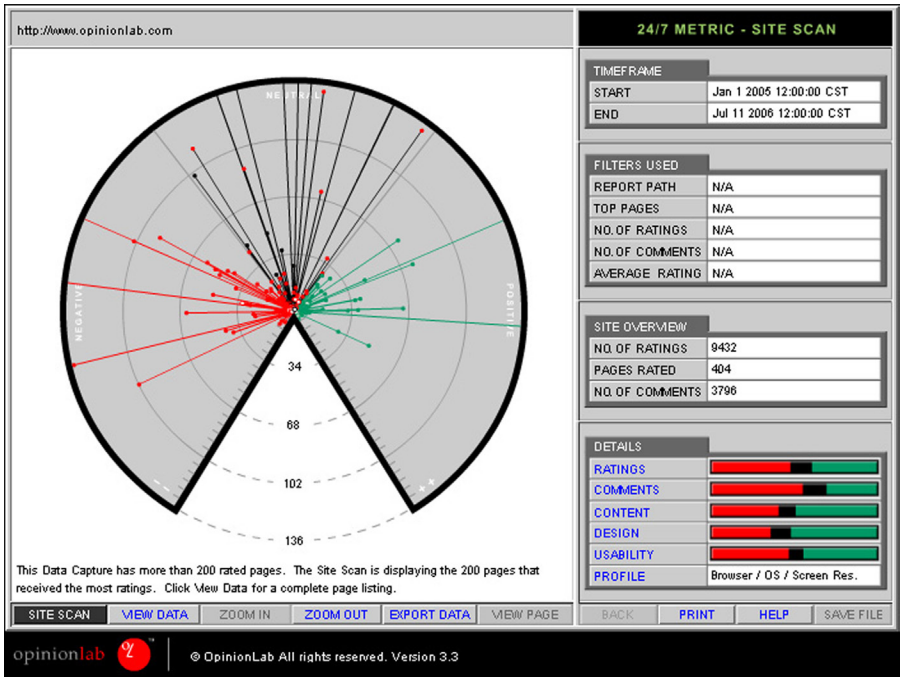


Figure 6.21 OpinionLab provides a variety of techniques for visualizing data for a website. In the visualization shown here, the most-rated 200 pages are represented graphically on the left. Pages receiving the most negative ratings are on the left, those with neutral ratings at the top, and those with the most positive ratings on the right.

OpinionLab keep the number of questions to a minimum. You need to try to strike a balance between getting the information you need and "scaring off" potential respondents. With every question you consider adding, ask yourself if you absolutely must have the information. Some researchers believe that about 20 is the maximum number of questions you should ask in this type of survey.

- *Self-selection of respondents*. Because respondents make a decision about whether or not to complete the survey, they are self-selecting. You should ask yourself if this biases the responses in any way. Some researchers argue that people who are unhappy with the website are more likely to respond than those who are happy (or at least satisfied). If your main purpose is to uncover areas of the site to improve, that may not be a problem.
- *Number of respondents*. Many of these services work on the basis of a percentage of visitors to offer the survey to. Depending on the amount of traffic your site gets, this percentage could be quite small and still generate a large number of responses. You should monitor responses closely to see if you need to increase or decrease the percentage.

- *Nonduplication of respondents.* Most of these services provide a mechanism for noting (typically via a browser cookie or IP address) when the survey has already been offered to someone. As long as the user doesn't clear their cookies and is using the same computer, the survey won't be presented to them again for a specified time period. This prevents duplicate responses from an individual and also prevents annoying those users who don't want to respond.

6.7 OTHER TYPES OF SELF-REPORTED METRICS

Many of the self-report techniques described so far have sought to assess users' reactions to products or websites as a whole or to tasks performed using them. But depending on the objectives of a usability study, you might want to assess users' reactions to specific *attributes* of the product overall or specific *parts* of the product.

6.7.1 Assessing Specific Attributes

Here are some of the attributes of a product or website that you might be interested in assessing:

- Visual appeal
- Perceived efficiency
- Confidence
- Usefulness
- Enjoyment
- Credibility
- Appropriateness of terminology
- Ease of navigation
- Responsiveness

Covering in detail the ways you might assess all the specific attributes you are interested in is beyond the scope of this book. Instead, we describe a few examples of usability studies that have focused on assessing specific attributes.

Gitte Lindgaard and associates at Carleton University were interested in learning how quickly users form an impression of the visual appeal of a web page (Lindgaard et al., 2006). They flashed images of web pages for either 50 or 500 msec to participants in their study. Each web page was rated on an overall scale of visual appeal and on the following bipolar scales: Interesting/Boring, Good Design/Bad Design, Good Color/Bad Color, Good Layout/Bad Layout, and Imaginative/Unimaginative. They found that the ratings on all five of these scales correlated very strongly with visual appeal ($r^2 = 0.86$ to 0.92). They also found that the results were consistent across the participants at both the 50- and the 500-msec exposure levels, indicating that even at 50 msec (or 1/20th of a second), users can form a consistent impression about the visual appeal of a web page.

Bill Albert and associates at Bentley University (Albert, Gribbons, & Almadas, 2009) extended this research to see if users could form an opinion quickly about

their trust of websites based on very brief exposures to images of web pages. They used 50 screenshots of popular financial and health-care websites. After viewing a page for only 50 msec, participants were asked to give a rating of their trust of the site on a 1 to 9 scale. After a break, they repeated the procedure in a second trial with the same 50 images. They found a significant correlation ($r = 0.81$, $p < 0.001$) between the trust ratings in the two trials.

Several years ago, Tullis conducted an online study of 10 different websites to learn more about what makes a website *engaging*. He defined an engaging website as one that (1) stimulates your interest and curiosity, (2) makes you want to explore the site further, and (3) makes you want to revisit the site. After exploring each site, participants responded to a single rating worded as "This website is: Not At All Engaging … Highly Engaging" using a five-point scale. The two sites that received the highest ratings on this scale are shown in Figure 6.22.

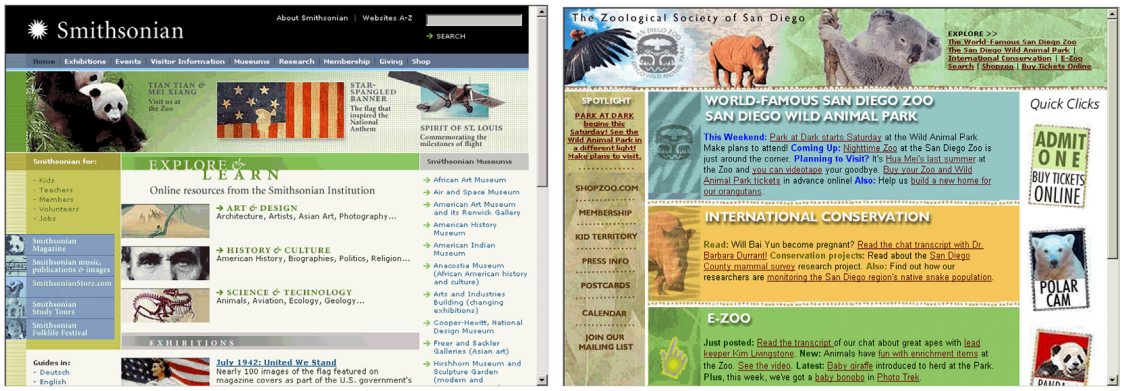


Figure 6.22 Websites rated as the most engaging of 10 sites studied.

One of the techniques often used in analyzing data from subjective rating scales is to focus on the responses that fall onto the extremes of the scale: the top one or two or bottom one or two values. As mentioned earlier, these are often referred to as "Top Box" or "Bottom Box" scores. We used this technique in an online study assessing users' reactions to various load times for an intranet homepage. We manipulated the load time artificially over a range of 1 to 11 seconds. Different load times were presented in a random order, and users were never told what the load time was. After experiencing each load time, users were asked to rate that load time on a five-point scale of "Completely Unacceptable" to "Completely Acceptable." In analyzing the data, we focused on the "Unacceptable" ratings (1 or 2) and the "Acceptable" ratings (4 or 5). These are plotted in Figure 6.23 as a function of the load time. Looking at the data this way makes it clear that a "crossover" from acceptable to unacceptable happened between 3 and 5 seconds.

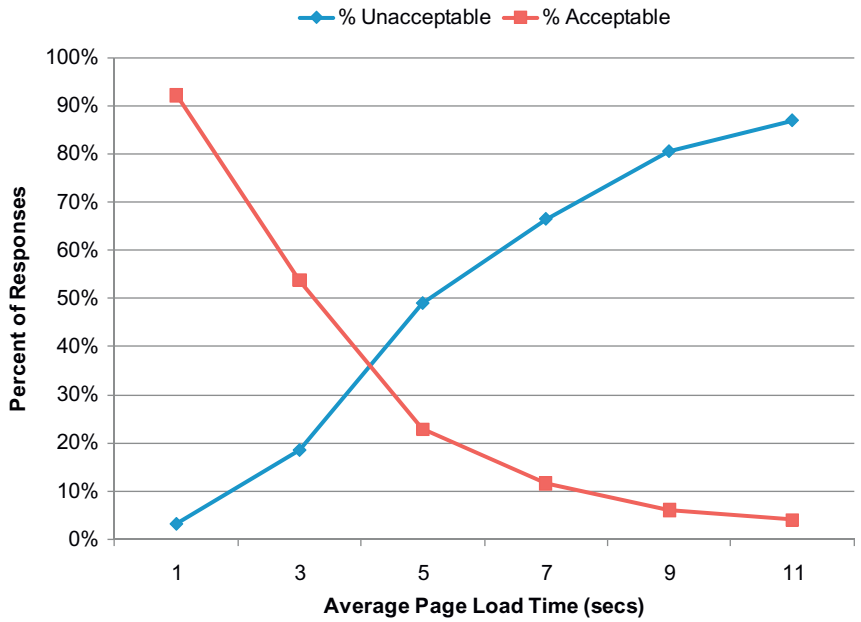


Figure 6.23 Data in which users rated the acceptability of various load times for an intranet homepage presented in a random order. Ratings were on a five-point scale, and data shown here are for the bottom two (Unacceptable) and top two (Acceptable) values only.

B. J. Fogg and associates at the Stanford Persuasive Technology Lab conducted a series of studies to learn more about what makes a website *credible* (Fogg et al., 2001). For example, they used a 51-item questionnaire to assess how believable a website is. Each item was a statement about some aspect of the site, such as "This site makes it hard to distinguish ads from content," and an associated seven-point scale from "Much less believable" to "Much more believable," on which users rated the impact of that aspect on how believable the site is. They found that data from the 51 items fell into seven scales, which they labeled as Real-World Feel, Ease of Use, Expertise, Trustworthiness, Tailoring, Commercial Implications, and Amateurism. For example, one of the 51 items that weighted strongly in the "Real-World Feel" scale was "The site lists the organization's physical address."

6.7.2 Assessing Specific Elements

In addition to assessing specific *aspects* of a product or website, you might be interested in assessing specific *elements* of it, such as instructions, FAQs, or online help; the homepage; the search function; or the site map. The techniques for assessing subjective reactions to specific elements are basically the same as for assessing specific aspects. You simply ask the user to focus on the specific element and then present some appropriate rating scales.

The Nielsen Norman Group (Stover, Coyne, & Nielsen, 2002) conducted a study that focused specifically on the site maps of 10 different websites. After

interacting with a site, users completed a questionnaire that included six statements related to the site map:

- The site map is easy to find
- The information on the site map is helpful
- The site map is easy to use
- The site map made it easy to find the information I was looking for
- The site map made it easy to understand the structure of the website
- The site map made it clear what content is available on the website

Each statement was accompanied by a seven-point Likert scale of "Strongly Disagree" to "Strongly Agree." They then averaged the ratings from the six scales to get an overall rating of the site map for each of the 10 sites. This is an example of getting more reliable ratings of a feature of a website by asking for several different ratings of the feature and then averaging them together.

Tullis (1998) conducted a study that focused on possible homepage designs for a website. (In fact, the designs were really just templates containing "placeholder" or "Lorem Ipsum" text.) One of the techniques used for comparing the designs was to ask participants in the study to rate the designs on three rating scales: page format, attractiveness, and use of color. Each was rated on a five-point scale (−2, −1, 0, 1, 2) of "Poor" to "Excellent." (Note to self and others: Don't use that scale again. It tends to bias respondents away from the ratings associated with the negative values and zero. But the results are still valid if the main thing we're interested in is the relative *comparison* of the ratings for the different designs.) Results for the five designs are shown in Figure 6.24. The

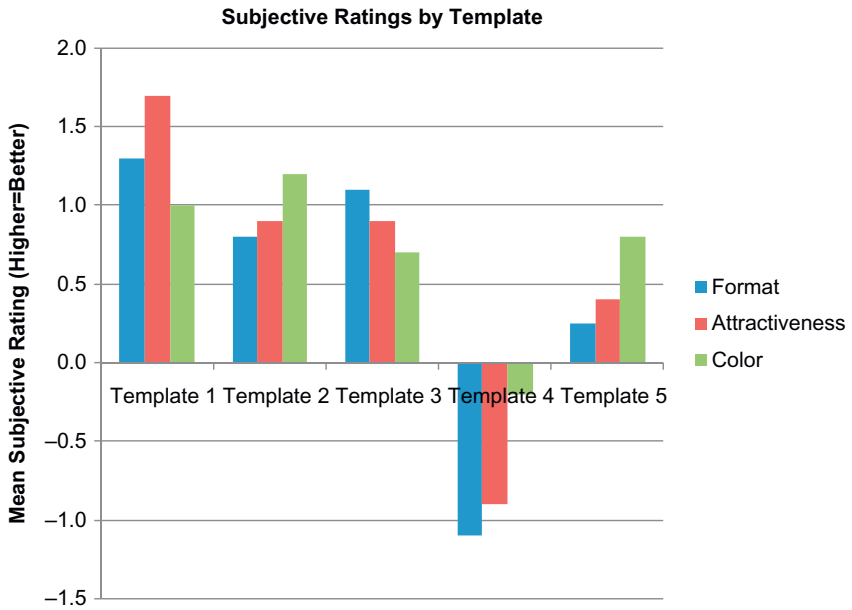


Figure 6.24 Data in which five different designs for a website's homepage were each rated on three scales: format, attractiveness, and use of color. Adapted from Tullis (1998); used with permission.

design that received the best ratings was Template 1, and the design that received the worst ratings was Template 4. This study also illustrates another common technique in studies that involve a comparison of alternatives. Participants were asked to rank-order the five templates from their most preferred to least preferred. In this study, 48% of the participants ranked Template 1 as their first choice, while 57% ranked Template 4 as their last choice.

6.7.3 Open-Ended Questions

Most questionnaires in usability studies include some open-ended questions in addition to the various kinds of rating scales that we've discussed in this chapter. In fact, one common technique is to allow the user to add comments related to any of the individual rating scales. Although the utility of these comments to the calculation of specific metrics may be limited, they can be very helpful in identifying ways to improve the product.

Another flavor of open-ended question used commonly in usability studies is to ask the users to list three to five things they like the *most* about the product and three to five things they like the *least*. These can be translated into metrics by counting the number of instances of essentially the same thing being listed and then reporting those frequencies. Of course, you could also treat the remarks that participants offer while thinking aloud as these kinds of verbatim comments.

Entire books have been written about analyzing these kinds of verbatim responses using what's generally called text mining (e.g., Miner et al., 2012), and a wide variety of tools are available in this space (e.g., Attensity, Autonomy, Clarabridge, to name a few). We will just describe a few simple techniques for collecting and summarizing these kinds of verbatim comments.

Summarizing responses from open-ended questions is always a challenge. We've never come up with a magic solution to doing this quickly and easily. One thing that helps is to be relatively specific in your open-ended questions. For example, a question that asks participants to describe anything they found confusing about the interface is going to be easier to analyze than a general "comments" field.

Figure 6.25 Word cloud created with Wordle.net of responses in an online study of the NASA website about the Apollo Space Program to a question asking for anything they found particularly frustrating or challenging about the site.

One very simple analysis method that we like is to copy all of the verbatim comments in response to a question into a tool for creating word clouds, such as Wordle.net. For example, Figure 6.25 shows a word cloud of responses to a question asking participants to describe anything they found particularly

challenging or frustrating about using the NASA website about the Apollo Space Program (Tullis, 2008b). In a word cloud, larger text is used to represent words that appear more frequently. It's apparent from this word cloud that participants were commenting frequently on the "search" on the site and the "navigation." (Some frequent words, such as "Apollo," are certainly not surprising given the subject matter.)

EXCEL TIP

Finding All Comments That Include a Specific Word

After studying a word cloud (and the accompanying word frequencies that most of these tools can generate), it's sometimes helpful to find all of the verbatim comments that included specific words anywhere in the comment. For example, after seeing the word cloud in Figure 6.25, it might be helpful to find all the comments that included the word "navigation." This can be done in Excel using the =SEARCH function. You can then sort on the column containing the results of the SEARCH function. Entries containing the target word will have numeric values (actually the character position where the target word starts) and those that don't contain the target word will give a "#VALUE!" error.

6.7.4 Awareness and Comprehension

A technique that somewhat blurs the distinction between self-reported data and performance data involves asking the users some questions about what they saw or remember from interacting with the application or website after they have performed some tasks with it and not being allowed to refer back to it. One flavor of this is a check for awareness of various features of a website. For example, consider the NASA homepage shown in Figure 6.26. First, the user would be given a chance to explore the site a little and complete a few very general tasks, such as reading the latest news from NASA and finding how to get images from the Hubble Space Telescope. Then, with the site no longer available to the user, a questionnaire is given that lists a variety of specific pieces of content that the site may or may not have had.

These would generally be content *not* related directly to the specific tasks that the user was asked to perform. You're interested in whether some of these other pieces of content "stood out" to the user. The user then indicates which of the pieces of content on the questionnaire he or she remembers seeing on the site. For example, two of the items on the questionnaire might be "When the ISS Crew is Due to Return" and "Satellite observation of the Western wildfires," both of which are links on the homepage. One of the challenges in designing such a questionnaire is that it must include logical "distracter" items as well—items that were not on the website (or page, if you limit the study to one page) but that look like they could have been.

A closely related technique involves testing for users' learning and comprehension related to some of the content of the website. After interacting with a

Figure 6.26 This NASA homepage illustrates one technique for assessing how "attention-grabbing" various elements of a web page are. After letting users interact with the site, you ask them to identify from a list of content items which ones were actually on the site.

site, users are given a quiz to test their comprehension of some of the information on the site. If the information is something that some of the participants might have already known prior to using the site, it would be necessary to administer a pretest to determine what they already know and then compare their results from the post-test back to that. When the users are not overtly directed to the information during their interaction with the site, this is usually called an "incidental learning" technique.

6.7.5 Awareness and Usefulness Gaps

One type of analysis that can be very valuable is to look at the difference between users' *awareness* of a specific piece of information or functionality and their perceived *usefulness* of that same piece of information or functionality once they are made aware of it. For example, if a vast majority of users are unaware of some specific functionality, but once they notice it they find it very useful, that suggests you should promote or highlight that functionality in some way.

To analyze awareness–usefulness gaps, you must have both an awareness and usefulness metric. We typically ask users about awareness as a yes/no question, for example, "Were you aware of this functionality prior to this study (yes or no)?" Then we ask, "On a 1 to 5 scale, how useful is this functionality to you (1 = Not at all useful; 5 = Very useful)?" This assumes that they have had a couple of minutes to explore the functionality. Next, you will need to convert the rating-scale data into a top-2-box score so that you have an apples-to-apples comparison. Simply plot the percentage of users who are aware of the functionality next to the percentage of users who found the functionality useful (percent top-2 box). The difference between the two bars is called the *awareness–usefulness gap* (see Figure 6.27).

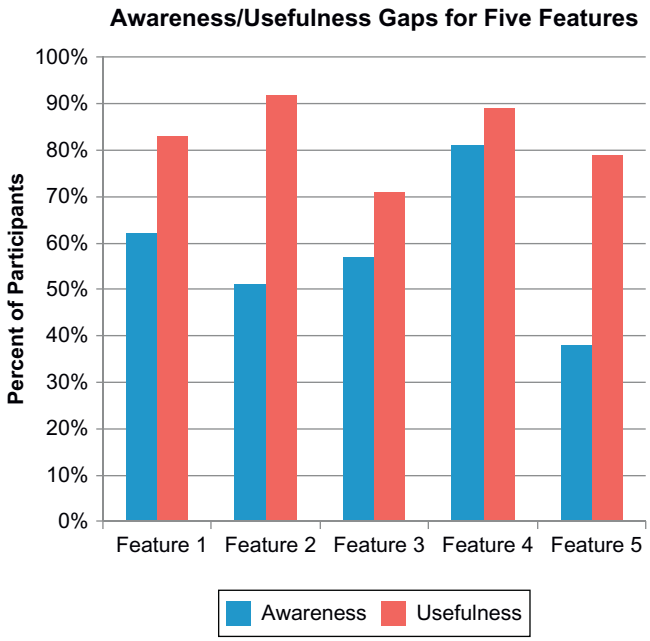


Figure 6.27 Data from a study looking at awareness–usefulness gaps. Items with the greatest difference between awareness and usefulness ratings, such as Tasks 2 and 5, are those you should consider making more obvious in the interface.

6.8 SUMMARY

Many different techniques are available for getting UX metrics from self-reported data. Here's a summary of some of the key points to remember.

1. Consider getting self-reported data at both a task level and at the end of a session. Task-level data can help you identify areas that need improvement. Session-level data can help you get a sense of overall usability.
2. When testing in a lab, consider using one of the standard questionnaires for assessing subjective reactions to a system. The SUS has been shown to be robust even with relatively small numbers of participants (e.g., 8–10).
3. When testing a live website, consider using one of the online services such as WAMMI or ACSI. The major advantage they provide is the ability to show you how the results for your site compare to a large number of sites in their reference database.
4. Be creative but also cautious in the use of other techniques in addition to simple rating scales. When possible, ask for ratings on a given topic in several different ways and average the results to get more consistent data. Carefully construct any new rating scales. Make appropriate use of open-ended questions and consider techniques such as checking for awareness or comprehension after interacting with the product.

This page intentionally left blank

CHAPTER 7

Behavioral and Physiological Metrics

CONTENTS

During a usability study, most participants do much more than complete tasks, respond to questions, and fill out questionnaires. They may laugh, groan, smirk, grimace, smile, fidget in their chair, look aimlessly around the room, or drum their fingers on the table. They feel a wide range of emotions such as stress, excitement, frustration, and surprise. Certain elements of the product grab their attention, while others are completely ignored. Many of these behaviors and emotions are measurable and offer valuable insights into the user experience of the product being tested. This chapter discusses metrics related to unprompted verbal expressions, eye tracking, emotional engagement, and stress.

7.1 OBSERVING AND CODING UNPROMPTED VERBAL EXPRESSIONS

Unprompted verbal expressions provide valuable insight into a participant's emotional and mental state while they are using a product. The participant will

Measuring the User Experience. DOI: http://dx.doi.org/10.1016/B978-0-12-415781-1.00007-8

probably make many comments without being asked, some negative ("This is hard" or "I don't like this design") and some positive ("Wow, this is much easier than I expected" or "I really like the way this looks"). Some comments are neutral or just hard to interpret, such as "This is interesting" or "This is not what I expected."

The most meaningful metric related to verbal expressions is the ratio of positive to negative comments. To do this type of analysis, you first need to catalog all verbal expressions or comments and then categorize each one as positive, negative, or neutral. Once this is complete, simply look at the ratio of positive to negative comments, as illustrated in Figure 7.1. Only knowing that positive comments outnumbered negative comments by a 2:1 ratio does not say a lot by itself. However, it's much more meaningful if the ratios are compared across different design iterations or between different products. For example, if the ratio of positive to negative comments has increased significantly with each new design iteration, this would be one indication of an improved design. Also, if a participant is interacting with more than one design, the same ratio can be calculated for each individual participant, assuming of course that the time spent with each product is the same.

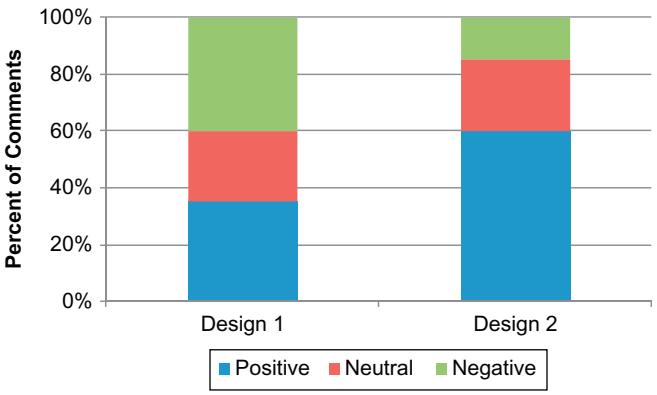


Figure 7.1 Example of coding the percentage of positive, neutral, and negative comments for two different designs

It's also possible to get more granular by differentiating among different types of unprompted verbal comments, such as the following:

- Strongly positive comments (e.g., "This is terrific!")
- Other positive comments (e.g., "That was pretty good.")
- Strongly negative comments (e.g., "This website is terrible!")
- Other negative comments (e.g., "I don't much like the way that worked.")
- Suggestions for improvement (e.g., "It would have been better if…")
- Questions (e.g., "How does this work?")
- Variation from expectation (e.g., "This isn't what I was expecting to get.")
- Stated confusion or lack of understanding (e.g., "This page doesn't make any sense.")
- Stated frustration (e.g., "At this point I'd just leave the website!")

These types of data are analyzed by examining the frequency of comments within each category. Like the previous example, comparing across design iterations or products is the most useful. Categorizing verbal comments beyond just the positive, negative, or neutral can be challenging. It's helpful to work

with another UX researcher to reach some level of agreement about categorizing each comment. Make good use of video recording. Even the best note takers can miss something important. Also, we recommend that these comments be viewed within a larger context. For example, if a participant said that they would never use the product, under any circumstance, yet say something positive about the colors, this needs to be accounted for in other metrics, as well as how the findings are presented. While these metrics are seldom collected because it is fairly time-consuming, they can offer valuable insight into the underlying feelings about a particular design.

7.2 EYE TRACKING

Eye tracking in user research has become more common over the past few years. This is in part due to the ease of use of the systems, particularly around analysis, accuracy, and mobile technology (in the form of goggles), as well as new webcam-based technology.

7.2.1 How Eye Tracking Works

Although a few different technologies are used, many eye-tracking systems, such as the one shown in Figure 7.2, use some combination of an infrared video camera and infrared light sources to track where the participant is looking. The

Figure 7.2 An eye-tracking system from SMI (www.smivision.com). Infrared light sources and an infrared video camera are directly below the monitor. The system tracks the participant's eyes automatically in real time.

infrared light sources create reflections on the surface of the participant's eye (called the corneal reflection), and the system compares the location of that reflection to the location of the participant's pupil. The location of the corneal reflection relative to the pupil changes as the participant moves his eyes.

The first activity in any eye-tracking study is to calibrate the system by asking the participant to look at a series of known points; then the system can subsequently interpolate where he is looking based on the location of the corneal reflection (see Figure 7.3). Typically the researcher can check the quality of the calibration, usually expressed as degrees that deviate from the *X* and *Y* visual planes. Deviations less than one degree are generally considered to be acceptable, and less than one-half of a degree is very good. It is critical that the calibration is satisfactory; otherwise all the eye movement data should not be recorded or analyzed. Without a good calibration there will be a disconnect between what the participant is *actually* looking at and what you assume he is looking at. Following calibration, the moderator makes sure the eye movement data are being recorded. The biggest issue tends to be participants who move around in their seat. Occasionally the moderator is required to ask the

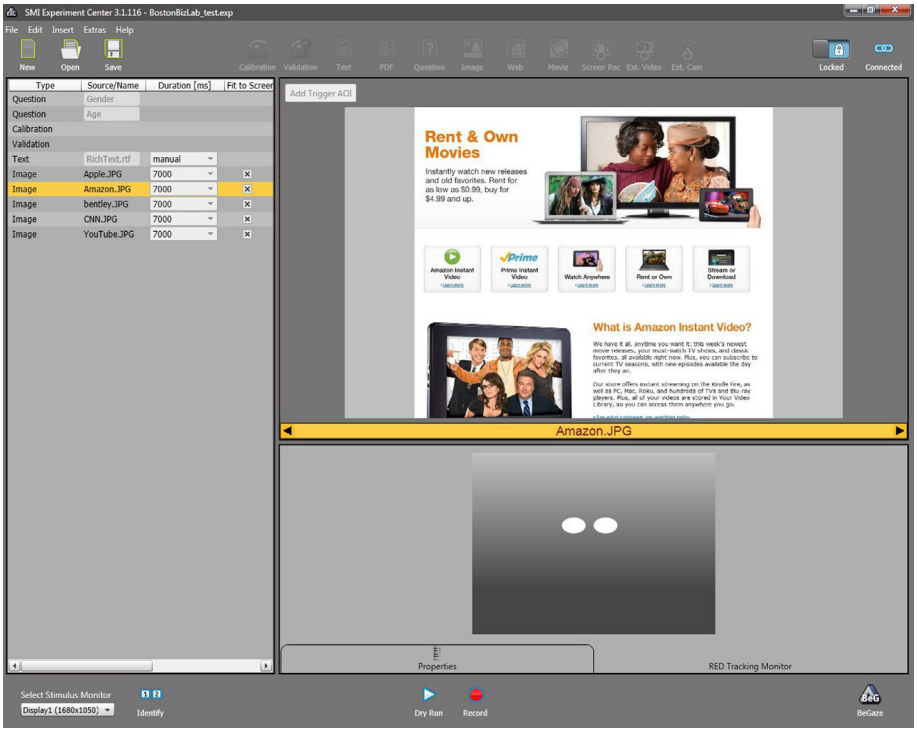


Figure 7.3 An example of SMI software used to run an eye-tracking study and monitor eye movements in real time. The three windows contain study details (left), stimuli being tracked (top right), and eye being tracked (bottom right).

participant to move back/forward, left/right, or raise/lower their seat to recapture the participant's eyes.

Information provided by an eye-tracking system can be remarkably useful in a usability test. Simply enabling observers to see where the participant is looking in real time is extremely valuable. Even if you do no further analyses of eye-tracking data, just this real-time display provides insight that would not be possible otherwise. For example, assume a participant is performing a task on a website and there's a link on the homepage that would take him directly to the page required to complete the task. The participant keeps exploring the website, going down dead ends, returning to the homepage, but never reaching the required page. In a situation like this, you would like to know whether the participant ever saw the appropriate link on the homepage or whether he saw the link but dismissed it as not what he wanted (e.g., because of its wording). Although you could subsequently ask participants that question, their memory may not be completely accurate. With an eye-tracking system you can tell whether the participant at least fixated on the link long enough to read it.

7.2.2 Visualizing Eye-Tracking Data

There are many ways to visualize eye-tracking data. These visualizations tell the story about where people were looking and when. They might be the only thing that your stakeholders really care about. All eye-tracking visualizations are either at an individual level, showing eye movements for one participant, or at an aggregate level, showing eye movements for more than one participant.

WEBCAM-BASED EYE TRACKING

New technology has been developed that allows UX researchers to run eye-tracking studies remotely by taking advantage of the participant's webcam. Webcam-based eye tracking operates under the same premise as more traditional systems use. However, instead of using an infrared signal, a webcam is recognizing the participant's eyes, specifically the movement of the pupil to determine the location on the stimuli the participant is fixating. Vendors such as EyeTrackShop (www.eyetrackshop) provide web-based eye-tracking services, which include setting up the study, storing the data, and providing the analysis and a report. Participants initially agree to allow their webcam to be used for the study and then go through a calibration step prior to running the study. Figure 7.4 is an example screen that the participant would see during the setup process. Similar to any eye-tracking study, different images or visual stimuli are shown to the participants, along with the option to add different survey questions. This technology has the potential to be very useful for UX researchers in that eye movement data can now be collected from a large number of participants, over a short amount of time, without respect to geography. For example, advertisers are now able to test ad effectiveness with a statistically reliable sample size, across many different markets. Data from an EyeTrackShop study clearly show that the "Devil Ad" is clearly more effective with respect to drawing visual attention than the other two ads (see Figure 7.5).

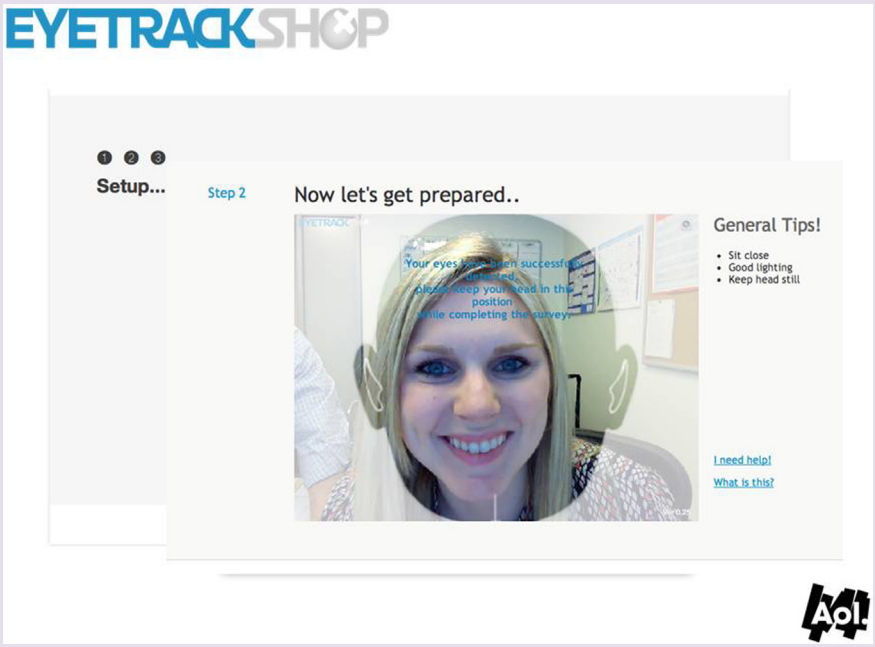


Figure 7.4 Example of the setup procedure using EyeTrackShop. Participants are required to have their face within the profile to ensure proper calibration.

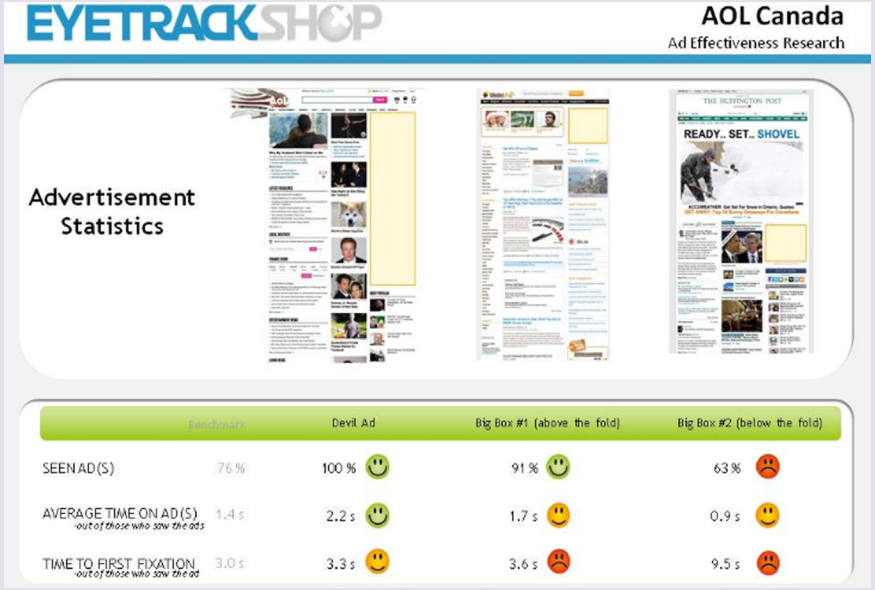


Figure 7.5 Ad effectiveness study using EyeTrackShop.com. The top of the screen shows stimuli (with areas of interest), and the bottom of the screen shows basic statistics, such as percentage who noticed each ad, amount of time spent looking at each ad, and how long it took to first notice each ad. The "devil ad" on the left was most effective.

Figure 7.6 shows the series or sequence of fixations that an individual participant made on the Amazon Video website, also known as a scan path. This is perhaps the most common way to visually represent the eye movements for a single participant. A fixation is defined by a pause in the eye's movement within a well-defined area. Normally these pauses last at least 100 msec (1/10th of a second) or longer. Fixations are usually numbered to indicate their sequence. The size of each circle is proportional to the length or duration of the fixation. The *saccades,* or movements between fixations, are shown by the lines. In Figure 7.6 it is easy to note that the participant was focused primarily on the faces, as well as the first "learn more" box (on the far left). Scan paths are an excellent way to show how a participant looked at the page and what elements they saw in what order.

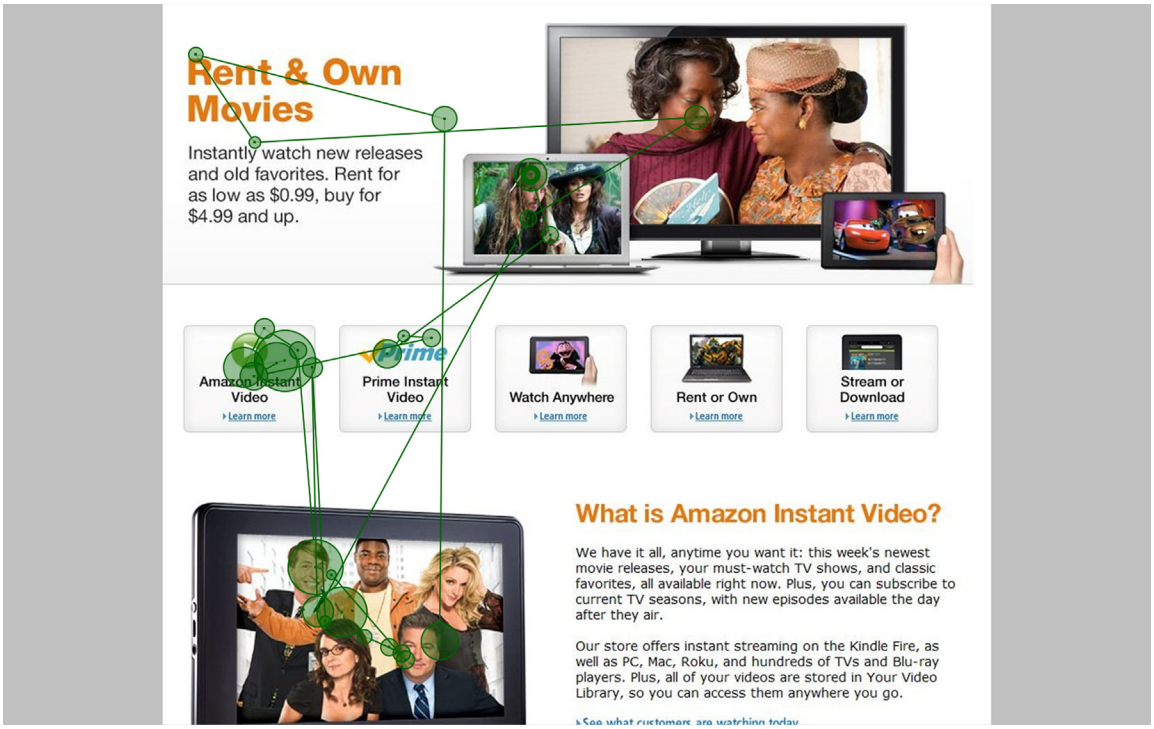


Figure 7.6 Example of one individual's scan path of eye movements on the Amazon Video website.

DID YOU KNOW?

During the saccades, when we're moving our eyes from one point to another, we're essentially blind. This is true whether we're scanning a webpage or reading a book like this one. Of course, we don't perceive it that way. Our brains are constantly integrating the information from the various fixations to give us the perception of a continuous visual stream of information.

By far the most common way to visually represent eye movement for multiple participants is through a heat map (see Figure 7.7). In this visualization, the brightest areas (red) represent a greater density of fixations. It is an excellent way to get a sense of what areas of the page attract more (and less) visual attention. It is important to keep in mind that the analysis software allows the researcher to define the scale of what is considered "red" versus "orange," etc. So, beware that the researcher can easily exaggerate the heat maps to show more or less color. We recommend using the default settings on most software; however, it is important to experiment with using different scales. The opposite visualization is called a focus map, which makes transparent those areas that received more visual attention and darkens those areas that received little or no visual attention. In some sense, a focus map is more intuitive, but a little less common since it is hard to see those areas that are ignored by users.

Figure 7.7 Example of a heat map of the Amazon Video website showing the distribution of eye movements across all participants in the study. The brighter areas as shown in red, orange, and yellow received relatively more visual attention.

7.2.3 Areas of Interest

The most common way to analyze eye-tracking data is by measuring visual attention on specific elements or regions. Most researchers are not just interested in how visual attention is distributed across a web page or scene, but whether participants noticed certain things and how much time was spent looking at them. This is particularly the case in marketing, whereby the success of an ad campaign

is tied directly to getting customers to notice something. Also, it's a concern when there are certain elements critical to task success or having a positive experience. When users don't see them, you can be sure that is a problem.

Figure 7.8 is an example of how to define specific regions on the page. These regions are typically referred to as "look zones" or "areas of interest" (AOIs). AOIs are essentially those things that you want to measure, as defined by a set of *x*,*y* coordinates. When analyzing time spent looking at different regions, keep the following in mind:

- Define each region carefully. Ideally, there will be a small amount of white space in between regions to make sure the eye movements don't get caught in between two AOIs right next to each other.
- Each region should be fairly homogeneous, such as navigation, content, ads, legal information, and so forth. If you prefer to subdivide your AOIs into individual elements, you can always aggregate the data later on.
- When presenting data by AOIs, the question about where participants actually looked within the region typically comes up. Therefore, we recommend including a heat map, as in Figure 7.6, that shows the continuous distribution of fixations.

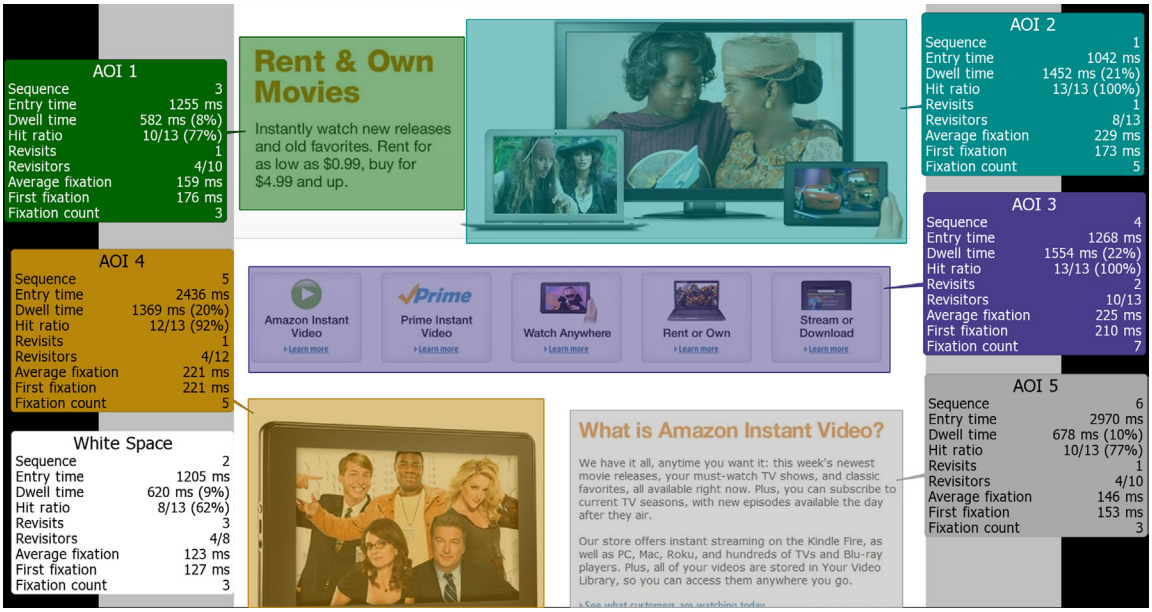


Figure 7.8 Example of the Amazon Movies website with AOIs showing summary statistics for each AOI.

Another useful way to analyze eye movement data by AOIs is through a binning chart (see Figure 7.9). A binning chart shows the percentage of time spent looking at each AOI by some time interval. Keep in mind that the percentages might not add up to 100% unless all the available space is represented within an AOI. Figure 7.9 shows that AOI 1 (green) received more visual attention in the first few

seconds relative to the last few seconds. Conversely, AOI 2 (gray) received more visual attention in the last few seconds compared to the first few seconds. This is a useful way to see the relative prominence of each AOI, not just expressed as a total amount of time. Figure 7.10 is a gridded AOI that shows the amount of visual attention given to equal-sized cells. This is a helpful visualization to see the visual attention across a page, particularly when the elements are not consistent across all pages. For example, the researcher may choose to aggregate data from more than one web page into a single gridded AOI to see generally where users are looking.

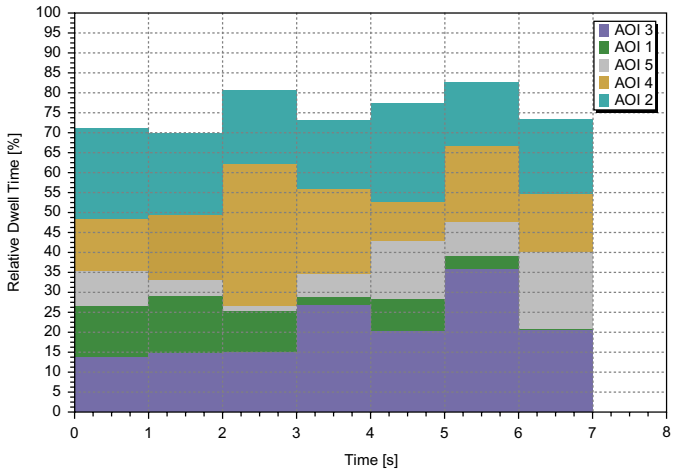


Figure 7.9 Example of a binning chart of the same Amazon Movies website. The binning chart shows the percentage of time spent looking at each AOI during each 1-second interval.

7.2.4 Common Eye-Tracking Metrics

There are many metrics associated with eye-tracking data. The following are some of the most common eye-tracking metrics used by UX researchers. It's important that all of these metrics are

Figure 7.10 Example of a gridded AOI for the Amazon Movies website. The gridded AOI shows the amount of visual attention given to equal-sized cells on the page.

associated with specific AOIs. Figure 7.11 is an example of the type of metrics derived from a single AOI.

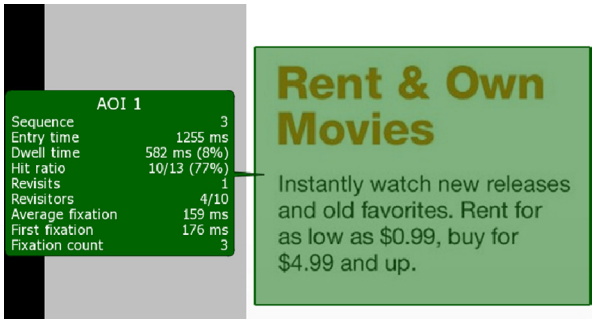


Figure 7.11 Example of common metrics calculated for a single AOI using the SMI software.

DWELL TIME

Dwell time is the total amount of time spent looking within an AOI. This includes all fixations and saccades within the AOI, including revisits. Dwell time is an excellent metric that conveys the level of interest with a certain AOI. Obviously, the greater the dwell time, the greater the level of interest in the AOI. As a general rule of thumb, dwell times less than 100 msec generally mean the participant processed a limited amount of information. A dwell time greater than 500 msec generally means the participant had an opportunity to process the information.

NUMBER OF FIXATIONS

The number of fixations is simply the total count of fixations with an AOI. The number of fixations, as expected, is strongly correlated with dwell time. Because of this, we typically just report dwell time.

FIXATION DURATION

Fixation duration is the average time for fixations. Fixation duration typically ranges from 150 to 300 msec. Fixation duration, similar to number of fixations and dwell time, represents the relative engagement with the object. The greater the average fixation duration, the greater the level of engagement.

SEQUENCE

The sequence represents the order or sequence in which each AOI is first fixated. The sequence tells the researcher the relative prominence of each AOI within the context of a given task. Sometimes it is very helpful to know which AOIs are jumping out to users initially and which AOIs are receiving attention later on. Typically, the sequence is calculated as the average order that each AOI was visited. Keep in mind that many participants may not have experienced that exact same order. Sequence is just a best estimate. We also recommend looking at a binning chart (see Figure 7.8) as another view on sequence of AOIs.

TIME TO FIRST FIXATION

In some situations it's helpful to know how long it takes users to first notice a particular element. For example, you may know that users spend only 7 seconds on average on a page, but you want to make sure that a specific element, such as a "continue" or "sign up" button, is noticed within the first 5 seconds. It's helpful that most eye-tracking systems time stamp each fixation (i.e., the exact time that each fixation occurred).

One way to analyze these data is to take an average of all the times at which the particular element was first fixated. Data should be treated as elapsed time, starting from the initial exposure. The average represents the amount of time taken to first notice the element, for all of those who *did* notice it. Of course, it's possible that some of the participants may not have noticed it all, let alone within the first 5 seconds. Therefore, you may come up with some misleading data showing an artificially quick time by not taking all the participants into account.

REVISITS

Revisits are the number of times that the eye fixates within an AOI, leaves the AOI, and returns back to fixate within the AOI. Revisits indicate the "stickiness" of the AOI. Do the users fixate and leave the AOI, never to return, or do they keep coming back with their eyes?

HIT RATIO

The hit ratio is very simply the percentage of participants who had at least one fixation within the AOI. In other words, this is the number of participants who saw the AOI. In Figure 7.10, 10 out of 13 participants (or 77%) fixated within this particular AOI.

7.2.5 Eye-Tracking Analysis Tips

Over the years we have learned a few things about how to analyze eye-tracking data. Above all else, we strongly recommend you plan your study carefully, as well as taking time to explore the data. It's very easy to draw the wrong

CAN YOU TRUST WHAT PEOPLE SAY THEY SAW IN A USABILITY TEST?

Albert and Tedesco (2010) ran an experiment in which they used eye tracking to test whether usability test participants report what they see accurately. In this study, participants looked at a series of website homepages. After being shown each homepage, the moderators pointed out a specific element. Half of the participants indicated if they had looked at specific elements based on three potential answers (did not look at the element, not sure if they looked at the element, or did look at the element). The other half of the participants used a five-point scale based on how much time was spent looking at that element (from "no time at all" up to "a lot of time"). Results showed that, in general, the eye movements were consistent with what the participants reported seeing. However, in about 10% of the cases, the participant claimed to have "definitely seen" an element, which the eye-movement data showed they did not fixate. In the second group of participants, about 5% of the cases the participants said they "spent a long time looking at an element," yet did not have any eye fixations on that element. Together, these results suggest that participants self-reporting what they looked at during a usability test are reasonably reliable but certainly not perfect.

conclusion based on a few heat maps. Here are a few other important tips to keep in mind as you dive into the data.

- Control the amount of exposure time for each participant. If they did not see the same image or stimuli for the same time, predefine the time to only include the first 10 or 15 seconds, or whatever duration makes the most sense given the context.
- If you are not able to control for exposure time, analyze the dwell time as a percentage, not as an absolute. If someone spent 10 seconds and the other person spent 1 minute, their eye movements will be different, as well as the actual amount of time spent looking at each element.
- Only look at time data when the participant is engaged with the task. Do not include any time data when the participant is debriefing about her experience and still being tracked.
- During the study, make sure that the participants are being tracked. Monitor their eye movements in real time. As soon as they start to slouch or turn their head, remind them gently to maintain their original position.
- Be careful when analyzing eye movements on dynamic websites. Websites that change considerably due to ads, flash, frames, and so on confuse most eye-tracking systems. Every new image is essentially treated as separate stimuli. We strong recommend that you consolidate as many web pages together as possible, knowing that not every page is exactly identical. Otherwise, you will end up with way too many web pages that were only viewed by a single participant. An alternative to this is to simply use static images. They are much easier to analyze, but lack an interactive experience.
- Consider using a trigger AOI to control where participants are initially looking at the start of the experiment. A trigger might say "look here to start the experiment." The text might be in the middle part of the page. After the participant has fixated on the text for a certain number of seconds, the experiment begins. This means that all participants start looking from the same location. This might be overkill for the typical usability test, but should be considered for more tightly controlled eye-tracking studies.

7.2.6 Pupillary Response

Closely related to the use of eye tracking in usability studies is the use of information about the response of the pupil. Most eye-tracking systems must detect the location of the participant's pupil and calculate its diameter to determine where he or she is looking. Consequently, information about pupil diameter is included in most eye-tracking systems. The study of pupillary response, or the contractions and dilations of the pupil, is called pupillometry. Most people know that the pupil contracts and dilates in response to the level of ambient light, but many people don't know that it also responds to cognitive processing, arousal, and increased interest. Typically the greater the level of arousal or interest, the larger the pupil size.

Because pupil dilation is correlated with so many different mental and emotional states, it's difficult to say whether pupillary changes indicate successes or failures in everyday usability testing. However, measuring pupil diameter may be useful in certain situations where the focus is on the amount of mental concentration or emotional arousal. For example, if you are interested mainly in eliciting an emotional response to a new graphic on a website, then measuring changes in pupil diameter (from baseline) may be very useful. To do this, simply measure the percentage deviation away from a baseline for each participant and then average those deviations across the participants. Alternatively, you can measure the percentage of participants who experienced dilated pupils (of a certain amount) while attending to a particular graphic or performing a specific function.

7.3 MEASURING EMOTION

Measuring emotion is difficult. Emotions are often fleeting, hidden, and conflicted. Asking a participant about what she is feeling through an interview or survey may not always be effective. Many participants tell us what they think we want to hear or simply have difficulty articulating what they are really feeling. Some are even hesitant or afraid to admit their true feelings to a perfect stranger.

Despite the difficulty in measuring emotions, it is still very important for the UX researcher to understand the emotional state of the participant. The participant's emotional state while experiencing something is almost always a concern. Most UX researchers use a combination of probing questions, as well as interpretation of their facial expressions, and even body language to infer the participant's emotional state. This may be acceptable for some products; however, it does not always suffice. Some products or experiences are relatively much more emotional and have a greater bearing on the overall user experience. Simply think about the range of emotions a participant might experience when calculating how much money he will have when he retires, reading about a health condition he has, or just playing an action game with friends.

There are essentially three different ways to measure emotions. Emotions can be inferred based on facial expressions, by skin conductance, or by use of EEG. This section highlights three different companies that used these three different approaches. All of these products and services are currently available commercially.

7.3.1 Affectiva and the Q-Sensor

Based on an interview with Daniel Bender, product manager, Affectiva (www.affectiva.com).

The Affective Computing Research group at MIT's Media Lab was founded in 1998 by Professor Rosalind Picard Sc.D. in an effort to develop technologies that advance understanding of emotions. The aim of the research group is to restore a proper balance between emotion and cognition in the design of technologies for addressing human needs (http://affect.media.mit.edu/). Picard and coinvestigator

Rana el Kaliouby, Ph.D., cofounded Affectiva in April 2009 to commercialize technologies developed at the MIT research group. The first product to come from Affectiva is called the Q Sensor (see Figure 7.12).

Figure 7.12 Affectiva's Q Sensor, a wearable, wireless biosensor.

The Q Sensor is a device worn on the wrist that measures the electrical conductance of the skin known as electrodermal activity (EDA). EDA increases when you sweat—small increases in moisture are associated with increased sympathetic nervous system activity indicating emotional activation or arousal. Three types of activation can lead to increases in arousal: increases in cognitive load, affective state, and/or physical activity. Emotional states associated with EDA increases include fear, anger, and joy. Arousal increases are also associated with cognitive demands and may be seen when you are engaged in problem-solving activity. Our state of arousal—and hence the conductivity of our skin—is lower when we are in a relaxed state or bored.

Researchers in a number of fields are using the Q Sensor to measure sympathetic nervous system activity objectively. One of the initial use cases for the Q Sensor has been in understanding the emotional state of students on the autism spectrum. Individuals with autism spectrum disorders often present neutral facial expressions, despite feeling threatened, confused, or otherwise emotionally distressed. Researchers working with autistic students are reviewing EDA data captured with the Q Sensor to better understand the triggers for emotional outbursts. Eventually, the technology will make its way into the classroom where it will serve teachers by providing early warning signals that students are becoming stressed without outward displays of duress. This will enable teachers to respond to their students in a timely and appropriate way.

In the area of user experience research, the Q Sensor can be used to help pinpoint moments of excitement, frustration, or increased cognitive load experienced by the participant. The UX researcher establishes a baseline for each participant. Experiences are then compared to their baseline, with particular attention given to the peaks, or places where there was a peak level in arousal.

While it is helpful knowing what may have triggered an increased level of arousal, it does not tell the researcher whether the experience was positive or negative. This is known as valence. Picard recognized the need to measure valence objectively as she brought Affectiva cofounder el Kaliouby to MIT in January 2007. El Kaliouby's research had been focused on measuring facial expressions using computer-vision and machine-learning techniques. This technology matured and was incorporated into Affectiva's second product, the Affdex facial expression recognition system. Affdex is a passive web-based platform that can take streaming video as an input and predict the presence of facial expressions in close to real time. Affdex is being used to measure emotional response to media

in online panels and in usability labs. Affdex facial-expression recognition provides an indication of the type of experience associated with the state of arousal.

Facial expressions are captured through a standard web camera on the participant's computer and time synchronized with data from the Q Sensor. This provides a rich data set, as peaks in arousal can be associated with a positive or negative valence. With Affdex, Affectiva is building the largest database of spontaneously generated facial expressions in the world. This will allow Affectiva to develop more advanced classifiers of different emotions, which will be used to predict increases in sales or brand loyalty. This powerful technology will arm the UX researcher with an additional set of tools to better understand emotional engagement across a wide variety of experiences. Case study 10.5 highlights use of the Q Sensor in the context of using an onscreen and tablet-based textbook.

RELATIONSHIP AMONG TASK PERFORMANCE, SUBJECTIVE RATINGS, AND SKIN CONDUCTANCE

In a study of participants playing a 3D video game (Super Mario 64), Lin, Hu, Omata, and Imamiya (2005) looked at the relationships among task performance, subjective ratings of stress, and skin conductance. Tasks involved playing three different parts of the game as quickly and accurately as possible. Participants played each part (task) for 10 minutes, during which period they could potentially complete the goal (succeed) multiple times. There was a strong correlation between participants' ratings of how stressful each of the tasks was and their normalized skin conductance (change relative to the participant's baseline) during the performance of each task. In addition, participants who had more successes during the performance of each task tended to have lower skin conductance levels, indicating that failure was associated with higher levels of stress (see Figure 7.13).

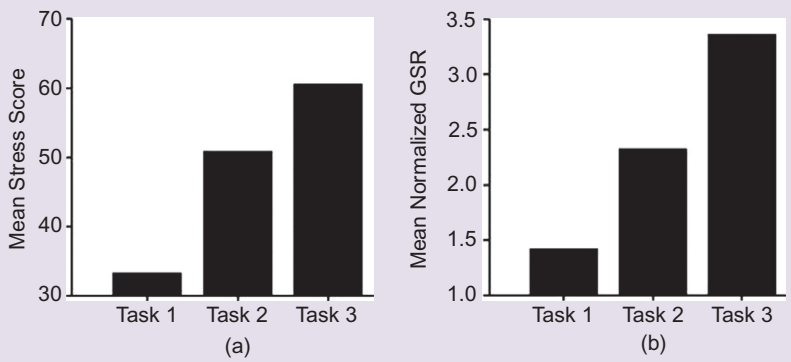


Figure 7.13 Data showing subjective ratings of stress (a) and normalized skin conductance (b) for three different tasks in a video game. Both show that Task 3 was the most stressful, followed by Task 2 and then Task 1. Adapted from Lin et al. (2005).

7.3.2 Blue Bubble Lab and Emovision

Based on an interview with Ben van Dongen, CEO and founder, BlueBubbleLab (www.bluebubblelab.com)

Blue Bubble Lab is a media and technology company based in Palo Alto and Amsterdam that focuses on bringing more relevant messages to consumers based on their emotions and behavior. ThirdSight (www.thirdsight.com), a subsidiary of Blue Bubble Lab, has developed a suite of technology products that bring together computer vision, facial expression analysis, and eye tracking. One product, Emovision, is an application that allows the researcher to understand the participants' emotional state while pinpointing what they are looking at. It is a powerful combination of technologies because the researcher can now draw a direct connection between visual stimuli and an emotional state at any moment in time. This will be invaluable in testing how different visual stimuli produce a range of emotional responses.

Emovision determines the emotional state based on the participants' facial expressions. In the 1970s, Paul Ekman and Wallace Friesen (1975) developed taxonomy for characterizing every conceivable facial expression. They called it the Facial Action Coding System, which included 46 specific actions involving the facial muscles. From his research, Ekman identified six basic emotions: happiness, surprise, sadness, afraid, disgust, and anger. Each of these emotions exhibits a distinct set of facial expressions that can be reliably identified automatically through computer vision algorithms. Emovision uses a webcam to identify the facial expressions at any moment in time and then classifies it into one of seven unique emotions: neutral, happy, surprise, sad, scared, disgusted, and puzzled. At the same time, the webcam is used to detect eye movements.

Figure 7.14 shows how the Emovision application works. On the left side of Figure 7.14 the participant's facial expressions are analyzed. Distinct facial

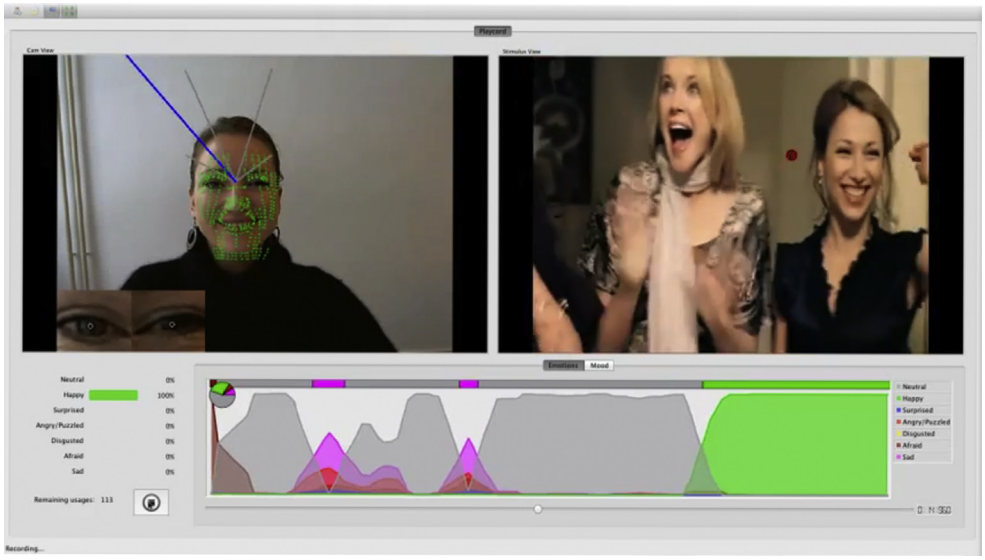


Figure 7.14 Example of EmoVision application that incorporates webcam-based eye tracking and facial expression analysis in real time.

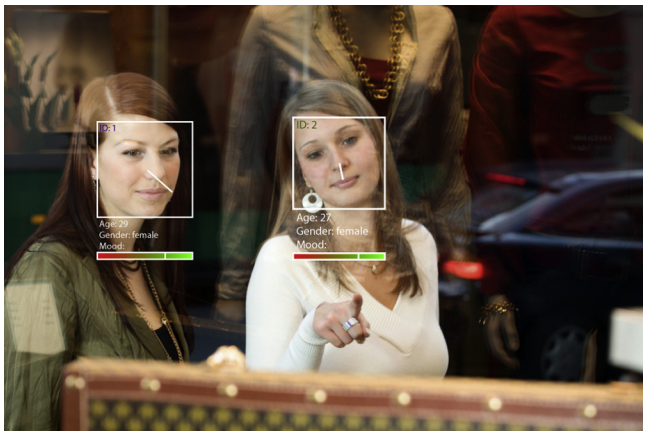


Figure 7.15 Example of how ThirdSight products can be used to deliver tailored messages to consumers based on their mood and other demographics.

muscles are identified and, depending on their shape and movement, an expression is identified. The right side of the application (Figure 7.14) shows the stimulus that is being viewed and the eye movements. In this case the participant is watching a TV commercial, and fixating (as represented by the red dot) in between the two women. The bottom of the screen shows the emotion (in this case it is happy) and assigns a percentage. The line graph depicts the change in emotion over time. When analyzing these data, the researcher can look at any moment in time and identify the associated emotion(s). Also, the researcher can view the overall mood of the experience by seeing the frequency distribution of all the emotions across the experiment. This might be valuable data when comparing different products.

One of the more fascinating applications of this technology is being able to target messages to consumers based on their mood. Figure 7.15 is an example of how this technology can be used to capture facial expressions in the real world, determine the overall mood (positive or negative), as well as demographics such as age and gender, and then deliver a targeted message on a digital billboard or other platform.

7.3.3 Seren and Emotiv

Based on an interview with Sven Krause, key account director, Seren (www.seren.com/)

Seren is a customer experience consultancy based in London. Sven Krause developed a way of measuring a user's emotional engagement and behavior by combining electroencephalography and eye-tracking data. Seren is applying this technology to a wide variety of contexts, including branding, gaming, service, and website design. Researchers at Seren feel this new technology allows them to gain a more complete picture of the user experience as it measures participants' unconscious responses to a stimulus.

Seren uses an EEG device developed by Emotiv (www.emotiv.com). EEG measures brain waves, specifically the amount of electrical activity on different parts of the participant's scalp. Electrical activity is associated with cognitive and emotional states. There is a certain pattern of electrical activity when the participant is in a more excited state relative to a calm state. Also, specific patterns of electrical activity have been associated with other emotional states, such as frustration, boredom, and engagement. EEG technology has been used for many years, for

example, helping diagnose patients with epilepsy, sleep disorders, strokes, and other neurological conditions. Only recently has it been applied within the field of marketing and customer experience.

Seren has worked with SMI (www.smivision.com) to integrate the Emotiv headset with their SMI eye tracker. This allows Seren's researchers to determine what participants are looking at and what triggers their emotional and cognitive state. The integration of both EEG and eye-tracking data is critical, as all data will have a consistent time stamp, allowing the researcher to explore both eye movement and EEG data for a specific event.

Setting up and using their system is fairly straightforward. Participants wear the EEG device on their head, with a series of small conductive pads that contact the scalp and forehead. The EEG device is connected wirelessly to the eye tracker. Baseline measures are taken for a few minutes to allow the participant to get comfortable with the setting. After the researcher feels she has achieved an acceptable baseline, the study begins. Figure 7.16 shows a typical setup. The researcher is monitoring both eye movements and EEG feedback in real time (as shown in Figure 7.17).

Figure 7.16 Typical setup at Seren using EEG technology.

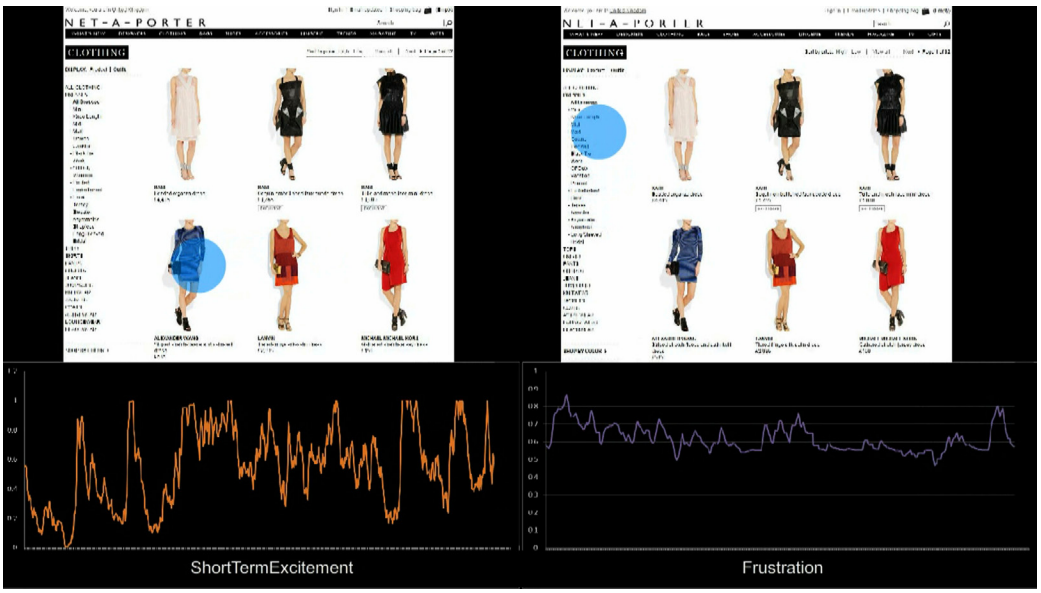


Figure 7.17 An SMI application that allows the researcher to observe EEG feedback and eye movements in real time.

Electroencephalography data are extremely useful in monitoring the emotional engagement of a participant throughout a session. Results can be used to prompt additional questions or to create "emotional heatmaps" that identify areas that led to a change of the emotional state.

7.4 STRESS AND OTHER PHYSIOLOGICAL MEASURES

Stress is unquestionably an important aspect of the user experience. Participants might feel stressed when they have difficulty finding important information or when they are unsure about a transaction they are going through. Measuring stress as part of a typical usability study is rarely done because it is hard to pinpoint the causes of stress. Perhaps the participants are nervous being in a lab environment, are worried about not doing well, or just don't like having their stress levels measured! Because it is hard to associate stress levels to the user experience, these metrics must be approached cautiously. However, they still might be valuable in certain situations.

7.4.1 Heart Rate Variance

One of the most common ways to measure stress is by heart rate, specifically heart rate variability (HRV). HRV measures the time intervals between heart beats. Somewhat counterintuitive, having a certain level of variability in heart rate is healthier than not having any variability at all. Measuring HRV has become much easier in the last few years, thanks primarily to the obsession with fitness and health and, of course, mobile technology. Many runners and other athletes are interested in measuring their heart rates when running. These athletes are likely to be wearing a device on their chest that measures their heart rate and pulse directly. This information can be sent directly to any device. People who aim to reduce stress in their life now can use a handful of smartphone apps to help them measure and monitor their stress levels. One popular app, called Azumio (www.azumio.com), allows users to measure their stress levels using their own smartphone. The user gently places their finger over the camera, and the software is able to detect the heat rate and calculate HRV (see Figure 7.18). HRV is calculated after about 2 minutes, and a stress score is calculated.

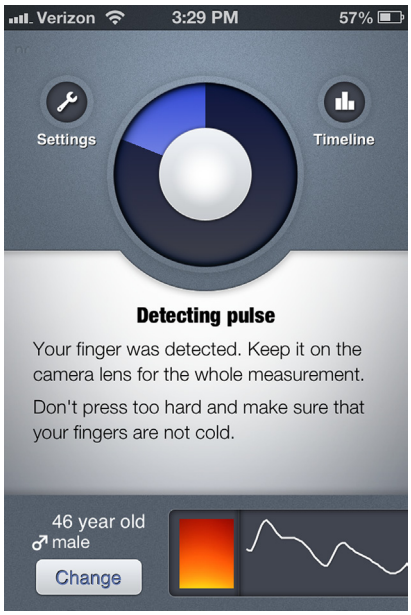


Figure 7.18 Example of Azumio Stress Checker app for the iPhone that measures stress through HRV by detecting heart rate through the camera.

These new apps might be useful for UX research, particularly when evaluating more emotionally based products such as dealing with a person's health or finances. It would be very easy to measure HRV before and after using different designs. It is quite possible that one design resulted in a greater overall HRV rate across participants than other designs. We certainly don't recommend this as a sole

measure of their experience, but it might offer an additional set of data points and potentially insight into causes of stress in their experience.

7.4.2 Heart Rate Variance and Skin Conductance Research

Several studies have sought to determine whether skin conductivity and heart rate could be used as indicators of stress or other adverse reactions in a usability setting. For example, Ward and Marsden (2003) used skin conductance and heart rate to measure user reactions to two versions of a website: a well-designed version and a poorly designed version. The poorly designed version included extensive use of drop-down lists on the homepage to "hide" much of the functionality, provided impoverished navigational cues, used gratuitous animation, and had occasional pop-up windows containing ads. Heart rate and skin conductance data were plotted as changes from the participant's baseline data established during the first minute of the session.

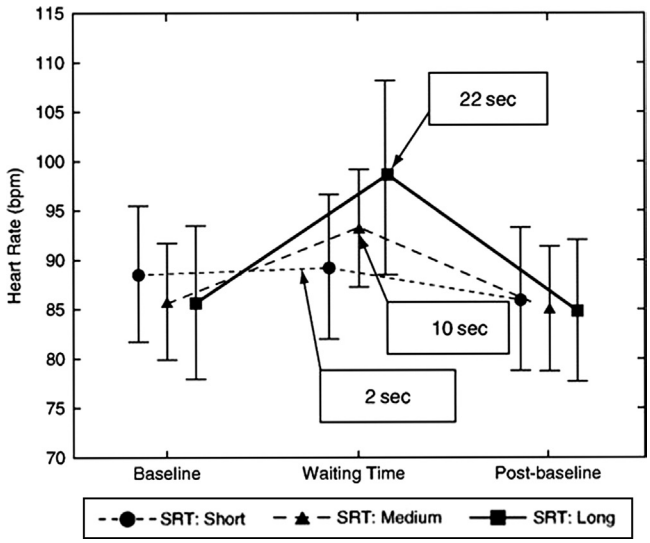


Figure 7.19 Data showing the heart rate of participants as they experienced different levels of response time waiting for web pages to load. Wait times of 10 and 22 seconds yielded progressively greater increases in heart rate relative to baseline, indicating physiological stress. Adapted from Trimmel et al. (2003) used with permission.

Both measures showed a decrease in heart rate and skin conductance for the well-designed website. For the poorly designed site, skin conductance data showed an increase over the first 5 minutes of the session, followed by a return to baseline over the final 5 minutes. Heart rate data for the poorly designed version showed some variability, but the overall trend was to stay at the same level as the baseline, unlike the well-designed version, which showed a decrease relative to baseline. Both measures appear to reflect greater stress in interacting with the poorly designed site.

Trimmel, Meixner-Pendleton, and Haring (2003) measured skin conductance and heart rate to assess the level of stress induced by the response times for web pages to load. They artificially manipulated page load times to be 2, 10, or 22 seconds. They found significant increases in heart rate as response time (page load time) increased, as shown in Figure 7.19. A similar pattern was found for skin conductance. This is evidence of physiological stress associated with longer response times.

7.4.3 Other Measures

A few creative researchers have come up with some other techniques that might be appropriate for assessing the user's level of frustration or engagement while

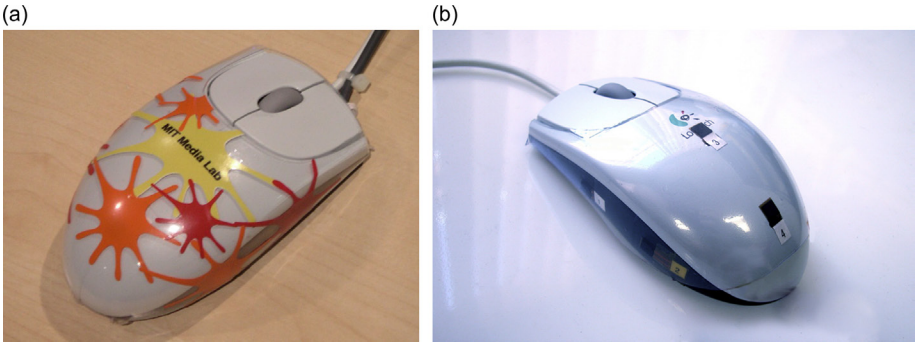


Figure 7.20 The PressureMouse is an experimental mouse that can detect how tightly the user is gripping it. The plastic overlay (a) transmits pressure to six sensors on the top and sides of the mouse (b). As users become frustrated with an interface, many of them subconsciously grip the mouse tighter. The pressure-sensitive mouse was developed by Carson Reynolds and Rosalind Picard of the MIT Media Lab.

interacting with a computer. Most notably, Rosalind Picard and her team in the Affective Computing Research Group at the MIT Media Lab have investigated a variety of techniques for assessing the user's emotional state during human–computer interaction. Two of these techniques that might have application to usability testing are the PressureMouse and the Posture Analysis Seat.

The PressureMouse (Reynolds, 2005), shown in Figure 7.20, is a computer mouse with six pressure sensors that detect how tightly the user is gripping the mouse. Researchers had users of the PressureMouse fill out a five-page web-based survey (Dennerlein et al., 2003). After submitting one of the pages, participants were given an error message indicating that something was wrong with their entries on that page. After acknowledging the error message, participants were then taken back to that page, but all the data they had entered had been deleted and they had to reenter it. As illustrated in Figure 7.21, participants who had been categorized as members of a "high-response" group (based on their negative ratings in a usability questionnaire about the online survey) gripped the mouse significantly tighter for the 15 seconds *after* their loss of data than they did for the 15 seconds *before.*

The Posture Analysis Seat measures the pressure that the user is exerting on the seat and back of the chair. Kapoor, Mota, and Picard (2001) found that they could reliably detect changes in posture on the part of the participant, such as sitting upright, leaning forward, slumping backward, or leaning sideways. These may be used to infer different levels of engagement or interest on the part of the participant. Of course, anyone who has taught can easily see a student's engagement based on how much they slouch in their seat!

These new technologies have yet to be used in everyday usability testing, but they look promising. As these or other technologies for measuring engagement or frustration become both affordable and unobtrusive, they can be used in many situations in which they could provide valuable metrics, such as designing products for children who have limited attention spans, evaluating users'

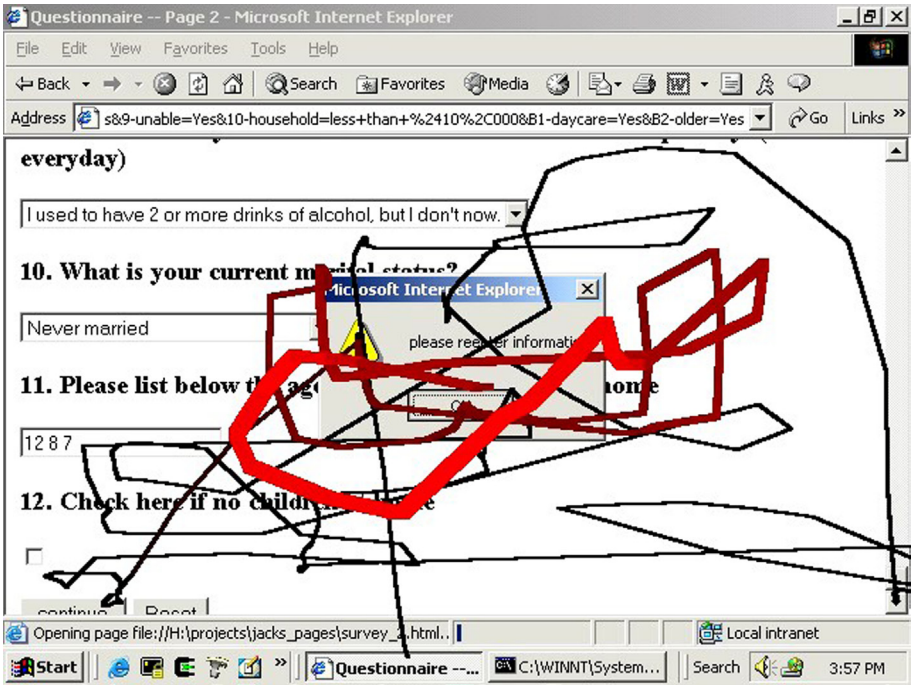


Figure 7.21 In this visualization of data from the PressureMouse, the mouse leaves a "trail" on the screen. The thickness of the trail indicates how tightly the participant is gripping the mouse. In this example, the participant is initially gripping with normal pressure while completing the online survey. When he clicked on the "Continue" button (#1), the pressure was still normal, until he started reading the error message, which caused him to grip the mouse tighter (#2). Finally, after dismissing the dialog box and seeing that the data he had entered was now gone, his grip on the mouse got even tighter (#3). Adapted from Reynolds (2005); used with permission.

patience for download times or error messages, or measuring teenagers' level of engagement with new social networking applications.

7.5 SUMMARY

This chapter covered a variety of ways to measure user behaviors and emotions. This provides potentially valuable insights into the deeper user experience that is often very easy to miss in the course of a usability test. These tools are becoming much easier to use, more accurate, more versatile and powerful, and even quite affordable. Despite these many advances, we strongly recommend taking advantage of other UX metrics and not relying solely on this technology to tell you everything about the user experience. Here's a summary of some of the key points to remember.

1. A structured approach to collecting unprompted verbal expressions during a usability test can be very helpful by tabulating the number of positive and negative comments made by participants during each of the tasks.

2. Eye tracking can be a significant benefit in many kinds of usability tests. The technology continues to improve, becoming more accurate, easier to use, and less intrusive. The value is being able to compare the effectiveness of different designs, as well as calculate metrics based on areas of interest. Key metrics include dwell time, time to first fixation, and hit ratio. There are many ways to visualize results from eye tracking, such as heat maps and gridded AOIs.
3. There are three ways to measure emotions: skin conductance, facial expressions, and EEG. Skin conductance measures level of arousal, facial expressions are classified and associated with six basic emotions, and EEG measures brain wave activity with unique signatures tied to specific emotional responses. There are new technologies based on each approach that even integrate eye movement data into their applications. These are powerful new tools used to gain insight into the emotional response of the user.
4. Stress is an important component of the user experience. It is measured most often as heart rate variance. New apps allow the researcher to calculate HRV very easily. However, there are many factors that impact stress beyond the user experience.
5. Other techniques for capturing information about the participant's behavior, such as a mouse that registers how tightly it is being gripped, are on the horizon and may become useful additions to the battery of tools available for use in usability testing.

CHAPTER 8

Combined and Comparative Metrics

CONTENTS

Usability data are building blocks. Each piece of usability data can be used to create new metrics. Raw usability data can include task completion rates, time on task, or self-reported ease of use. All of these usability data can be used to derive new metrics that were not available previously, such as an overall usability metric or some type of "usability scorecard." Why might you want to do this? We think the most compelling reason is to have an easy-to-understand score or summary of all the metrics you've collected in a study. This can be very handy when presenting to senior managers, for tracking changes across iterations or releases, and for comparing different designs.

Two of the common ways to derive new usability metrics from existing data are by (1) combining more than one metric into a single usability measure and (2) comparing existing usability data to expert or ideal results. Both methods are reviewed in this chapter.

8.1 SINGLE USABILITY SCORES

In many usability tests, you collect more than one metric, such as task completion rate, task time, and perhaps a self-reported metric such as a system usability scale (SUS) score. In most cases, you don't care so much about the results for each these metrics individually as you do about the total picture of the user experience as reflected by *all* of these metrics. This section covers the various

Measuring the User Experience. DOI: http://dx.doi.org/10.1016/B978-0-12-415781-1.00008-X

ways you can combine or represent different metrics to get an overall view of the usability of a product, or different aspects of a product, perhaps as revealed by different tasks.

The most common question asked after a usability test is "How did it do?" People who ask this question (often the product manager, developer, or other members of the project team) usually don't want to hear about task completion rates, task times, or questionnaire scores. They want an overall score of some type: Did it pass or fail? How did it do in comparison to the last round of usability testing? Making these kinds of judgments in a meaningful way involves combining the metrics from a usability test into some type of overall score. The challenge is figuring out how to combine scores from different scales in a meaningful way (e.g., task completion rates in percentages and task times in minutes or seconds).

8.1.1 Combining Metrics Based on Target Goals

Perhaps the easiest way to combine different metrics is to compare each data point to a target goal and represent one single metric based on the percentage of users who achieved a combined set of goals. For example, assume that the goal is for users to complete at least 80% of their tasks successfully in no more than 70 seconds each on the average. Given that goal, consider the data in Table 8.1, which shows task completion rate and average time per task for each of eight participants in a usability test.

Table 8.1 shows some interesting results. The average values for task completion (82%) and task time (67 seconds) would seem to indicate that the goals for this test were met. Even if you look at the number of users who met the task completion goal (six participants, or 75%) or the task time goal (five participants, or 62%), you would still find the results reasonably encouraging. However, the most appropriate way to look at the results is to see if each individual participant

Participant #	Task Completion	Task Time (secs)	Goal Met?
1	85%	68	1
2	70%	59	0
3	80%	79	0
4	75%	62	0
5	90%	72	0
6	80%	60	1
7	80%	56	1
8	95%	78	0
Average:	82%	67	38%

Table 8.1 Sample task completion and task time data from eight participants[a].
[a]Also shown are averages for task completion and time and an indication of whether each participant met the objective of completing at least 80% of the tasks in no more than 70 seconds.

met the stated goal (i.e., the *combination* of completing at least 80% of the tasks in no more than 70 seconds each). It turns out, as shown in the last column of Table 8.1, that only three, or 38%, of the participants actually met the goal. This demonstrates the importance of looking at individual participant data rather than just looking at averages. This can be particularly true when dealing with relatively small numbers of participants.

This method of combining metrics based on target goals can be used with any set of metrics. The only real decision is what target goals to use. Target goals can be based on business goals and/or comparison to ideal performance. The math is easy (each person just gets a 1 or 0), and the interpretation is easy to explain (the percentage of users who had an experience that met the stated goal during the test).

8.1.2 Combining Metrics Based on Percentages

Although we're well aware that we should have measurable target goals for our usability tests, in practice many of us don't have them. So what can you do to combine different metrics when you don't have target goals? One simple technique for combining scores on different scales is to convert each score to a percentage and then average them. For example, consider the data in Table 8.2, which shows results of a usability test with 10 participants.

One way to get an overall sense of the results from this study is to first convert each of these metrics to a percentage. In the case of the number of tasks completed and the subjective rating, it's easy because we know the maximum ("best") possible value for each of those scores: there were 15 tasks, and the maximum

Participant #	Time Per Task (sec)	Tasks Completed (of 15)	Rating (0–4)
1	65	7	2.4
2	50	9	2.6
3	34	13	3.1
4	70	6	1.7
5	28	11	3.2
6	52	9	3.3
7	58	8	2.5
8	60	7	1.4
9	25	9	3.8
10	55	10	3.6

Table 8.2 Sample data from a usability test with 10 participants[a].

[a]Time per task is the average time to complete each task, in seconds. Tasks completed are number of tasks (out of 15) that the user completed successfully. Rating is the average of a five-point task ease rating for each task, where higher is better.

possible subjective rating on the scale was 4. So we just divide the score obtained for each participant by the corresponding maximum to get the percentage.

In the case of time data, it's a little trickier, as there's not a predefined "best" or "worst" time—the ends of the scale are not known beforehand. One way of handling this would be to have several experts do the task and treat the average of their times as the "best" time. Another way is to treat the fastest time obtained from the participants in the study as the "best" (25 seconds, in this example), the slowest time as the "worst" (70 seconds, in this example), and then express other times in relation to those. Specifically, you divide the difference between the longest time and the observed time by the difference between longest and shortest times. This way, the shortest time becomes 100% and the longest becomes 0%. Using that method of transforming the data, you get the percentages shown in Table 8.3.

Participant #	Time	Tasks	Rating	Average
1	11%	47%	60%	39%
2	44%	60%	65%	56%
3	80%	87%	78%	81%
4	0%	40%	43%	28%
5	93%	73%	80%	82%
6	40%	60%	83%	61%
7	27%	53%	63%	48%
8	22%	47%	35%	35%
9	100%	60%	95%	85%
10	33%	67%	90%	63%

Table 8.3 Data from Table 8.2 transformed to percentages[a].

[a]For task completion data, the score was divided by 15. For rating data, the score was divided by 4. For time data, the difference between the longest time (70) and the observed time was divided by the difference between longest (70) and shortest (25) times.

We're grateful to David Juhlin, of the Bentley University Design and Usability Center, for suggesting this transformation of time data. In the first edition of this book we used a different method, which resulted in a nonlinear transformation. This new approach is linear and more appropriate.

TRANSFORMING TIME DATA IN EXCEL

Here are the steps for transforming time data to percentages using these rules in Excel:

1. Enter the raw times into a single column in Excel. For this example, we will assume they are in column "A" and that you started on row "1". Make sure there are no other values in this column, such as an average at the bottom.

2. In the cell to the right of the first time, enter the formula:

   = (MAX(A:A) − A1)/(MAX(A:A) − MIN(A:A))

3. Copy this formula down as many rows as there are times to be transformed.

Table 8.3 also shows the average of these percentages for each of the participants. If any one participant had completed all the tasks successfully in the shortest average time and had given the product a perfect score on the subjective rating scales, that person's average would have been 100%. However, if any one participant had failed to complete any of the tasks, had taken the longest time per task, and had given the product the lowest possible score on the subjective rating scales, that person's average would have been 0%. Of course, rarely do you see either of those extremes. Like the sample data in Table 8.3, most participants fall between those two extremes. In this case, averages range from a low of 28% (Participant 4) to a high of 85% (Participant 9), with an overall average of 58%.

CALCULATING PERCENTAGES ACROSS ITERATIONS OR DESIGNS

One of the valuable uses of this kind of overall score is in making comparisons across iterations or releases of a product or across different designs. But it's important to do the transformation across *all* of the data at once, not separately for each iteration or design. This is particularly important for time data, where the times that you've collected are determining the best and worst times. That selection of the best and worst times should be done by looking across all of the conditions, iterations, or designs that you want to compare.

So if you had to give an "overall score" to the product whose test results are shown in Tables 8.2 and 8.3, you could say it got 58% overall. Most people wouldn't be too happy with 58%. Many years of grades from school have probably conditioned most of us to think of a percentage that low as a "failing grade." But you should also consider how accurate that percentage is. Because it's an average based on individual scores from 10 different participants, you can construct a confidence interval for that average, as explained in Chapter 2. The 90% confidence interval in this case is ±11%, meaning that the confidence interval extends from 47 to 69%. Running more participants would probably give you a more accurate estimate of this value, whereas running fewer would probably have made it less accurate.

One thing to be aware of is that when we averaged the three percentages together (from task completion data, task time data, and subjective ratings), we gave equal weight to each of those measures. In many cases, that's a perfectly reasonable thing to do, but sometimes the business goals of the product

may indicate a different weighting. In this example, we're combining two performance measures (task completion and task time) with one self-reported measure (rating). By giving equal weight to each, we're actually giving twice as much weight to performance as to the self-reported measure. That can be adjusted by using weights in calculating the averages, as shown in Table 8.4.

Participant #	Time	Weight	Tasks	Weight	Rating	Weight	Weighted Average
1	38%	1	47%	1	60%	2	51%
2	50%	1	60%	1	65%	2	60%
3	74%	1	87%	1	78%	2	79%
4	36%	1	40%	1	43%	2	40%
5	89%	1	73%	1	80%	2	81%
6	48%	1	60%	1	83%	2	68%
7	43%	1	53%	1	63%	2	55%
8	42%	1	47%	1	35%	2	40%
9	100%	1	60%	1	95%	2	88%
10	45%	1	67%	1	90%	2	73%

Table 8.4 Calculation of weighted averages[a].
[a]Each individual percentage is multiplied by its associated weight, these products are summed, and that sum is divided by the sum of the weights (4, in this example).

In Table 8.4, the subjective rating is given a weight of 2, and each of the two performance measures is given a weight of 1. The net effect is that the subjective rating gets as much weight in the calculation of the average as the two performance measures together. The result is that these weighted averages for each participant tend to be closer to the subjective ratings than the equal-weight averages in Table 8.3. The exact weights you use for any given product should be determined by the business goals for the product. For example, if you're testing a website for use by the general public, and the users have many other competitors' websites to choose from, you might want to give more weight to self-reported measures because you probably care more about the users' *perception* of the product than anything else.

However, if you're dealing with an application where speed and accuracy are more important, such as a stock-trading application, you would probably want to give more weight to performance measures. You can use any weights that are appropriate for your situation, but remember to divide by the sum of those weights in calculating the weighted average.

These basic principles apply to transforming any set of metrics from a usability test. For example, consider the data in Table 8.5, which includes number of tasks completed successfully (out of 10), number of web page visits, an overall satisfaction rating, and an overall usefulness rating.

Participant #	Tasks Completed (of 10)	# of Page Visits (min = 20)	Satis-faction Rating (0–6)	Useful-ness Rating (0–6)	Tasks	Page Visits	Satis-faction	Useful-ness	Average
1	8	32	4.7	3.9	80%	63%	78%	65%	71%
2	6	41	4.1	3.8	60%	49%	68%	63%	60%
3	7	51	3.4	3.7	70%	39%	57%	62%	57%
4	5	62	2.4	2.3	50%	32%	40%	38%	40%
5	9	31	5.2	4.2	90%	65%	87%	70%	78%
6	5	59	2.7	2.9	50%	34%	45%	48%	44%
7	10	24	5.1	4.8	100%	83%	85%	80%	87%
8	8	37	4.9	4.3	80%	54%	82%	72%	72%
9	7	65	3.1	2.5	70%	31%	52%	42%	49%

Table 8.5 Sample data from a usability test with nine participants[a].

[a]Tasks completed are the number of tasks (out of 10) that the user completed successfully. Number of page visits is the total number of web pages that the user visited in attempting the tasks.(Typically, each revisit to the same page is counted as another visit.) The two ratings are average subjective ratings of satisfaction and usefulness, each on a seven-point scale (0–6)

Calculating percentages from these scores is very similar to the previous example. The number of tasks completed is divided by 10, and the two subjective ratings are each divided by 6 (the maximum rating). The other metric, number of web page visits, is somewhat analogous to the time metric in the previous example. But in the case of web page visits, it is usually possible to calculate the minimum number of page visits that would be required to accomplish the tasks. In this example, it was 20. You can then transform the number of page visits by dividing 20 (the fewest possible) by the actual number of page visits. The closer the number of page visits is to 20, the closer the percentage will be to 100%. Table 8.5 shows original values, percentages, and then equal-weight averages. In this case, note that equal weighting (normal average) results in the same weight being given to performance data (task completion and page visits) and self-reported data (the two ratings).

CONVERTING RATINGS TO PERCENTAGES

What if the subjective ratings you used were on a scale that started at 1 instead of 0? Would that make a difference in how you transform the ratings to a percentage? Most definitely. Let's assume the ratings were on a scale of 1–7 instead of 0–6, with higher numbers being better. Both are seven-point scales. In both cases, you want the lowest possible rating to become 0% and the highest possible rating to become 100%. When the ratings are on a 0–6 scale, simply dividing each rating by 6 (the highest possible rating) gives the desired range (0 to 100%). But when the ratings are on a 1–7 scale,

there's a problem. If you divide each rating by 7 (the highest possible rating), you get a maximum score of 100%, which is okay, but the minimum score is 1/7, or 14%, not the 0% that you want. The solution is to first subtract 1 from each rating (rescaling it to 0–6) and then divide by the new maximum score (6, in this case). So, the lowest score becomes (1−1)/6, or 0%, and the highest becomes (7−1)/6, or 100%.

To look at transforming another set of metrics, consider the data in Table 8.6. In this case, the number of errors is listed, which would include specific errors the users made, such as data-entry errors. Obviously, it is possible (and desirable) for a user to make no errors, so the minimum possible is 0. But there's usually no predefined maximum number of errors that a user could make. In a case like this, the best way to transform the data is to divide the number of errors obtained by the maximum number of errors and then subtract from 1. In this example, the maximum is 5, the number of errors made by participant 4. This is how the error percentages in Table 8.6 were obtained. If any user had no errors (optimum), their percentage would be 100%. The percentage for the user(s) with the highest number of errors would be 0%. Note that in calculating any of these percentages, we always want higher percentages to be better—to reflect better usability. So, in the case of errors, it makes more sense to think of the resulting percentage as an "accuracy" measure.

Participant #	Tasks Completed (of 10)	# of Errors	Satisfaction Rating (0–6)	Tasks	Accuracy	Satisfaction	Average
1	8	2	4.7	80%	60%	78%	73%
2	6	4	4.1	60%	20%	68%	49%
3	7	0	3.4	70%	100%	57%	76%
4	5	5	2.4	50%	0%	40%	30%
5	9	2	5.2	90%	60%	87%	79%
6	5	4	2.7	50%	20%	45%	38%
7	10	1	5.1	100%	80%	85%	88%
8	8	1	4.9	80%	80%	82%	81%
9	7	3	3.1	70%	40%	52%	54%
10	9	2	4.2	90%	60%	70%	73%
11	7	1	4.5	70%	80%	75%	75%
12	8	3	5.0	80%	40%	83%	68%

Table 8.6 Sample data from a usability test with 12 participants[a].

[a]Tasks completed are the number of tasks (out of 10) that the user completed successfully. Number of errors is the number of specific errors that the user made, such as data-entry errors. Satisfaction rating is on a scale of 0 to 6.

WATCH OUT FOR OUTLIERS

When transforming any data where you're letting observed values determine the minimum or maximum (e.g., times or errors), you need to be particularly cautious about outliers. For example, in the data shown in Table 8.6, what if Participant #4 had made 20 errors instead of 5? The net effect would have been that his transformed percentage would still have been 0% but all of the others would have been pushed much higher. One of the standard ways of detecting outliers is by calculating the mean and standard deviation of all your data and then considering any values more than *twice* or *three times* the standard deviation away from the mean as outliers. (Most people use twice the standard deviation, but if you want to be really conservative, use three times.) For the purpose of transforming data, those outliers should be excluded. In this modified example, the mean plus twice the standard deviation of the number of errors is 14.2, while the mean plus three times the standard deviation is 19.5. By either criterion, you should treat 20 errors as an outlier and exclude it.

When transforming any usability metric to a percentage, the general rule is to first determine the minimum and maximum values that the metric can possibly have. In many cases this is easy; they are predefined by the conditions of the usability test. Here are the various cases you might encounter.

- If the minimum possible score is 0 and the maximum possible score is 100 (e.g., a SUS score), then you've basically already got a percentage. Just divide by 100 to make it a true percentage.
- In many cases, the minimum is 0 and the maximum is known, such as the total number of tasks or the highest possible rating on a rating scale. In that case, simply divide the score by the maximum to get the percentage. (This is why it's generally easier to code rating scales starting with 0 as the worst value.)
- In some cases, the minimum is 0 but the maximum is not known, such as the example of errors. In that situation, the maximum would need to be defined by the data—the highest number of errors any participant made. Specifically, the number of errors would be transformed by dividing the number of errors obtained by the maximum number of errors any participant made and subtracting that from 1.
- Finally, in some cases, neither minimum nor maximum possible scores are predefined, as with time data. In this case, you can use your data to determine the minimum and maximum values. Assuming higher values are worse, as is the case with time data, you would divide the difference between the highest value and the observed value by the difference between the highest and the lowest values.

WHAT IF HIGHER NUMBERS ARE WORSE?

Although higher numbers are better in cases such as task success rates, in other cases they're worse, such as time or errors. Higher numbers could also be worse in a rating scale if it was defined that way (e.g., 0–6, where 0 = Very Easy and 6 = Very Difficult).

In any of these cases, you must reverse the scale before averaging these percentages with other percentages where higher numbers are better. For example, with the rating scale just shown, you would subtract each value from 6 (the maximum) to reverse the scale. So 0 becomes 6 and 6 becomes 0.

8.1.3 Combining Metrics Based on *Z* Scores

Another technique for transforming scores on different scales so that they can be combined is using *z* scores. (See, for example, Martin & Bateson, 1993, p. 124.) These are based on the normal distribution and indicate how many units above or below the mean of the distribution any given value is. When you transform a set of scores to their corresponding *z* scores, the resulting distribution by definition has a mean of 0 and standard deviation of 1. This is the formula for transforming any raw score to its corresponding *z* score:

$$z = (x - \mu)/\sigma,$$

where x is the score to be transformed, μ is the mean of the distribution of those scores, and σ is the standard deviation of the distribution of those scores.

This transformation can also be done using the "=STANDARDIZE" function in Excel. Data in Table 8.2 could also be transformed using *z* scores, as shown in Table 8.7.

Participant #	Time Per Task (sec)	Tasks Completed (of 15)	Rating (0–4)	*z* Time	*z* Time* (–1)	*z* Tasks	*z* Rating	Average
1	65	7	2.4	0.98	–0.98	–0.91	–0.46	–0.78
2	50	9	2.6	0.02	–0.02	0.05	–0.20	–0.06
3	34	13	3.1	–1.01	1.01	1.97	0.43	1.14
4	70	6	1.7	1.30	–1.30	–1.39	–1.35	–1.35
5	28	11	3.2	–1.39	1.39	1.01	0.56	0.99
6	52	9	3.3	0.15	–0.15	0.05	0.69	0.20
7	58	8	2.5	0.53	–0.53	–0.43	–0.33	–0.43
8	60	7	1.4	0.66	–0.66	–0.91	–1.73	–1.10
9	25	9	3.8	–1.59	1.59	0.05	1.32	0.98
10	55	10	3.6	0.34	–0.34	0.53	1.07	0.42
Mean				0.0	0.0	0.0	0.00	0.00
Standard deviation				1.0	1.0	1.0	1.00	0.90

Table 8.7 Sample data from Table 8.2 transformed using *z* scores[a].

[a]For each original score, the z score was determined by subtracting the mean of the score's distribution from it and then dividing by the standard deviation. This z score tells you how many standard deviations above or below the mean that score is. Since you need all the scales to have higher numbers better, the scale of the z scores of times is reversed by multiplying by (–1).

EXCEL TIP

Step-by-Step Guide to Calculating *z* Scores

Here are the steps for transforming any set of raw scores (times, percentages, clicks, whatever) into *z* scores:

1. Enter raw scores into a single column in Excel. For this example, we will assume they are in column "A" and that you started on row "1". Make sure there are no other values in this column, such as an average at the bottom.
2. In the cell to the right of the first raw score, enter the formula:

   = STANDARDIZE(A1, AVERAGE(A:A), STDEV(A:A))

3. Copy this "standardize" formula down as many rows as there are raw scores.
4. As a double check, calculate the mean and standard deviation for this *z*-score column. The average should be 0, and the standard deviation should be 1 (both within rounding error).

The bottom two rows of Table 8.7 show the mean and standard deviation for each set of *z* scores, which should always be 0 and 1, respectively. Note that in using *z* scores, we didn't have to make any assumptions about the maximum or minimum values that any of the scores could have. In essence, we let each set of scores define its own distribution and rescale them so those distributions would each have a mean of 0 and standard deviation of 1. In this way, when they are averaged together, each of the *z* scores makes an equal contribution to the average *z* score. Note that when averaging the *z* scores together, each of the scales must be going the same direction—in other words, higher values should always be better. In the case of time data, the opposite is almost always true. Since *z* scores have a mean of 0, this is easy to correct simply by multiplying the *z* score by (–1) to reverse its scale.

If you compare the *z*-score averages in Table 8.7 to the percentage averages in Table 8.3, you will find that the ordering of the participants based on those averages is nearly the same: Both techniques yield the same top three participants (9, 5, and 3) and the same bottom three participants (4, 8, and 1).

One disadvantage of using *z* scores is that you can't think of the overall average of the *z* scores as some type of overall usability score, as by definition that overall average will be 0. So when would you want to use *z* scores? They mainly are useful when you want to compare one set of data to another, such as data from iterative usability tests of different versions of a product, data from different groups of users in the same usability test, or data from different conditions or designs within the same usability test. You should also have a reasonable sample size (e.g., at least 10 participants per condition) to use the *z*-score method.

For example, consider the data shown in Figure 8.1 from Chadwick-Dias, McNulty, and Tullis (2003), which shows *z* scores of performance for two iterations of a prototype. This research studied the effects of age on performance in using a website. Study 1 was a baseline study. Based on their observations of the participants in Study 1, especially the problems encountered by the older participants, they made changes to the prototype and then conducted Study 2 with a new group of participants. The *z* scores were equal-weighted combinations of task time and task completion rate.

It's important to understand that the *z*-score transformations were done using the *full set* of data from Study 1 and Study 2 combined. They were then plotted appropriately to indicate from which study each *z* score was derived. The key finding was that the performance *z* scores for Study 2 were significantly higher than the performance *z* scores for Study 1; the effect was the same regardless of age (as reflected by the fact that the two lines are parallel to each other). If the *z*-score transformations had been done *separately* for Study 1 and Study 2, the results would have been meaningless because the means for Study 1 and Study 2 would both have been forced to 0 by the transformations.

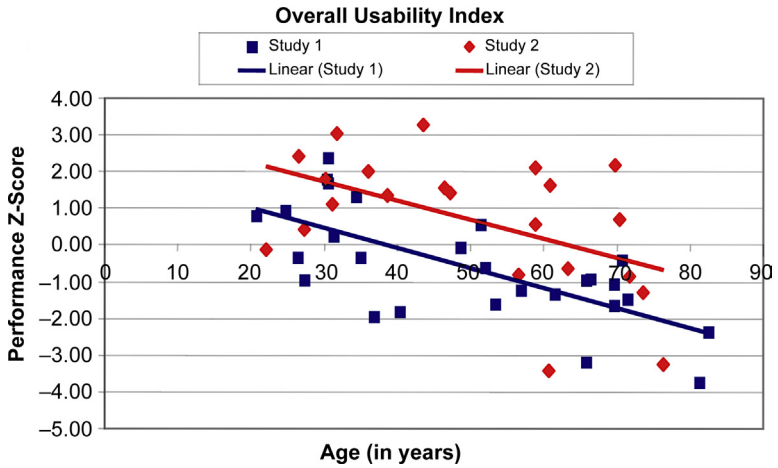


Figure 8.1 Data showing performance *z* scores from two studies of a prototype with participants over a wide range of ages. The performance *z* score was an equal-weighted combination of task time and task completion rate. Changes were made to the prototype between Study 1 and Study 2. The performance *z* scores were significantly better in Study 2, regardless of the participant's age. Adapted from Chadwick-Dias et al. (2003); used with permission.

8.1.4 Using Single Usability Metric

Jeff Sauro and Erika Kindlund (2005) developed a quantitative model for combining usability metrics into a single usability score. Their focus is on task completion, task time, error counts per task, and post-task satisfaction rating

(similar to ASQ described in Chapter 6). Note that all of their analyses are at the task level, whereas the previous sections have described analyses at the "usability test" level. At the task level, task completion is typically a binary variable for each participant: that person either completed the task successfully or did not. At the usability-test level, task completion, as shown in previous sections, indicates how many tasks each person completed, and it can be expressed as a percentage for each participant.

Sauro and Kindlund used techniques derived from Six Sigma methodology (e.g., Breyfogle, 1999) to standardize their four usability metrics (task completion, time, errors, and task rating) into a SUM. Conceptually, their techniques are not that different from the *z* score and percentage transformations described in the previous sections. In addition, they used Principal Components Analysis, a statistical technique that looks at correlations between variables, to determine if all four of their metrics were contributing significantly to the overall calculation of the single metric. They found that all four were significant and, in fact, that each contributed about equally. Consequently, they decided that each of the four metrics (once standardized) should contribute equally to the calculation of the SUM score.

An online tool for entering data from a usability test and calculating the SUM score is available from Jeff Sauro's "Usability Scorecard" website at http://www.usabilityscorecard.com/. For each task and each participant in the usability test, you must enter the following:

- Whether the participant completed the task successfully (0 or 1).
- Number of errors committed on that task by that participant. (You also specify the number of error opportunities for each task.)
- Task time in seconds for that participant.
- Post-task satisfaction rating, which is an average of three post-task ratings on five-point scales of task ease, satisfaction, and perceived time—similar to ASQ.

After entering these data for all the tasks, the tool standardizes the scores and calculates the SUM score for each task. Standardized data shown for each task are illustrated in Table 8.8. Note that a SUM score is calculated for each task, which allows for overall comparisons of tasks. In these sample data, participants did best on the "Cancel reservation" task and worst on the "Check restaurant hours" task. An overall SUM score, 68% in this example, is also calculated, as is a 90% confidence interval (53 to 88%), which is the average of the confidence intervals of the SUM score for each task.

	SUM						
Task	**Low**	**Mean**	**High**	**Completion**	**Satisfaction**	**Time**	**Errors**
Reserve a room	62%	75%	97%	81%	74%	68%	76%
Find a hotel	38%	58%	81%	66%	45%	63%	59%
Check room rates	49%	66%	89%	74%	53%	63%	74%
Cancel reservation	89%	91%	99%	86%	91%	95%	92%
Check restaurant hours	22%	46%	68%	58%	45%	39%	43%
Get directions	56%	70%	93%	81%	62%	66%	71%
Overall SUM	53%	68%	88%				

Table 8.8 Sample standardized data from a usability test[a].
[a]After entering data for each participant and each task, these are the standardized scores calculated by SUM, including an overall SUM score and a confidence interval for it.

The online tool also provides the option to graph task data from a usability study, including the SUM scores. Figure 8.2 shows a sample graph from the tool.

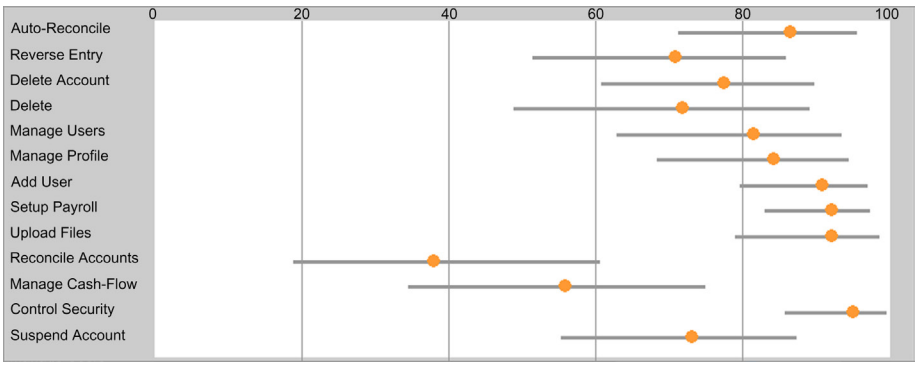


Figure 8.2 Sample graph of SUM scores from http://www.usabilityscorecard.com/. The tasks of this usability test are listed down the left. For each task, the orange circle shows the mean SUM score and bars show the 90% confidence interval for each. In this example, it's apparent that the "Reconcile Accounts" and "Manage Cash-Flow" tasks are the most problematic.

8.2 USABILITY SCORECARDS

An alternative to combining different metrics to derive an overall usability score is to present the results of the metrics graphically in a summary chart. This type of chart is often called a Usability Scorecard. The goal is to present data from

the usability test in such a way that overall trends and important aspects of the data can be detected easily, such as tasks that were particularly problematic for the users. If you only have two metrics that you're trying to represent, a simple combination graph from Excel may be appropriate. For example, Figure 8.3 shows the task completion rate and task ease rating for each of 10 tasks in a usability test.

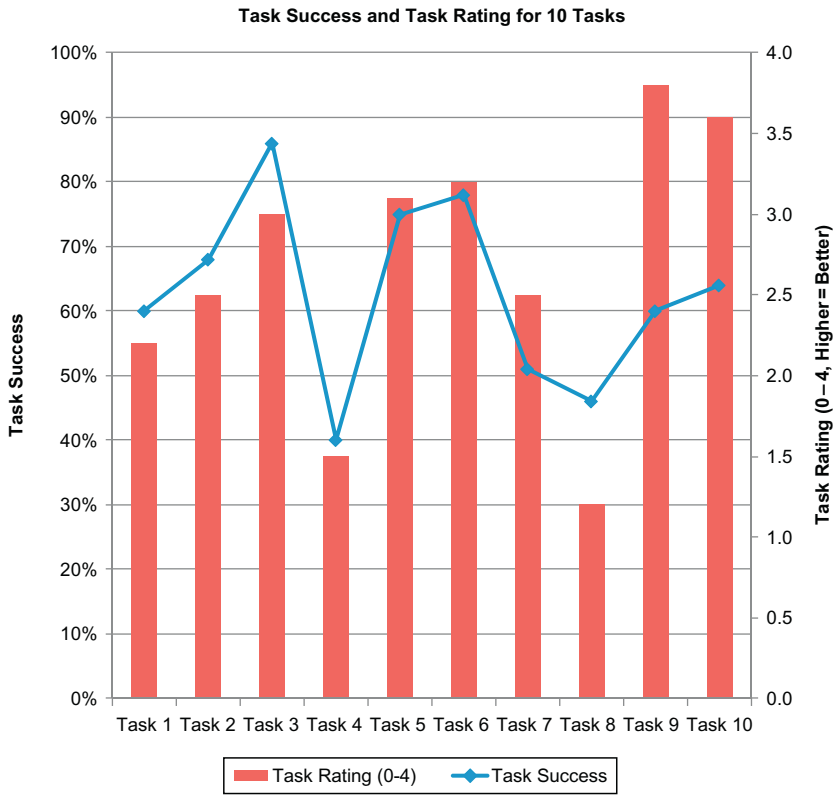


Figure 8.3 A sample combination column and line chart for 10 tasks. Task rating is shown via the columns and labeled on the right axis. Task success is shown via the lines and is labeled on the left axis.

The combination chart in Figure 8.3 has some interesting features. It clarifies which tasks were the most problematic for the participants (Tasks 4 and 8) because they have the lowest values on both scales. It's also obvious where there were significant disparities between task success data and task ease ratings, such as Tasks 9 and 10, which had only moderate task completion rates but the highest task ratings. (This is an especially troubling finding because it might indicate that some of the users did not complete the task successfully but thought they did.) Finally, it's easy to distinguish the tasks that had reasonably high values for both metrics, such as Tasks 3, 5, and 6.

HOW TO CREATE A COMBINATION CHART IN EXCEL

Older versions of Excel made it easy to create this type of combination chart with two axes, but it's a bit more challenging in the newer versions (2007 and higher). Here's what you do:

1. Enter your data into two columns in the spreadsheet (e.g., one column for task success and the other for task rating). Create a column chart like you normally would for both variables. This will look strange because the two variables will be plotted on the same axis, with one scale overshadowing the other greatly.
2. Right-click on one of the columns in the chart and choose "Format Data Series." In the resulting dialog box, choose "Series Options." In the "Plot Series On" area, choose "Secondary Axis."
3. Close that dialog box. The chart will still look odd because now the two columns are on top of each other.
4. Right click on a column being charted on the primary (left) axis and select "Change Series Chart Type."
5. Change that variable to a line graph. Close that dialog box.

(Yes, we know this type of combination chart breaks the rule about only using line graphs for continuous data, like times. But you have to break the rule to make it work in Excel. And rules are made to be broken anyway!)

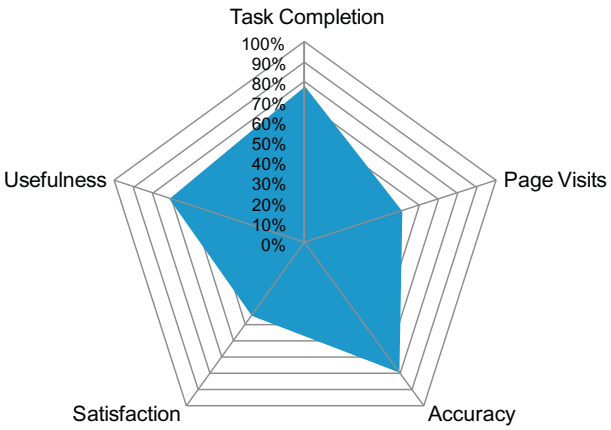


Figure 8.4 A sample radar chart summarizing task completion, page visits, accuracy (lack of errors), satisfaction rating, and usefulness rating from a usability test. Each has been transformed to a percentage using the techniques outlined earlier in this chapter.

This type of combination chart works well if you have only two metrics to represent, but what if you have more? One way of representing summary data for three or more metrics is using radar charts (which were also illustrated in Chapter 6). Figure 8.4 shows an example of a radar chart for summarizing the results of a usability test with five factors: task completion, page visits, accuracy (lack of errors), satisfaction rating, and usefulness rating. In this example, although task completion, accuracy, and usefulness rating were relatively high (good), the page visits and satisfaction rating were relatively low (poor).

Although radar charts can be useful for a high-level view, it's not really possible to represent task-level information in them. The example in Figure 8.4 averaged data across the tasks. What if you want to represent summary data for three or more metrics but also maintain task-level information? One technique for doing that is using what are called Harvey Balls. A variation on this technique has been popularized by *Consumer Reports*. For

example, consider the data shown earlier in Table 8.7, which presents the results for six tasks in a usability test, including task completion, time, satisfaction, and errors. These data could be summarized in a comparison chart as shown in Figure 8.5. This type of comparison chart allows you to see at a glance how the participants did for each of the tasks (by focusing on the rows) or how the participants did for each of the metrics (by focusing on the columns).

Task	SUM Score	Completion	Satisfaction	Time	Errors
Cancel reservation	91%				
Reserve a room	75%				
Get directions	70%				
Check room rates	66%				
Find a hotel	58%				
Check restaurant hours	46%				

Figure 8.5 A sample comparison chart using data from Table 8.7. Tasks have been ordered by their SUM score, starting with the highest. For each of the four standardized scores (task completion, satisfaction, task time, and errors), the value has been represented by coded circles (known as Harvey Balls), as shown in the key.

WHAT ARE HARVEY BALLS?

Harvey Balls are small, round pictograms used typically in a comparison table to represent values for different items:

They're named for Harvey Poppel, a Booz Allen Hamilton consultant who created them in the 1970s as a way of summarizing long tables of numeric data. There are five levels, progressing from an open circle to a completely filled circle. Typically, the open circle represents the worst values, and the completely filled circle represents the best values. Links to images of Harvey Balls of different sizes can be found on our website, www.MeasuringUX.com. Harvey Balls shouldn't be confused with Harvey Ball, who was the creator of the smiley face ☺!

8.3 COMPARISON TO GOALS AND EXPERT PERFORMANCE

Although the previous section focused on ways to summarize usability data without reference to an external standard, in some cases you may have an external standard that can be used for comparison. The two main flavors of an external standard are predefined goals and expert, or optimum, performance.

8.3.1 Comparison to Goals

Perhaps the best way to assess the results of a usability test is to compare those results to goals that were established before the test. These goals may be set at the task level or an overall level. Goals can be set for any of the metrics we've discussed, including task completion, task time, errors, and self-reported measures. Here are some examples of task-specific goals:

- At least 90% of representative users will be able to reserve a suitable hotel room successfully.
- Opening a new account online should take no more 8 minutes on average.
- At least 95% of new users will be able to purchase their chosen product online within 5 minutes of selecting it.

Similarly, examples of overall goals could include the following:

- Users will be able to complete at least 90% of their tasks successfully.
- Users will be able to complete their tasks in less than 3 minutes each, on average.
- Users will give the application an average SUS rating of at least 80%.

Typically, usability goals address task completion, time, accuracy, and/or satisfaction. The key is that the goals must be measurable. You must be able to determine whether the data in a given situation supports the attainment of the goal. For example, consider the data in Table 8.9.

	Target # of Page Visits	Actual # of Page Visits
Task 1	5	7.9
Task 2	8	9.3
Task 3	3	7.3
Task 4	10	11.5
Task 5	4	7
Task 6	6	6.9
Task 7	9	9.8
Task 8	7	10.2

Table 8.9 Sample data from eight tasks showing target number of page visits and mean of actual number of page visits.

Table 8.9 shows data for eight tasks in a usability study of a website. For each task, a target number of page visits has been predetermined (ranging from 4 to 10). Figure 8.6 depicts the target and actual page views for each task graphically. This chart is useful because it allows you to visually compare the actual number of page visits for each task, and its associated confidence interval, to the target number of page views. In fact, all the tasks had significantly more page views than the targets. What's perhaps not so obvious is how the various tasks performed relative to each other—in other words, which ones came out better and which ones worse. To make that kind of comparison easier, Figure 8.7 shows the ratio of the target to actual page views for each task. This can be thought of as a "page view efficiency" metric: the closer it is to 100%, the more efficient the participants were being. This makes it easy to spot tasks where the participants had trouble (e.g., Task 3) versus tasks where they did well (e.g., Task 7). This technique could be used to represent the percentage of participants who met any particular objective (e.g., time, errors, SUS rating) either at the task level or at the overall level.

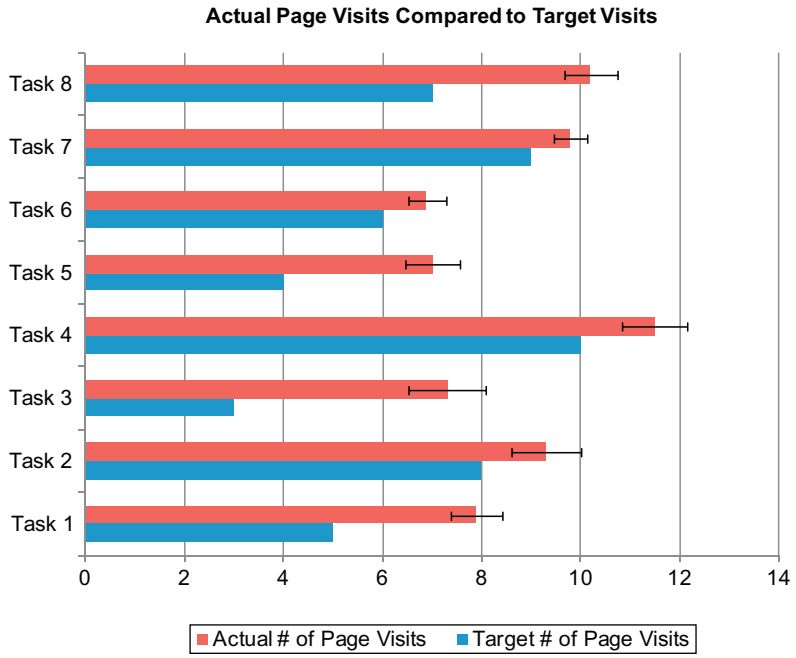


Figure 8.6 Target and actual number of page visits for each of eight tasks. Error bars represent the 90% confidence interval for the actual number of page visits.

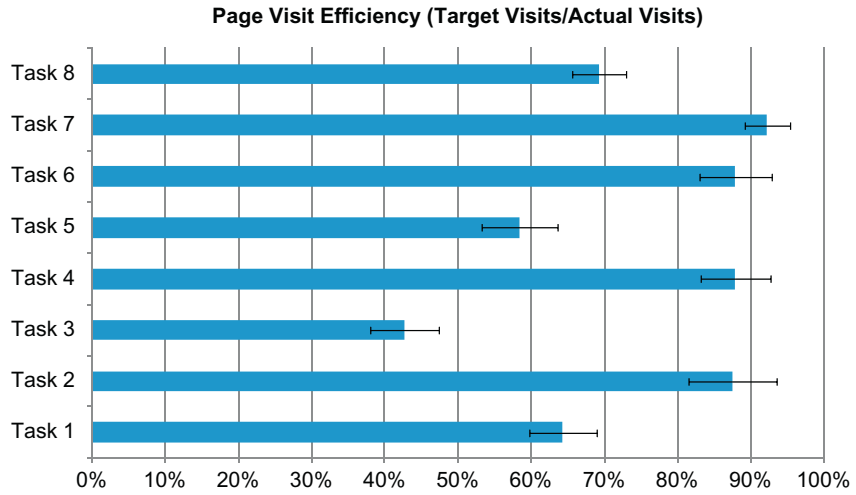


Figure 8.7 Ratio of target to actual page views for each of the eight tasks.

8.3.2 Comparison to Expert Performance

An alternative to comparing the results of a usability test to predefined goals is to compare the results to the performance of experts. The best way to determine the expert performance level is to have one or more presumed experts actually perform the tasks and to measure the same things that you're measuring in the usability test. Obviously your experts really need to be experts—people with subject-matter expertise, in-depth familiarity with the tasks, and in-depth familiarity with the product, application, or website being tested. And your data will be better if you can average the performance results from more than one expert. Comparing results of a usability test to results for experts allows you to compensate for the fact that certain tasks may be inherently more difficult or take longer, even for an expert. The goal, of course, is to see how close the performance of the participants in the test actually comes to the performance of the experts.

Although you could theoretically do a comparison to expert performance for any performance metric, it's used most commonly for time data. With task success data, the usual assumption is that a true expert would be able to perform all the tasks successfully. Similarly, with error data, the assumption is that an expert would not make any errors. But even an expert would require some amount of time to perform the tasks. For example, consider the task time data shown in Table 8.10.

Task	Actual Time	Expert Time	Expert/Actual
1	124	85	69%
2	101	50	50%
3	89	70	79%
4	184	97	53%
5	64	40	63%
6	215	140	65%
7	70	47	67%
8	143	92	64%
9	108	98	91%
10	92	60	65%

Table 8.10 Sample time data from 10 tasks in a usability test showing average actual time per task (in seconds), expert time per task, and ratio of expert to actual time.

Graphing the ratio of expert to actual times, as shown in Figure 8.8, makes it easy to spot tasks where the test participants did well in comparisons to the experts (Tasks 3 and 9) and tasks where they did not do so well (Tasks 2 and 4).

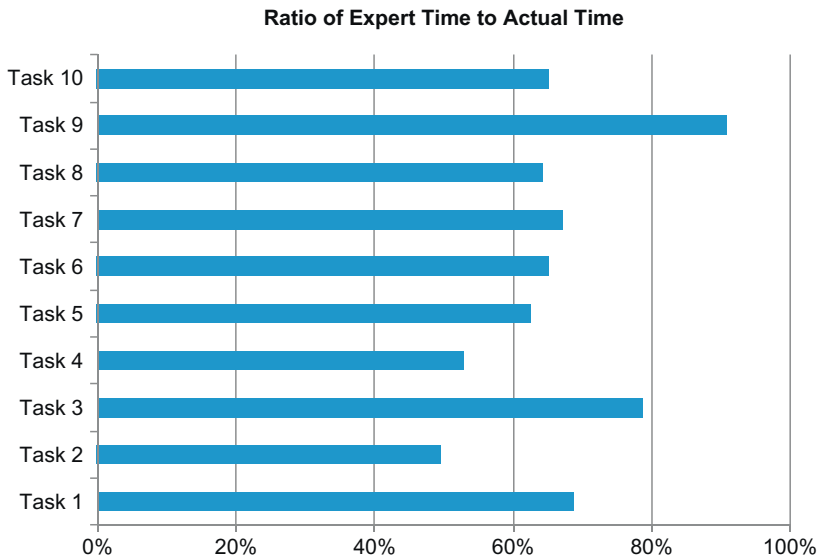


Figure 8.8 Graph of the ratio of expert to actual times from Table 8.10.

8.4 SUMMARY

Some of the key takeaways from this chapter are as follow.

1. An easy way to combine different usability metrics is to determine the percentage of users who achieve a combination of goals. This tells you the overall percentage of users who had a good experience with your product (based on the target goals). This method can be used with any set of metrics and is understood easily by management.
2. One way of combining different metrics into an overall "usability score" is to convert each of the metrics to a percentage and then average them together. This requires being able to specify, for each metric, an appropriate minimum and maximum value.
3. Another way to combine different metrics is to convert each metric to a z score and then average them together. Using z scores, each metric gets equal weight when they are combined. But the overall average of the z scores will always be 0. This metric is useful in comparing different subsets of the data to each other, such as data from different iterations, different groups, or different conditions.
4. The SUM technique is another method for combining different metrics, specifically task completion, task time, errors, and task-level satisfaction rating. The method requires entry of individual task and participant data for the four metrics. Calculations yield a SUM score, as a percentage, for each task and across all tasks, including confidence intervals.
5. Various types of graphs and charts can be useful for summarizing the results of a usability test in a "usability scorecard." A combination line and column chart is useful for summarizing the results of two metrics for tasks in a test. Radar charts are useful for summarizing the results of three or more metrics overall. A comparison chart using Harvey Balls to represent different levels of the metrics can summarize effectively the results for three or more metrics at the task level.
6. Perhaps the best way to determine the success of a usability test is to compare the results to a set of predefined usability goals. Typically these goals address task completion, time, accuracy, and satisfaction. The percentage of users whose data met the stated goals can be a very effective summary.
7. A reasonable alternative to comparing to predefined goals, especially for time data, is to compare actual performance data to data for experts. The closer the actual performance is to expert performance, the better.

CHAPTER 9

Special Topics

CONTENTS

This chapter introduces a number of topics related to the measurement or analysis of user experience data not traditionally thought of as part of "mainstream" UX data. These include information you can glean from live data on a production website, data from card-sorting studies, data related to the accessibility of a website, and UX return on investment (ROI). These topics didn't fit neatly into the other chapters, but we believe they are an important part of a complete UX metrics toolkit.

9.1 LIVE WEBSITE DATA

If you're dealing with a live website, there's a potential treasure trove of data about what the visitors to your site are actually doing—what pages they're visiting, what links they're clicking on, and what paths they're following through the site. The challenge usually isn't getting raw data but making sense of it. Unlike lab studies with perhaps a dozen participants or online studies with perhaps 100 participants, live sites have the potential to yield data from thousands or even hundreds of thousands of users.

Entire books have been written on only the subject of web metrics and web analytics (e.g., Burby & Atchison, 2007; Clifton, 2012; Kaushik, 2009). There's even a "For Dummies" book on the topic (Sostre & LeClaire, 2007). So obviously

Measuring the User Experience. DOI: http://dx.doi.org/10.1016/B978-0-12-415781-1.00009-1

we won't be able to do justice to the topic in just one section of one chapter in this book. What we'll try to do is introduce you to some of the things you can learn from live website data and specifically some of the implications they might have for the usability of your site.

9.1.1 Basic Web Analytics

Some websites get huge numbers of visitors every day. But regardless of how many visitors your site gets (assuming it gets some), you can learn from what they're doing on the site.

A number of tools are available for capturing web analytics. Most web-hosting services provide basic analytics as part of the hosting service, and other web

SOME WEB ANALYTICS TERMS

Here are the meanings of some of the terms used commonly in web analytics.

- **Visitors.** The people who have visited your website. Usually a visitor is counted only once during the time period of a report. Some analytics packages use the term "unique visitor" to indicate that they're not counting the same person more than once. Some also report "new visitors" to distinguish them from ones who have been to your site before.
- **Visits.** The individual times that your website was accessed; sometimes also called "sessions." An individual *visitor* can have multiple *visits* to your site during the time period of the report.
- **Page views.** The number of times individual pages on your site are viewed. If a visitor reloads a page, that typically counts as a new page view; likewise, if visitors navigate to another page in your site and then return to a page, that will count as a new page view. Page views let you see which pages on your site are the most popular.
- **Landing page or entrance page.** The first page that a visitor visits on your site. This is often the home page, but might be a lower level page if they found it through a search engine or had bookmarked it.
- **Exit page.** The last page that a visitor visits on your site.
- **Bounce rate.** The percentage of visits in which the visitor views only one page on your site and then leaves the site. This could indicate a lack of engagement with your site, but it could also mean that they found what they were looking for from that one page.
- **Exit rate (for a page).** The percentage of visitors who leave your site from a given page. *Exit rate*, which is a metric at an individual page level, is often confused with *bounce rate*, which is an overall metric for a site.
- **Conversion rate.** The percentage of visitors to a site who convert from being simply a casual visitor to taking some action, such as making a purchase, signing up for a newsletter, or opening an account.

analytics services are available for free. Perhaps the most popular free analytics service is Google Analytics (http://www.google.com/analytics/). Figure 9.1 shows a screenshot from Google Analytics.

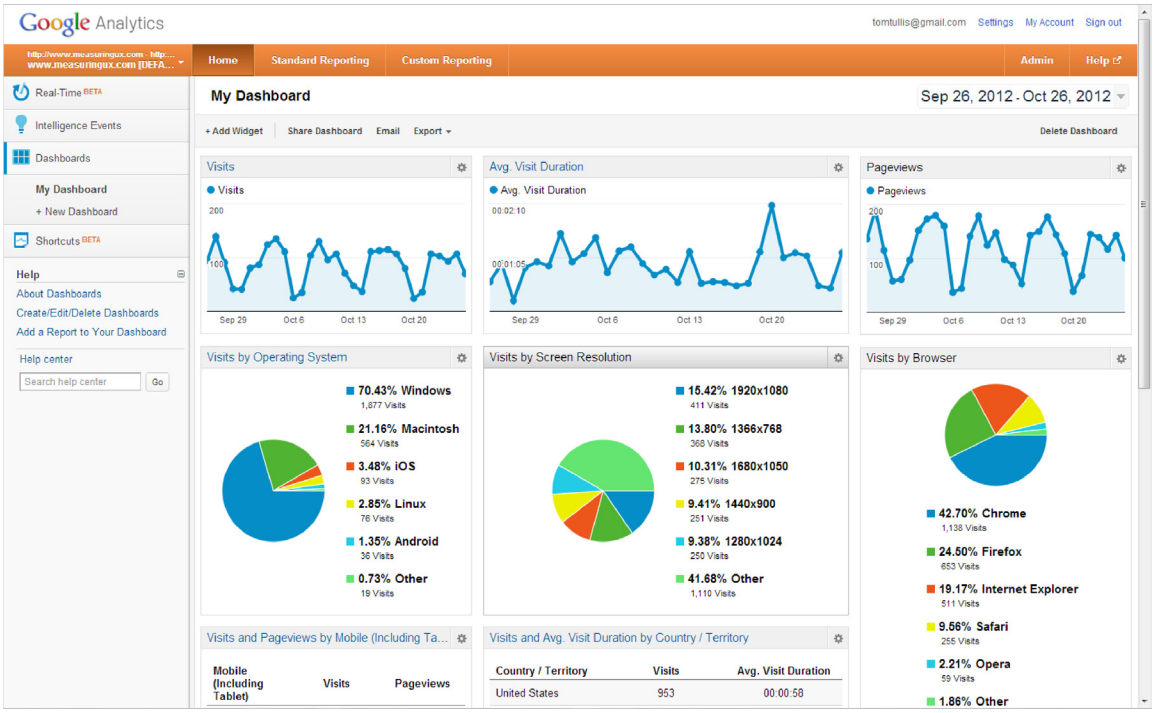


Figure 9.1 Sample Google Analytics screen for the MeasuringUX.com site.

As can be seen in Figure 9.1, you can look at many of the metrics for your site over time, such as line graphs for visits, average visit duration, and page views. These graphs of visits and page views show a pattern that's typical for some websites, which is a difference in the number of visitors, visits, and page views for the weekend vs the weekdays. You also can capture some basic information about the visitors to your site, such as the operating system they're running, their screen resolution, and the browsers they're using, as illustrated in the pie charts.

Simply looking at the number of page views for various pages in your site can be enlightening, especially over time or across iterations of the site. For example, assume that a page about Product A on your site was averaging 100 page views per day for a given month. Then you modified the homepage for your site, including the description of the link to Product A's page. Over the next month, the Product A page then averaged 150 page views per day. It would certainly appear that the changes to the homepage significantly increased the number of visitors accessing the Product A page. But you need to be careful that other factors didn't cause the increase. For example, in the financial-services world,

certain pages have seasonal differences in their number of page views. A page about contributions to an Individual Retirement Account (IRA), for example, tends to get more visits in the days leading up to April 15 because of the deadline in the United States for contributing to the prior year's IRA.

It's also possible that something caused your site as a whole to start getting more visitors, which certainly could be a good thing. But it could also be due to factors not related to the design or usability of your site, such as news events related to the subject matter of your site. This also brings up the issue of the impact that search "bots" can have on your site's statistics. Search bots, or spiders, are automated programs used by most of the major search engines to "crawl" the web by following links and indexing the pages they access. One of the challenges, once your site becomes popular and is being "found" by most of the major search engines, is filtering out the page views due to these search bots. Most bots (e.g., Google, Yahoo!) usually identify themselves when making page requests and thus can be filtered out of the data.

What analyses can be used to determine if one set of page views is significantly different from another set? Consider the data shown in Table 9.1, which shows the number of page views per day for a given page over two different weeks. Week 1 was before a new homepage with a different link to the page in question was launched, and Week 2 was after.

	Week 1	**Week 2**
Sun	237	282
Mon	576	623
Tue	490	598
Wed	523	612
Thu	562	630
Fri	502	580
Sat	290	311
Averages	454	519

Table 9.1 Number of page views for a given web page over 2 different weeks[a].
[a]Week 1 was before a new homepage was launched, and Week 2 was after. The new homepage contained different wording for the link to this page.

These data can be analyzed using a paired *t* test to see if the average for Week 2 (519) is significantly different from the average for Week 1 (454). It's important to use a paired *t* test because of the variability due to the days of the week; comparing each day to itself from the previous week takes out the variability due to days. A paired *t* test shows that this difference is statistically significant ($p < 0.01$). If you had not used a paired *t* test, and just used a *t* test for two

independent samples, the result ($p = 0.41$) would not have been anywhere near significant. (See Chapter 2 for details on how to run a paired t test in Excel.)

9.1.2 Click-Through Rates

Click-through rates can be used to measure the effectiveness of different ways of presenting a link or button. They indicate the percentage of visitors who are shown a particular link or button who then actually click on it. If a link is shown 100 times and it is clicked on 1 of those times, its click-through rate is 1%. Most commonly the term is used to measure the effectiveness of web ads, but the concept applies to any link, button, or clickable image. For example, Holland (2012a) describes a study of two different buttons on a product page for an ecommerce site, as shown in Figure 9.2. The only difference between the two pages was the wording of the green button: "Personalize Now" vs "Customize It". The click-through rate for the button labeled "Personalize Now" was 24% higher. They continued tracking through to actual sales and found that clicks on that version of the button also resulted in a 48% higher revenue per visitor. Why did the "Personalize Now" text yield more clicks and more sales? We could speculate, but we don't really know. That's one of the limitations of live-site data.

Figure 9.2 Example of two button designs tested on a product page.

What analyses can be used to determine if the click-through rate for one link is significantly different from that for another link? One appropriate analysis is the χ^2 test. A χ^2 test lets you determine whether an observed set of frequencies is significantly different from an expected set of frequencies. (See Chapter 2 for more details.) For example, consider the data shown in Table 9.2 that represent

	Click	No Click
Link #1	145	10289
Link #2	198	11170

Table 9.2 Click rates for two different links: The number of times each link was clicked and the number of times each was presented but not clicked.

click rates for two different links. The click-through rate for Link #1 is 1.4% [145/(145 + 10,289)]. The click-through rate for Link #2 is 1.7% [198/(198 + 11,170)]. But are these two significantly different from each other? Link #2 got more clicks, but it was also presented more times. To do a χ^2 test, you must first construct a table of expected frequencies as if there were no difference in the click-through rates of Link #1 and Link #2. This is done using the sums of the rows and columns of the original table, as shown in Table 9.3.

Observed	Click	No Click	Sum
Link #1	145	10289	10434
Link #2	198	11170	11368
Sum	343	21459	21802

Table 9.3 Same data as Table 9.2 but with sums of rows and columns added[a].
[a]*These are used to calculate expected frequencies if there were no differences in the click-through rates.*

By taking the product of each pair of row and column sums and dividing that by the grand total you get the expected values as shown in Table 9.4. For example, the expected frequency for a "Click" on "Link #1" (164.2) is the product of the respective row and column sums divided by the grand total: (343×10,434)/21,802. The "CHITEST" function in Excel can then be used to compare the actual frequencies in Table 9.2 to the expected frequencies in Table 9.4. The resulting value is $p = 0.04$, indicating that a significant difference exists between the click-through rates for Link #1 and Link #2.

Expected	Click	No Click
Link #1	164.2	10269.8
Link #2	178.8	11189.2

Table 9.4 Expected frequencies if there were no differences in click-through rates for Link #1 and Link #2, derived from sums shown in Table 9.3.

You should keep two important points about the χ^2 test in mind. First, the χ^2 test must be done using raw frequencies or counts, *not* percentages. You commonly think of click-through rates in terms of percentages, but that's not how you test for significant differences between them. Also, the categories used must be *mutually exclusive* and *exhaustive*, which is why the preceding example used "Click" and "No Click" as the two categories of observations for each link. Those two categories are mutually exclusive and account for all possible actions that could be taken on the link—either the user clicked on it or didn't.

9.1.3 Drop-Off Rates

Drop-off rates can be a particularly useful way of detecting where there might be some usability problems on your site. The most common use of drop-off rates is to identify where in a sequence of pages users are dropping out of or abandoning a process, such as opening an account or completing a purchase. For example, assume that the user must fill out the information on a sequence of five pages to open some type of account. Table 9.5 reflects the percentage of users who started the process that actually completed each of the five pages.

In this example, all of the percentages are relative to the number of users who started the entire process—that is, who got to Page #1. So 89% of the users who got to Page #1 completed it successfully, 80% of that original number completed Page #2, and so on. Given the data in Table 9.5, which of the five pages do the users seem to be having the most trouble with? The key is to look at how many users dropped off from each page—in other words, the difference between how many got to the page and how many completed it. Those "drop-off percentages" for each of the pages are shown in Table 9.6.

Page #1	89%
Page #2	80%
Page #3	73%
Page #4	52%
Page #5	49%

Table 9.5 Percentage of users who started a multipage process that actually completed each of the steps.

Page #1	11%
Page #2	9%
Page #3	7%
Page #4	21%
Page #5	3%

Table 9.6 Drop-off percentages for each page shown in Table 9.5: The difference between percentage who got to the page and percentage who completed it successfully.

This makes it clear that the largest drop-off rate, 21%, is associated with Page #4. If you're going to redesign this multipage process, you would be well advised to learn what's causing the drop-off at Page #4 and then try to address that in the redesign.

9.1.4 A/B Tests

A/B tests are a special type of live-site study in which you manipulate elements of the pages that are presented to the users. The traditional approach to A/B testing on a website involves posting two alternative designs for a given page or elements of a page. Some visitors to the site see the "A" version whereas others see the "B" version. In many cases, this assignment is random, so about the same number of visitors sees each version. In some cases, the majority of visitors see the existing page, and a smaller percentage see an experimental version that's being tested. Although these studies are typically called A/B tests, the same concept applies to any number of alternative designs for a page.

WHAT MAKES A GOOD A/B TEST?

A good A/B test requires careful planning. Here are some tips to keep in mind:

- Make sure the method you're using to "split" visitors between "A" and "B" versions really is random. If someone tells you it's good enough to just send all visitors in the morning to version "A" and all visitors in the afternoon to version "B", don't believe it. There could be something different about morning visitors vs afternoon visitors.
- Test small changes, especially at first. It might be tempting to design two completely different versions of a page, but you'll learn much more by testing small differences. If the two versions are completely different from each other, and one performs significantly better than the other, you still don't know why that one was better. If the only difference is, for example, wording of the call-to-action button, then you know the difference is due to that wording.
- Test for significance. It might look like one version is beating the other one, but do a statistical test (e.g., χ^2) to make sure.
- Be agile. When you're confident that one version is outperforming the other, then "promote" the winning version (i.e., send all visitors to it) and move on to another A/B test.
- Believe the data, not the HIPPO (Highest Paid Person's Opinion). Sometimes the results of A/B tests are surprising and counterintuitive. One of the advantages that UX researchers bring to the mix is that you can follow up on these surprising findings using other techniques (e.g., surveys, lab, or online studies) to try to understand them better.

Technically, visitors to a page can be directed to one of the alternative pages in a variety of ways, including based on random number generation, the exact time (e.g., an even or odd number of seconds since midnight), or several other techniques. Typically, a cookie is set to indicate which version the visitor was shown so that if he or she returns to the site within a specified time period, the same version will be shown again. Keep in mind that it's important to test the alternative versions *at the same time* because of the external factors mentioned before that could affect the results if you tested at different times.

Holland (2012b) described an A/B test of the page layout for an online newspaper. As shown in Figure 9.3, the two versions differed in the relationship of the photos to the articles they accompanied. In Version A, photos alternated between the left and the right of the articles. In Version B, photos were always to

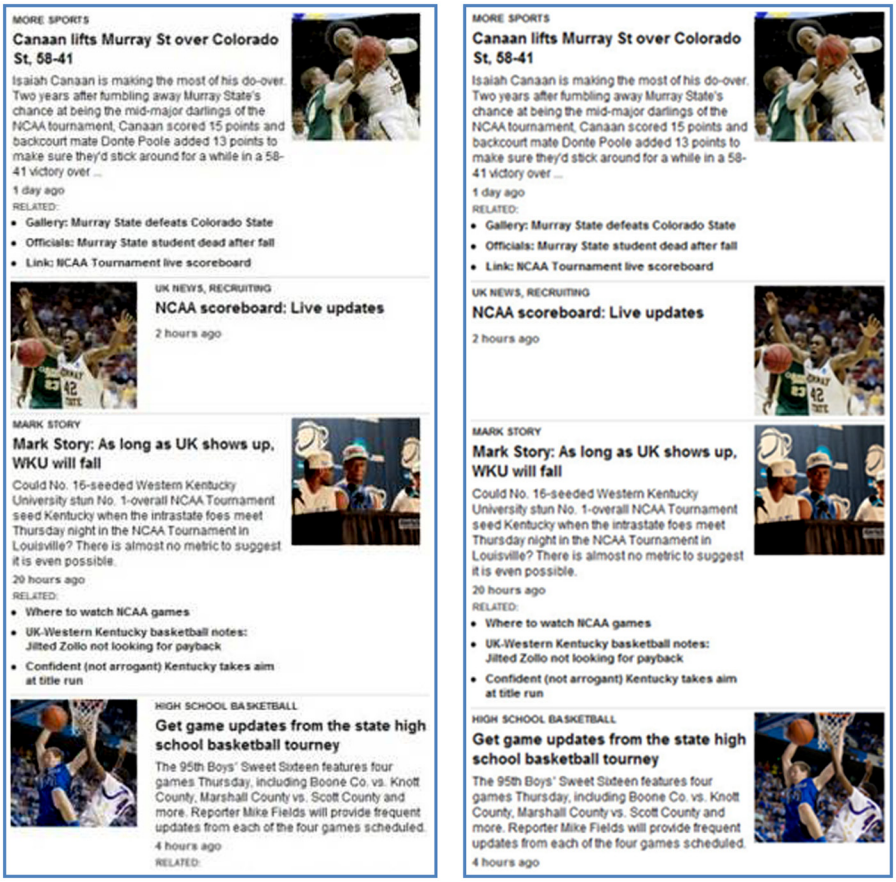


Figure 9.3 Two versions of page layout for an online newspaper. In Version A (left), photos for each article alternated between the left and the right of the article. In Version B (right), they were always to the right.

the right. They measured the click rates on the articles (to read the full version of the article). Version B, with article photos always to the right, increased clicks by 20% and total site pages viewed by 11%.

WHICH TESTWON.COM

Anne Holland runs a website called "Which Test Won" that's a treasure trove of examples of A/B tests. As of the writing of the second edition of this book, she has about 300 examples of different A/B tests on her site, ranging from tests that manipulated entire page designs to tests where the only difference was the color of a single button. She posts a new test every week, encouraging readers to guess whether the A or the B version of the test won. She also has a free e-mail newsletter alerting readers to new tests on the site.

Carefully designed A/B tests can give you significant insight into what works and what doesn't work on your website. Many companies, including Amazon, eBay, Google, Microsoft, Facebook, and others, are constantly doing A/B tests on their live sites, although most users don't notice it (Kohavi, Crook, & Longbotham, 2009; Kohavi, Deng, Frasca, Longbotham, Walker, & Xu, 2012; Tang, Agarwal, O'Brien, & Meyer, 2010). In fact, as Kohavi and Round (2004) explained, A/B testing is constant at Amazon, and experimentation through A/B testing is the main way they make changes to their site.

9.2 CARD-SORTING DATA

Card sorting as a technique for organizing the elements of an information system in a way that makes sense to the users has been around at least since the early 1980s. For example, Tullis (1985) used the technique to organize the menus of a mainframe operating system. More recently, the technique has become popular as a way of informing decisions about the information architecture of a website (e.g., Maurer & Warfel, 2004; Spencer, 2009). Over the years the technique has evolved from a true card-sorting exercise using index cards to an online exercise using virtual cards. Although many UX professionals seem to be familiar with the basic card-sorting techniques, fewer seem to be aware that various metrics can be used in the analyses of card-sorting data.

The two major types of card-sorting exercises are (1) open card sorts, where you give users the cards that are to be sorted but let them define their own groups that the cards will be sorted into, and (2) closed card sorts, where you give users the cards to be sorted as well as the names of the groups to sort them into. Although some metrics apply to both, others are unique to each.

CARD-SORTING TOOLS

A number of tools are available for conducting card-sorting exercises. Some are desktop applications, whereas others are web based. Most of these include basic analysis capabilities (e.g., hierarchical cluster analysis). Here are some of the ones we're familiar with:

- CardZort (http://www.cardzort.com/cardzort/)(a Windows application)
- OptimalSort (http://www.optimalworkshop.com/optimalsort.htm)(a web-based service)
- UsabiliTest Card sorting (http://www.usabilitest.com/CardSorting)(a web-based service)
- UserZoom Card sorting (http://www.userzoom.com/products/card-sorting)(a web-based service)
- UzCardSort (http://uzilla.mozdev.org/cardsort.html)(a Mozilla extension)
- Websort (http://www.websort.net/)(a web-based service)
- XSort (http://www.xsortapp.com/)(a Mac OS X application)

Although not a card-sorting tool, you could also use PowerPoint or similar programs to do card-sorting exercises when the number of cards is relatively small. Create a slide that has the cards to be sorted along with empty boxes and then e-mail that to participants, asking them to put the cards into the boxes and to name the boxes. Then they simply email the file back. Of course, you're on your own for the analysis in this case.

9.2.1 Analyses of Open Card-Sort Data

One way to analyze data from an open card sort is to create a matrix of the "perceived distances" (also called a dissimilarity matrix) among all pairs of cards in the study. For example, assume you conducted a card-sorting study using 10 fruits: apples, oranges, strawberries, bananas, peaches, plums, tomatoes, pears, grapes, and cherries. Assume one participant in the study created the following names and groupings:

- "Large, round fruits": apples, oranges, peaches, tomatoes
- "Small fruits": strawberries, grapes, cherries, plums
- "Funny-shaped fruits": bananas, pears

You can then create a matrix of "perceived distances" among all pairs of fruits for each participant using the following rules:

- If this person put a pair of cards in the same group, it gets a distance of 0.
- If this person put a pair of cards into different groups, it gets a distance of 1.

Using these rules, the distance matrix for the preceding participant would look like what's shown in Table 9.7.

	Apples	Oranges	Strawberries	Bananas	Peaches	Plums	Tomatoes	Pears	Grapes	Cherries
Apples	—	0	1	1	0	1	0	1	1	1
Oranges		—	1	1	0	1	0	1	1	1
Strawberries			—	1	1	0	1	1	0	0
Bananas				—	1	1	1	0	1	1
Peaches					—	1	0	1	1	1
Plums						——	1	1	0	0
Tomatoes							—	1	1	1
Pears								—	1	1
Grapes									—	0
Cherries										—

Table 9.7 Distance matrix for one participant in the fruit card-sorting example.

CARD-SORT ANALYSIS SPREADSHEETS

Donna Maurer has developed an Excel spreadsheet for the analysis of card-sorting data. She uses some very different techniques for exploring the results of a card-sorting exercise than the more statistical techniques we're describing here, including support for the person doing the analysis to standardize the categories by grouping the ones that are similar. The spreadsheet and instructions can be downloaded from http://www.rosenfeldmedia.com/books/cardsorting/blog/card_sort_analysis_spreadsheet/.

In addition, Mike Rice has developed a spreadsheet for creating a co-occurrence matrix from card-sorting data. This type of analysis allows you to see how often any two cards were sorted into the same group. His analysis spreadsheet works with the same spreadsheets that Donna Maurer uses for her analyses. Mike's analysis spreadsheet, and the instructions for using it, can be found at http://www.informoire.com/co-occurrence-matrix/.

We're only showing the top half of the matrix for simplicity, but the bottom half would be exactly the same. The diagonal entries are not meaningful because the distance of a card from itself is undefined. (Or it can be assumed to be zero if needed in the analyses.) So for any one participant in the study, the entries in this matrix will only be 0's or 1's. The key is to then combine these matrices for all the participants in the study. Let's assume you had 20 participants do the card-sorting exercise with the fruits. You can then sum the matrices for the 20 participants. This will create an overall distance matrix whose values can, in theory, range from 0 (if all participants put that pair into the same group) to 20 (if all participants put that pair into different groups). The higher the number, the

greater the distance. Table 9.8 shows an example of what that might look like. In this example, only 2 of the participants put the oranges and peaches in different groups, whereas all 20 of the participants put the bananas and tomatoes into different groups.

	Apples	Oranges	Strawberries	Bananas	Peaches	Plums	Tomatoes	Pears	Grapes	Cherries
Apples	—	5	11	16	4	10	12	8	11	10
Oranges		—	17	14	2	12	15	11	12	14
Strawberries			—	17	16	8	18	15	4	8
Bananas				—	17	15	20	11	14	16
Peaches					—	9	11	6	15	13
Plums						—	12	10	9	7
Tomatoes							—	16	18	14
Pears								—	12	14
Grapes									—	3
Cherries										—

Table 9.8 Overall distance matrix for 20 participants in the fruit card-sorting study.

This overall matrix can then be analyzed using any of several standard statistical methods for studying distance (or similarity) matrices. Two that we find useful are hierarchical cluster analysis (e.g., Aldenderfer & Blashfield, 1984) and multidimensional scaling (MDS)(e.g., Kruskal & Wish, 2006). Both are available in a variety of commercial statistical analysis packages, including SAS (http://www.sas.com), IBM SPSS (http://www.spss.com), and NCSS (http://www.ncss.com/), as well as some add-on packages for Excel (e.g., Unistat, http://www.unistat.com; XLStat, http://www.xlstat.com).

HIERARCHICAL CLUSTER ANALYSIS

The goal of hierarchical cluster analysis is to build a tree diagram where the cards that were viewed as most similar by the participants in the study are placed on branches that are close together. For example, Figure 9.4 shows the result of a hierarchical cluster analysis of the data in Table 9.8. The key to interpreting a hierarchical cluster analysis is to look at the point at which any given pair of cards "join together" in the tree diagram.

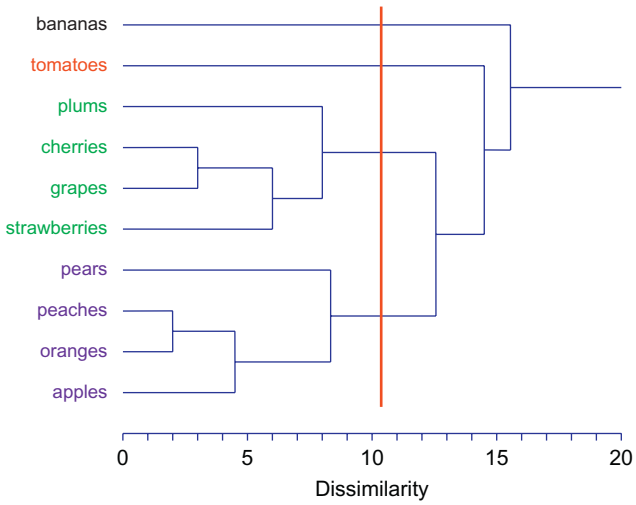


Figure 9.4 Result of a hierarchical cluster analysis of data shown in Table 9.8.

Cards that join together sooner are more similar to each other than those that join together later. For example, the pair of fruits with the lowest (shortest) distance in Table 9.8 (peaches and oranges; distance = 2) join together first in the tree diagram.

Several different algorithms can be used in hierarchical cluster analysis to determine how the "linkages" are created. Most of the commercial packages that support hierarchical cluster analysis let you choose which method to use. The linkage method we think works best is one called the Group Average method. But you might want to experiment with some of the other linkage methods to see what the results look like; there's no absolute rule saying one is better than another.

One thing that makes hierarchical cluster analysis so appealing for use in the analysis of card-sorting data is that you can use it to directly inform how you might organize the cards (pages) in a website. One way to do this is to take a vertical "slice" through the tree diagram and see what groupings that creates. For example, Figure 9.4 shows a four-cluster "slice": The vertical line intersects four horizontal lines, forming the four groups whose members are color coded. How do you decide how many clusters to create when taking a "slice" like this? Again, there's no fixed rule, but one method we like is to calculate the average number of groups of cards created by the participants in the card-sorting study and then try to approximate that.

After taking a "slice" through the tree diagram and identifying the groups created by that, the next thing you might want to do is determine how those groups compare to the original card-sorting data—in essence, to come up with a "goodness-of-fit" metric for your derived groups. One way of doing that is to compare the pairings of cards in your derived groups with the pairings created by each participant in the card-sorting study and to identify what percentage of the pairs match. For example, for the data in Table 9.7, only 7 of the 45 pairs do *not* match those identified in Figure 9.4. The 7 nonmatching pairings are apples–tomatoes, apples–pears, oranges–tomatoes, oranges–pears, bananas–pears, peaches–tomatoes, and peaches–pears. That means 38 pairings do match, or 84% (38/45). Averaging these matching percentages across all the participants will give you a measure of the goodness of fit for your derived groups relative to the original data.

MULTIDIMENSIONAL SCALING

Another way of analyzing and visualizing data from a card-sorting exercise is using multidimensional scaling (MDS). Perhaps the best way to understand MDS is through an analogy. Imagine that you had a table of the mileages between all pairs of major U.S. cities but not a map of where those cities are located. An MDS analysis could take that table of mileages and derive an approximation of the map showing where those cities are relative to each other. In essence, MDS tries to create a map in which the distances between all pairs of items match the distances in the original distance matrix as closely as possible.

The input to an MDS analysis is the same as the input to hierarchical cluster analysis—a distance matrix, like the example shown in Table 9.8. The result of an MDS

analysis of the data in Table 9.8 is shown in Figure 9.5. The first thing that's apparent from this MDS analysis is how the tomatoes and bananas are isolated from all the other fruit. That's consistent with the hierarchical cluster analysis, where those two fruits were the last two to join all the others. In fact, our four-cluster "slice" of the hierarchical cluster analysis (Figure 9.4) had these two fruits as groups unto themselves. Another thing apparent from the MDS analysis is how the strawberries, grapes, cherries, and plums cluster together on the left, and the apples, peaches, pears, and oranges cluster together on the right. That's also consistent with the hierarchical cluster analysis.

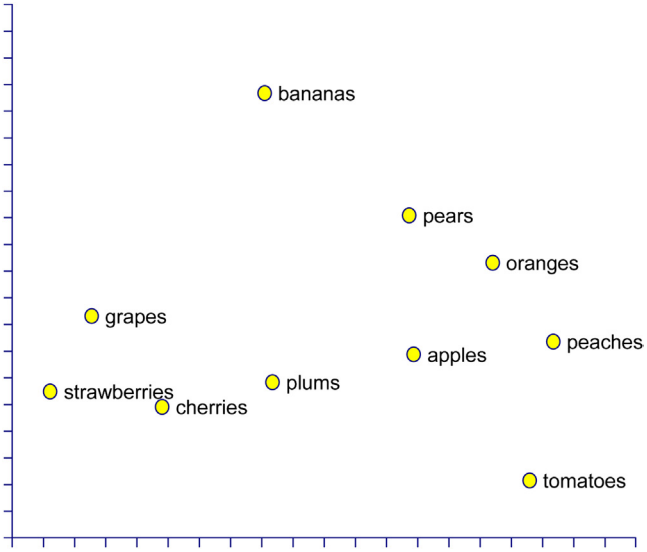
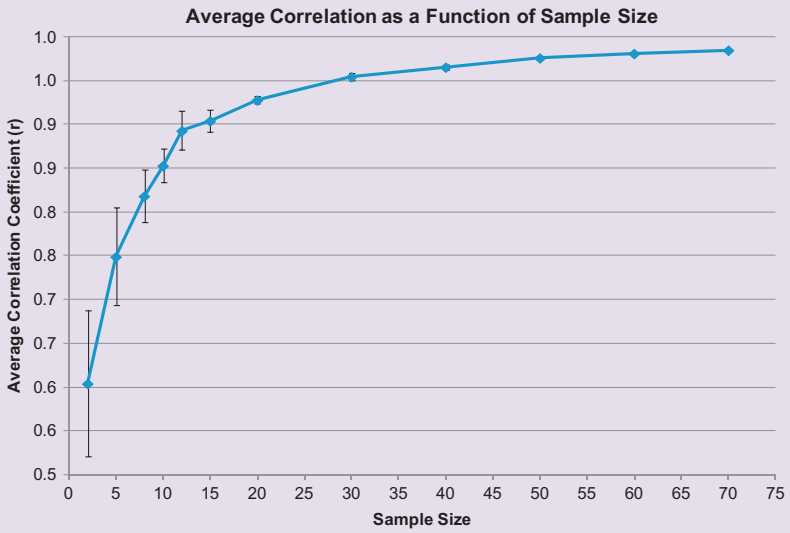


Figure 9.5 Multidimensional scaling analysis of the distance matrix in Table 9.8.

Note that it's also possible to use more than two dimensions in an MDS analysis, but we've rarely seen a case where adding

HOW MANY PARTICIPANTS ARE ENOUGH FOR A CARD-SORTING STUDY?

Tullis and Wood (2004) conducted a card-sorting study in which they addressed the question of how many people are needed for a card-sorting study if you want reliable results from your analyses. They did an open sort with 46 cards and 168 participants. They then analyzed the results for the full data set (168 participants), as well as many random subsamples of the data from 2 to 70 participants. Correlations of the results for those subsamples to the full data set looked like the chart here.



The "elbow" of that curve appears to be somewhere between 10 and 20, with a sample size of 15 yielding a correlation of 0.90 with the full data set. Although it's hard to know how well these results would generalize to other card-sorting studies with different subject matter or different numbers of cards, they at least suggest that about 15 may be a good target number of participants.

even just one more dimension yields particularly useful insights into card-sorting data. Another point to keep in mind is that the orientation of the axes in an MDS plot is arbitrary. You could rotate or flip the map any way you want and the results would still be the same. The only thing that's actually important is the relative distances between all pairs of the items.

The most common metric that's used to represent how well an MDS plot reflects the original data is a measure of "stress" that's sometimes referred to as *Phi*. Most of the commercial packages that do MDS analysis can also report the stress value associated with a solution. Basically, it's calculated by looking at all pairs of items, finding the difference between each pair's distance in the MDS map and its distance in the original matrix, squaring that difference, and summing those squares. That measure of stress for the MDS map shown in Figure 9.5 is 0.04. The smaller the value, the better. But how small does it really need to be? A good rule of thumb is that stress values under 0.10 are excellent, whereas stress values above 0.20 are poor.

We find that it's useful to do both a hierarchical cluster analysis and an MDS analysis. Sometimes you see interesting things in one that aren't apparent in the other. Because they are different statistical analysis techniques, you shouldn't expect them to give exactly the same answers. For example, one thing that's sometimes easier to see in an MDS map is which cards are "outliers"—those that don't obviously belong with a single group. There are at least two reasons why a card could be an outlier: (1) It could truly be an outlier—a function that really is different from all the others, or (2) it could have been "pulled" toward two or more groups. When designing a website, you would probably want to make these functions available from *each* of those clusters.

9.2.2 Analyses of Closed Card-Sort Data

Closed card sorts, where you not only give participants the cards but also the names of the groups in which to sort them, are probably done less often than open card sorts. Typically, you would start with an open sort to get an idea of the kinds of groups that users would naturally create and the names they might use for them. Sometimes it's helpful to follow up an open sort with one or more closed sorts, mainly as a way of testing your ideas about organizing the functions and naming the groupings. With a closed card sort you have an idea about how you want to organize the functions, and you want to see how close users come to matching the organization you have in mind.

We used closed card sorting to compare different ways of organizing the functions for a website (Tullis, 2007). We first conducted an open sort with 54 functions. We then used those results to generate six different ways of organizing the functions that we then tested in six simultaneous closed card-sorting exercises. Each closed card sort used the same 54 functions but presented different groups to sort the functions into. The number of groups in each "framework" (set of group names) ranged from three to nine. Each participant only saw and used one of the six frameworks.

In looking at the data from a closed card sort, the main thing you're interested in is how well the groups "pulled" the cards to them that you intend to belong to those groups. For example, consider the data in Table 9.9 which shows the percentage of participants in a closed card-sorting exercise who put each card into each of the groups.

The other percentage, shown on the right in Table 9.9, is the highest percentage for each card. This is an indicator of how well the "winning" group pulled the appropriate cards to it. What you hope to see are cases like Card #10 in this table, which was pulled very strongly to Group C, with 92% of the participants putting it in that group. Ones that are more troubling are cases such as Card #7, where 46% of the participants put it in Group A, but 37% put it in Group C—participants were very "split" in terms of deciding where that card belonged in this set of groups.

One metric you could use for characterizing how well a particular set of group names fared in a closed card sort is the average of these maximum values for all the

Card	Group A	Group B	Group C	Max
Card #1	17%	78%	5%	78%
Card #2	15%	77%	8%	77%
Card #3	20%	79%	1%	79%
Card #4	48%	40%	12%	48%
Card #5	11%	8%	81%	81%
Card #6	1%	3%	96%	96%
Card #7	46%	16%	37%	46%
Card #8	57%	38%	5%	57%
Card #9	20%	75%	5%	75%
Card #10	4%	5%	92%	92%
			Average:	73%

Table 9.9 Percentage of participants in a closed card sort who put each of 10 cards into each of the three groups provided.

cards. For the data in Table 9.9, that would be 73%. But what if you want to compare results from closed card sorts that had the same cards but different sets of groups? That average maximum percentage will work well for comparisons as long as each set contained the same number of groups. But if one set had only three groups and another had nine groups, as in the Tullis (2007) study, it's not a fair metric for comparison. If participants were simply acting randomly in doing the sorting with only three groups, by chance they would get a maximum percentage of 33%. But if they were acting randomly in doing a sort with nine groups, they would get a maximum percentage of only 11%. So using this metric, a framework with more groups is at a disadvantage in comparison to one with fewer groups.

We experimented with a variety of methods to correct for the number of groups in a closed card sort. The one that seems to work best is illustrated in Table 9.10. These are the same data as shown earlier in Table 9.9 but with two additional columns. The "2nd place" column gives the percentage associated with the group that had the next-highest percentage. The "Difference" column is simply the difference between the maximum percentage and the 2nd-place percentage. A card that was pulled strongly to one group, such as Card #10, gets a relatively small penalty in this scheme. But a card that was split more evenly, such as Card #7, takes quite a hit.

The average of these differences can then be used to make comparisons between frameworks that have different numbers of groups. For example, Figure 9.6 shows the data from Tullis (2007) plotted using this method. We call this a measure of the percent agreement among the participants about which group each card belongs to. Obviously, higher values are better.

Card	Category A	Category B	Category C	Max	2nd Place	Difference
Card #1	17%	78%	5%	78%	17%	61%
Card #2	15%	77%	8%	77%	15%	62%
Card #3	20%	79%	1%	79%	20%	60%
Card #4	48%	40%	12%	48%	40%	8%
Card #5	11%	8%	81%	81%	11%	70%
Card #6	1%	3%	96%	96%	3%	93%
Card #7	46%	16%	37%	46%	37%	8%
Card #8	57%	38%	5%	57%	38%	18%
Card #9	20%	75%	5%	75%	20%	55%
Card #10	4%	5%	92%	92%	5%	87%
			Average:	73%		52%

Table 9.10 Same data as shown in Table 9.9 but with an additional two columns[a].

[a] *"2nd place" refers to the next-highest percentage after the maximum percentage, and "Difference" indicates the difference between the maximum percentage and the 2nd-place percentage.*

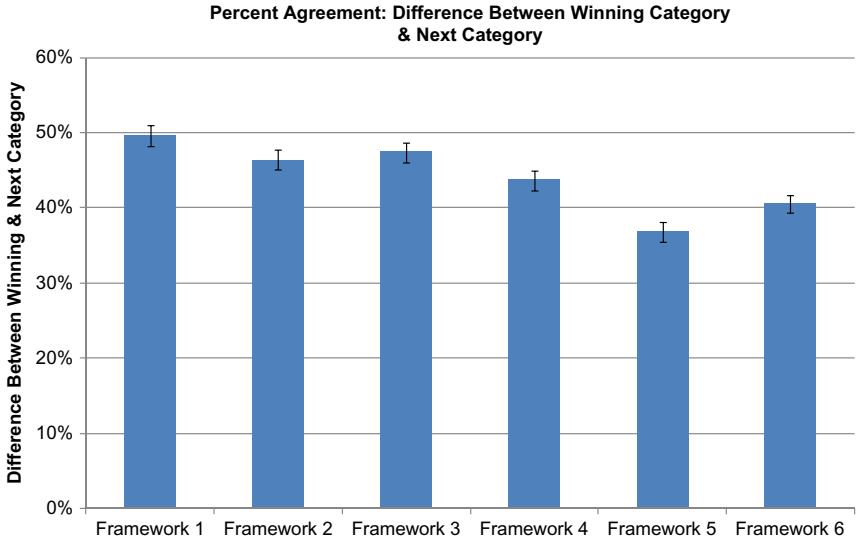


Figure 9.6 Comparison of six frameworks in six parallel closed card sorts. Because the frameworks had different numbers of groups, a correction was used in which the percentage associated with the 2nd-place group was subtracted from the winning group. Adapted from Tullis (2007); used with permission.

Data from a closed card sort can also be analyzed using hierarchical cluster analysis and MDS analysis, just like data from an open card sort. These give you visual representations of how well the framework you presented to the participants in the closed card sort actually worked for them.

9.2.3 Tree Testing

A technique that's closely related to closed card sorting is tree testing. This is a technique where you provide an interactive representation of the proposed information organization for a site, typically in the form of menus that let the user traverse the information hierarchy. For example, Figure 9.7 shows a sample study in Treejack (http://www.optimalworkshop.com/treejack.htm) from the participant's perspective.

Although the interface is very different, conceptually this is similar to a closed card-sorting exercise. In a tree test, each task is similar to a "card" in that the participants are telling you where they would expect to find that element in the tree structure.

Figure 9.8 shows an example of data for one task provided by Treejack, including the following:

- *Task success data*. You tell Treejack which nodes in the tree you consider to be successful for each task.
- *Directness*. This is the percentage of participants who didn't backtrack up the tree at any point during the task. This can be a useful indication of how confident the participants are in making their selections.
- *Time taken*. Average time taken by participants to complete the task.

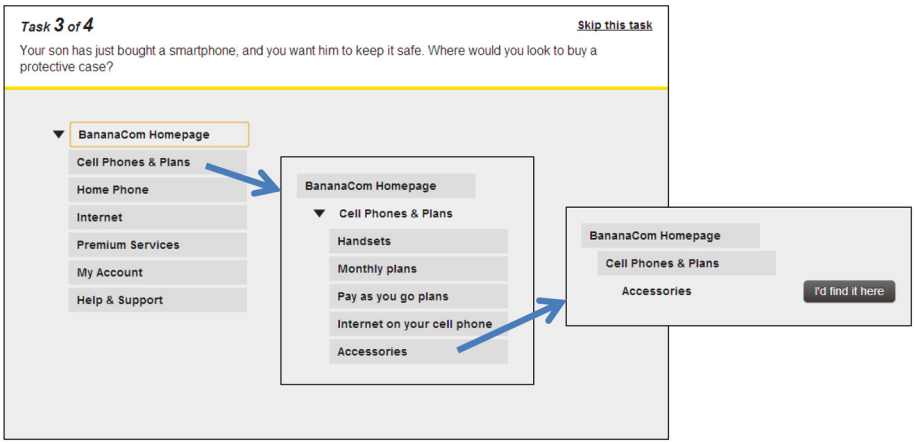


Figure 9.7 Sample study in Treejack. The task is shown at the top. Initially the participant sees only the menu on the left. After selecting "Cell Phones & Plans" from that menu, a submenu is shown. This continues until the participant chooses the "I'd find it here" button. The participant can go back up the tree at any time.

And all three of these metrics are shown with 95% confidence intervals!

Treejack also provides an interesting visualization of data for each task called a "PieTree," shown in Figure 9.9. In this visualization, the size of each node reflects the number of participants who visited that node for this task. Colors within each node reflect the percentage of participants who continued down a correct path, an incorrect path, or nominated a "leaf" node as the correct answer. In the online version of the PieTree, hover information for each node gives you more details about what the participants did at that node.

SOME TREE-TESTING TOOLS

The following are some of the tree-testing tools that we're aware of:

- C-Inspector (http://www.c-inspector.com)
- Optimal Workshop's Treejack (http://www.optimalworkshop.com/treejack.htm)
- PlainFrame (http://uxpunk.com/plainframe/)
- UserZoom Tree Testing (http://www.userzoom.com/products/tree-testing)

9.3 ACCESSIBILITY DATA

Accessibility usually refers to how effectively someone with disabilities can use a particular system, application, or website (e.g., Cunningham, 2012; Henry, 2007; Kirkpatrick et al., 2006). We believe that accessibility is really just usability for a particular set of users. When viewed that way, it becomes obvious that most of

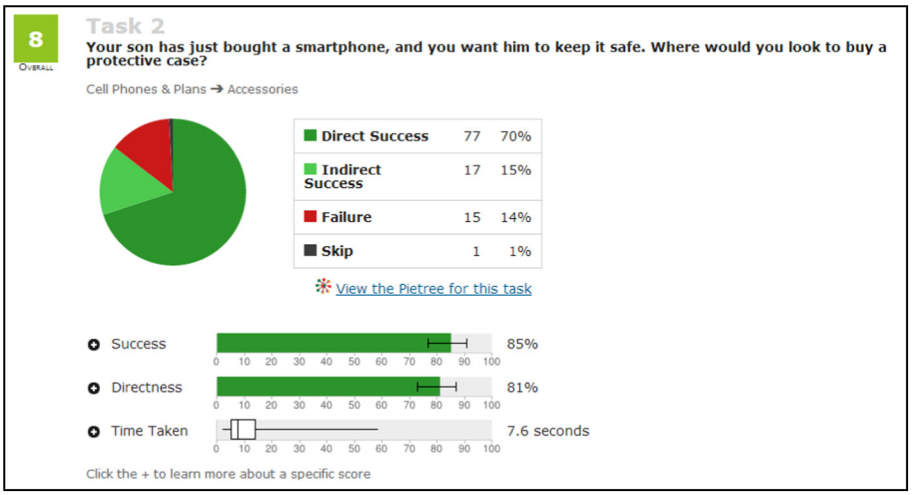


Figure 9.8 Sample data for one task in Treejack, including task success, directness, and time taken.

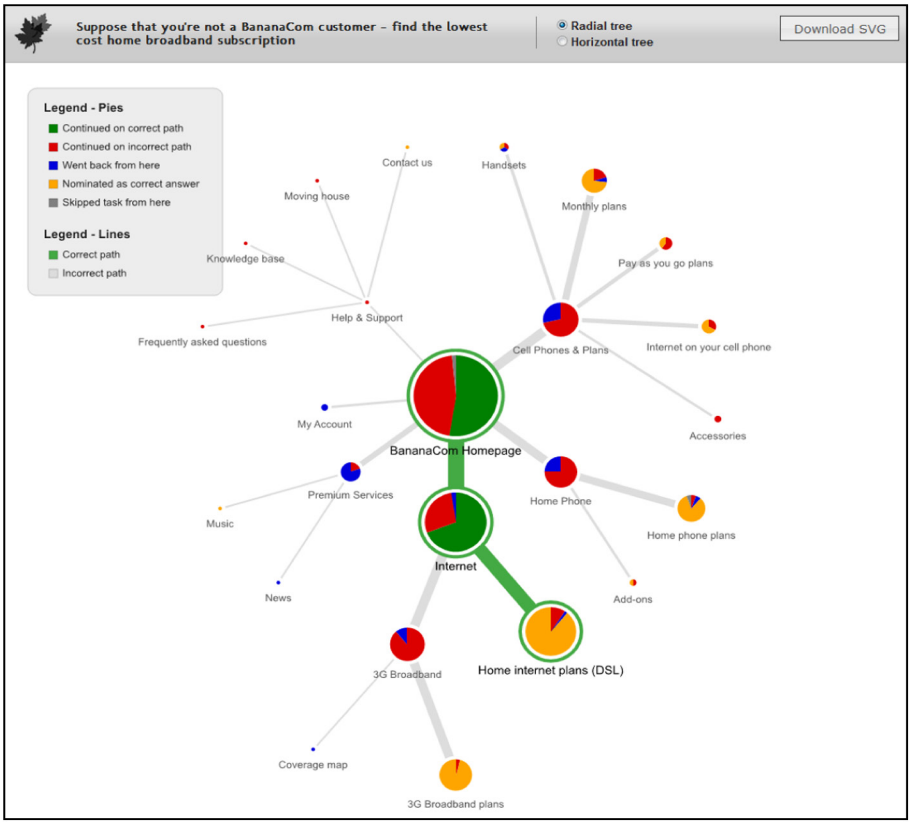


Figure 9.9 A "PieTree" from Treejack showing the paths that participants took in performing a single task, in this case indicating where they would expect to find information about lowest-cost home Internet plans. Green highlights the correct path (starting from the center).

the other metrics discussed in this book (e.g., task completion rates and times, self-reported metrics) can be applied to measure the usability of any system for users with different types of disabilities. For example, Nielsen (2001) reported four usability metrics from a study of 19 websites with three groups of users: blind users, who accessed the sites using screen-reading software; low-vision users, who accessed the sites using screen-magnifying software; and a control group who did not use assistive technology. Table 9.11 shows results for the four metrics.

	Screen Reader Users	Screen Magnifier Users	Control Group (No Disabilities)
Success rate	12.5%	21.4%	78.2%
Time on task	16:46	15:26	7:14
Errors	2.0	4.5	0.6
Subjective rating (1–7 scale)	2.5	2.9	4.6

Table 9.11 Data from usability tests of 19 websites with blind users, low-vision users, and users with normal vision[a].

[a]Adapted from Nielsen (2001); used with permission.

These results point out that the usability of these sites is far worse for the screen-reader and screen-magnifier users than it is for the control users. But the other important message is that the best way to measure the usability of a system or website for users with disabilities is to actually test with representative users. Although that's a very desirable objective, most designers and developers don't have the resources to test with representative users from all the disability groups that might want to use their product. That's where accessibility guidelines can be helpful.

Perhaps the most widely recognized web accessibility guidelines are the Web Content Accessibility Guidelines (WCAG), Version 2.0, from the World-Wide Web Consortium (W3C)(http://www.w3.org/TR/WCAG20/). These guidelines are divided into four major categories:

1. Perceivable
    a. Provide text alternatives for nontext content.
    b. Provide captions and other alternatives for multimedia.
    c. Create content that can be presented in different ways, including assistive technologies, without losing meaning.
    d. Make it easier for users to see and hear content.
2. Operable
    a. Make all functionality available from a keyboard.
    b. Give users enough time to read and use content.
    c. Do not use content that causes seizures.
    d. Help users navigate and find content.

3. Understandable
    a. Make text readable and understandable.
    b. Make content appear and operate in predictable ways.
    c. Help users avoid and correct mistakes.
4. Robust
    a. Maximize compatibility with current and future user tools.

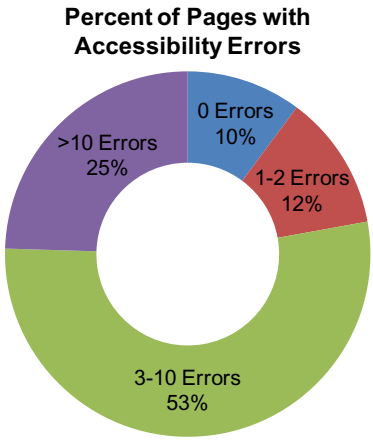


Figure 9.10 Results of analysis of a website against the WCAG guidelines.

One way of quantifying how well a website meets these criteria is to assess how many of the pages in the site fail one or more of each of these guidelines.

Some automated tools can check for certain obvious violations of these guidelines (e.g., missing "Alt" text on image). Although the errors they detect are generally true errors, they also commonly miss many errors. Many of the items that the automated tools flag as *warnings* may in fact be true errors, but it takes a human to find out. For example, if an image on a web page has null Alt text defined (ALT=""), that may be an error if the image is informational or it may be correct if the image is purely decorative. The bottom line is that the only really accurate way to determine whether accessibility guidelines have been met is by manual inspection of the code or by evaluation using a screen reader or other appropriate assistive technology. Often both techniques are needed.

AUTOMATED ACCESSIBILITY-CHECKING TOOLS

Some of the tools available for checking web pages for accessibility errors include the following:

- Cynthia Says (http://www.contentquality.com/)
- Accessibility Valet Demonstrator (http://valet.webthing.com/access/url.html)
- WebAIM's WAVE tool (http://wave.webaim.org/)
- University of Toronto Web Accessibility Checker (http://achecker.ca/checker/)
- TAW Web Accessibility Test (http://www.webdevstuff.com/103/taw-web-accessibility-test.html)

Once you've analyzed the pages against the accessibility criteria, one way of summarizing the results is to count the number of pages with errors. For example, Figure 9.10 shows results of a hypothetical analysis of a website against the WCAG guidelines. This shows that only 10% of the pages have no errors, whereas 25% have more than 10 errors. The majority (53%) have 3–10 errors.

In the United States, another important set of accessibility guidelines is the so-called Section 508 guidelines or, technically, the 1998 Amendment to Section 508 of the 1973 Rehabilitation Act (Section 508, 1998; also see Mueller, 2003).

This law requires federal agencies to make their electronic and information technology accessible to people with disabilities, including their websites. The law applies to all federal agencies when they develop, procure, maintain, or use electronic and information technology. Section 508 specifies 16 standards that websites must meet. The Section 508 requirements are essentially a subset of the full WCAG 2.0 guidelines. We believe the most useful metric for illustrating Section 508 compliance is a page-level metric, indicating whether the page passes all 16 standards or not. You can then chart the percentage of pages that pass versus those that fail.

UPDATE TO SECTION 508

At the time of the writing of this second edition, an update of Section 508 is expected shortly. A 2011 draft has been released, commented on, and public hearings have been held. The new version is much more complete and closely reflects WCAG 2.0. The latest information can be found at http://www.access-board.gov/508.htm.

9.4 RETURN-ON-INVESTMENT DATA

A book about usability metrics wouldn't be complete without at least some discussion of return on investment, as the usability metrics discussed in this book often play a key role in calculating ROI. But because entire books have been written on this topic (Bias & Mayhew, 2005; Mayhew & Bias, 1994), our purpose is to just introduce some of the concepts.

The basic idea behind usability ROI, of course, is to calculate the financial benefit attributable to usability enhancements for a product, system, or website. These financial benefits are usually derived from such measures as increased sales, increased productivity, or decreased support costs that can be attributed to the usability improvements. The key is to identify the cost associated with the usability improvements and then compare those to the financial benefits.

As Bias and Mayhew (2005) summarize, there are two major categories of ROI, with different types of returns for each:

- Internal ROI:
    - Increased user productivity
    - Decreased user errors
    - Decreased training costs
    - Savings gained from making changes earlier in the design life cycle
    - Decreased user support
- External ROI:
    - Increased sales
    - Decreased customer support costs

- Savings gained from making changes earlier in the design life cycle
- Reduced cost of providing training (if training is offered by the company)

To illustrate some of the issues and techniques in calculating usability ROI, we'll look at an example from Diamond Bullet Design (Withrow, Brinck, & Speredelozzi, 2000). This case study involved the redesign of a state government web portal. They conducted usability tests of the original website and a new version that had been created using a user-centered design process. The same 10 tasks were used to test both versions. A few of them were as follows:

- You are interested in renewing a {state} driver's license online.
- How do nurses get licensed in {the state}?
- To assist in traveling, you want to find a map of {state} highways.
- What 4-year colleges are located in {the state}?
- What is the state bird of {the state}?

Twenty residents of the state participated in the study, which was a between-subjects design (with half using the original site and half using the new). Data collected included task times, task completion rates, and various self-reported metrics. They found that the task times were significantly shorter for the redesigned site, and the task completion rates were significantly higher. Figure 9.11 shows task times for the original and redesigned sites. Table 9.12 shows a summary of the task completion rates and task times for both versions of the site, as well as an overall measure of efficiency for both (task completion rate per unit time).

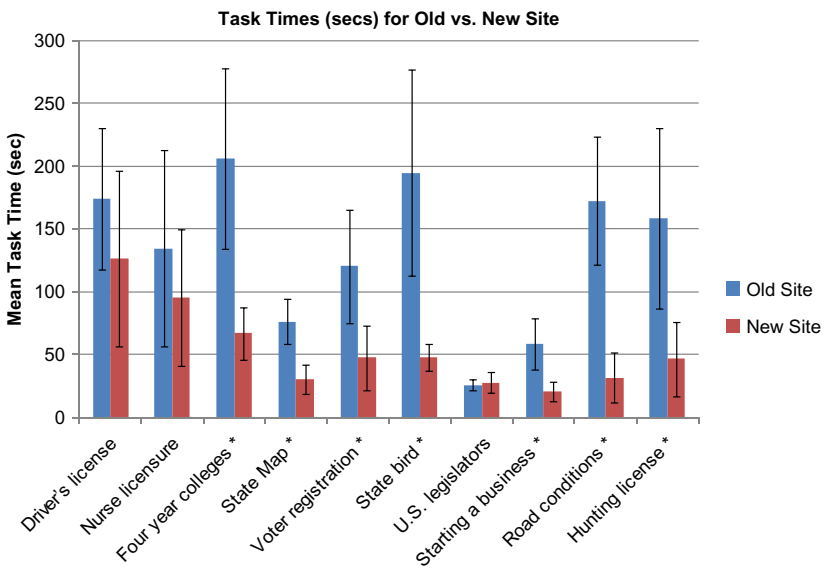


Figure 9.11 Task times for the original and the redesigned sites (an asterisk indicates significant difference). Adapted from Withrow et al. (2000); used with permission.

	Original Site	Redesigned Site
Average Task Completion Rate	72%	95%
Average Task Time (mins)	2.2	0.84
Average Efficiency	33%	113%

Table 9.12 Summary of task performance data[a].

[a]*Average efficiency is the task completion rate per unit of time (Task Completion Rate/Task Time). Adapted from Withrow et al. (2000); used with permission.*

So far, everything is very straightforward and simply illustrates some of the usability metrics discussed in this book. But here's where it gets interesting. To begin calculating ROI from the changes made to the site, Withrow et al. (2000) made the following assumptions and calculations related to *time savings*:

- Of the 2.7 million residents of the state, we might "conservatively estimate" a quarter of them use the website at least once per month.
- If each of them saved 79 seconds (as was the average task savings in this study), then about 53 million seconds (14,800 hours) are saved per year.
- Converting this to labor costs, we find 370 person-weeks (at 40 hours per week) or 7 person-years are saved per month; 84 person-years are saved each year.
- On average, a citizen in the target state had an annual salary of $14,700.
- This leads to a yearly benefit of *$1.2 million* based only on the time savings.

Note that this chain of reasoning had to start with a pretty big assumption: that a quarter of the residents of the state use the site at least once per month. So that assumption, which all the rest of the calculations hinge upon, is certainly up for debate. A better way of generating an appropriate value with which to start these calculations would have been from actual usage data for the current site.

They went on to calculate an increase in revenue due to the increased task completion rate for the new site:

1. The task failure rate of the old portal was found to be 28%, whereas the new site was 5%.
2. We might assume that 100,000 users would pay a service fee on the order of $2 per transaction at least once a month.
3. Then the 23% of them who are succeeding on the new site, whereas formerly they were failing, are generating an additional $552,000 in revenue per year.

Again, a critical assumption had to be made early in the chain of reasoning: that 100,000 users would pay a service fee to the state on the order of $2 per transaction at least once a month. A better way of doing this calculation would have been to use data from the live site specifically about the frequency of fee-generating transactions (and amounts of the fees). These could then have been adjusted to reflect the higher task completion rate for the redesigned site. If you

agree with their assumptions, these two sets of calculations yield a total of about $1.75 million annually, either in time savings to the residents or increased fees to the state. Although Withrow and colleagues (2000) don't specify how much was spent on the redesign of this portal, we can safely assume it was dramatically less than $1.75 million!

This example points out some of the challenges associated with calculating usability ROI. In general, there are two major classes of situations where you might try to calculate a usability ROI: when users of the product are employees of your company and when users of the product are your customers. It tends to be much more straightforward to calculate ROI when the users are employees of your company. You generally know how much employees are paid, so time savings in completing certain tasks (especially highly repetitive ones) can be translated directly to monetary savings. In addition, you may know the costs involved in correcting certain types of errors, so reductions in the rates of those errors could also be translated to monetary savings.

Calculating usability ROI tends to be much more challenging when the users are your customers (or really anyone not an employee of your company). Your benefits are much more indirect. For example, it might not make any real difference to your bottom line that your customers can complete a key income-generating transaction in 30% less time than before. It probably does *not* mean that they will then be performing significantly more of those transactions. But what it *might* mean is that over time those customers will remain your customers and others will become your customers who might not have otherwise (assuming the transaction times are significantly shorter than they are for your competitors), thus increasing revenue. A similar argument can be made for increased task completion rates.

SOME ROI CASE STUDIES

A variety of other case studies of usability ROI are available. Here's just a sampling.

- The Nielsen Norman Group did a detailed analysis of 72 usability ROI case studies and found increases in key performance indicators of 0% to over 6000%. The case studies covered a wide variety of websites, including Macy's, Bell Canada, New York Life, Open Table, a government agency, and a community college (Nielsen, Berger, Gilutz, & Whitenton, 2008).
- A redesign of the BreastCancer.org discussion forums resulted in a 117% increase in site visitors, a 41% increase in new memberships, a 53% reduction in time taken to register, and a 69% reduction in monthly help desk costs (Foraker, 2010).
- After a redesign of Move.com's home search and contact an agent features, users' ability to find a home increased from 62 to 98%, sales lead generation to real estate agents increased over 150%, and their ability to sell advertising space on the site improved significantly (Vividence, 2001).

- A user-centered redesign of Staples.com resulted in 67% more repeat customers and a 10% improvement in ratings of ease of placing orders, overall purchasing experience, and likelihood of purchasing again. Online revenues went from $94 million in 1999 to $512 million after implementation of the new site (Human Factors International, 2002).
- A major computer company spent $20,700 on usability work to improve the sign-on procedure in a system used by several thousand employees. The resulting productivity improvement saved the company $41,700 the first day the system was used (Bias & Mayhew, 1994).
- After a redesign of the navigational structure of Dell.com, revenue from online purchases went from $1 million per day in September 1998 to $34 million per day in March 2000 (Human Factors International, 2002).
- A user-centered redesign of a software product increased revenue by more than 80% over the initial release of the product (built without usability work). The revenues of the new system were 60% higher than projected, and many customers cited usability as a key factor in deciding to buy the new system (Wixon & Jones, 1992).

9.5 SUMMARY

Here are some of the key takeaways from this chapter.

1. If you're dealing with a live website, you should be studying what your users are doing on the site as much as you can. Don't just look at page views. Look at click-through rates and drop-off rates. Whenever possible, conduct live A/B tests to compare alternative designs (typically with small differences). Use appropriate statistics (e.g., χ^2) to make sure any differences you're seeing are statistically significant.
2. Card sorting can be immensely helpful in learning how to organize some information or an entire website. Consider starting with an open sort and then following up with one or more closed sorts. Hierarchical cluster analysis and multidimensional scaling are useful techniques for summarizing and presenting the results. Closed card sorts can be used to compare how well different information architectures work for the users. Tree-testing tools can also be a useful way to test a candidate organization.
3. Accessibility is just usability for a particular group of users. Whenever possible, try to include older users and users with various kinds of disabilities in your usability tests. In addition, you should evaluate your product against published accessibility guidelines or standards, such as WCAG or Section 508.
4. Calculating ROI data for usability work is sometimes challenging, but it usually can be done. If the users are employees of your company, it's generally easy to convert usability metrics such as reductions in task times into dollar savings. If the users are external customers, you generally have to extrapolate usability metrics such as improved task completion rates or improved overall satisfaction to decreases in support calls, increases in sales, or increases in customer loyalty.

CHAPTER 10

Case Studies

CONTENTS

Measuring the User Experience. DOI: http://dx.doi.org/10.1016/B978-0-12-415781-1.00010-8

This chapter presents five case studies showing how other UX researchers and practitioners have used metrics in their work. These case studies highlight the amazing breadth of products and UX metrics. We thank the authors of these case studies: Erin Bradner from Autodesk; Mary Theofanos, Yee-Yin Choong, and Brian Stanton from the National Institute of Standards and Technology (NIST); Tanya Payne, Grant Baldwin, and Tony Haverda from Open Text; Viki Stirling and Caroline Jarrett from Open University; and Amanda Davis, Elizabeth Rosenzweig, and Fiona Tranquada from Bentley University.

10.1 NET PROMOTER SCORES AND THE VALUE OF A GOOD USER EXPERIENCE

By Erin Bradner, Autodesk

Net Promoter is a measure of customer satisfaction that grew out of the Customer Loyalty research by Frederick Reichheld (2003). Reichheld developed

the Net Promoter Score (NPS) to simplify the characteristically long and cumbersome surveys that typified customer satisfaction research at that time. His research found a correlation between a company's revenue growth and their customers' willingness to recommend them. The procedure used to calculate the NPS is decidedly simple and is outlined here. In short, Reichheld argued that revenues grow as the percentage of customers willing to recommend a product or company increases actively relative to the percentage likely to recommend against it. (Note: Net Promoter is a registered trademark of Satmetrix, Bain and Reichheld.)

At Autodesk we've been using the Net Promoter method to analyze user satisfaction with our products for 2 years (Bradner, 2010). We chose Net Promoter as model for user satisfaction because we wanted more than an average satisfaction score. We wanted to understand how the overall ease-of-use and feature set of an established product factor into our customers' total product experience (Sauro & Kindlund, 2005). Through multivariate analysis—used frequently in conjunction with Net Promoter—we identified the experience attributes that inspire customers to promote our product actively. These attributes include the user experience of the software (ease of use), customer experience (phone calls to product support), and purchase experience (value for the price).

This case study explains the specific steps we followed to build this model of user satisfaction and outlines how we used it to quantify the value of a good user experience.

10.1.1 Methods

In 2010, we launched a survey aimed at measuring user satisfaction with the discoverability, ease of use, and relevance of a feature of our software we'll refer to here as the *L&T* feature. We used an 11-point scale and asked users' satisfaction with the feature, along with their likelihood to recommend the product. The recommend question is the question that is the defining feature of the Net Promoter model. To calculate the NPS, we:

1. Asked customers if they'd recommend our product using a scale from 0 to 10, where 10 means *extremely likely* and 0 means *extremely unlikely*.
2. Segmented the responses into three buckets:
   Promoters: Responses from 9 to 10
   Passives: Responses from 7 to 8
   Detractors: Responses from 0 to 6
3. Calculated the percentage of promoters and percentage of detractors.
4. Subtracted the percentage of detractors from the percentage of promoter responses to get the NPS.

This calculation gave us a NPS. Knowing that we had 40% more customers promoting than detracting our product does mean something. But it also begged the question: is 40% a good score?

Industry benchmarks do exist for NPS. For example, the consumer software industry (Sauro, 2011) has an average NPS of 21%—meaning a 20% is about

average for products such as Quicken, QuickBooks, Excel, Photoshop, and iTunes. Common practice at Autodesk is to place less stock in benchmarks but rather focus carefully on the aspects of the user experience that drive up the promoters while reducing the detractors.

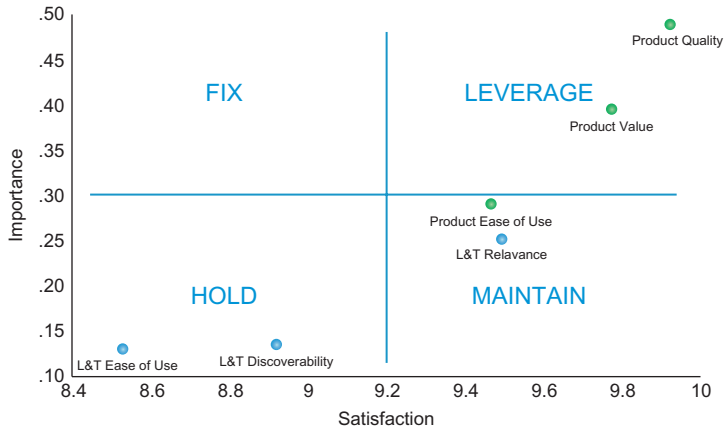


Figure 10.1 Anatomy of a Key Driver Analysis. Note that some graph data are simulated.

To isolate the "drivers" of a good user experience, we also included rating questions in our survey that asked about the overall product quality, product value, and product ease of use. We asked these questions on the same 11-point scale used for the recommendation question. We then calculated mean satisfaction scores for each experience variable. Satisfaction is plotted along the *x* axis shown in Figure 10.1.

Next we ran a multiple regression analysis with Net Promoter as the dependent variable and the attributes as independent variables. This analysis showed us which experience attributes were significant contributors to users' likelihood to recommend the product. Because it uses the beta coefficient, the analysis takes into account the correlation between each variable. Those correlations are plotted against the *y* axis in Figure 10.1. The *y* axis represents the standardized beta coefficient. We call the *y* axis "Importance" because correlation to the question "would you recommend this product?" is what tells us how important each experience variable is to our users. Plotting satisfaction against importance gives us insight into which experience attributes (interface, quality, or price) are most important to our users.

10.1.2 Results

According to Reichheld (2003), no one is going to recommend a product without really liking it. When we recommend something, especially in a professional setting, we put our reputations on the line. Recommending a product is admitting we are more than satisfied with the product. It signifies we are willing to do a little marketing and promotion on behalf of this product.

This altruistic, highly credible, and free promotion from enthusiastic customers is what makes the recommend question meaningful to measure. Promoters are going to actively encourage others to purchase our product and, according to Reichheld's research, are more likely to repurchase.

We wanted to determine how a customer's likelihood to recommend a given product was driven by specific features and by the overall ease of use of that

product. A new feature (we'll call L&T) was included in the product we were studying. When we plotted users' satisfaction with the L&T feature against their willingness to recommend the product containing the L&T feature, we found that the L&T feature was lower on the *y* axis relative to other aspects of the interface (as shown in Figure 10.1). Using the L&T feature (L&T Ease of Use) and locating it (L&T Discoverability) scored lower in satisfaction than Product Quality, Product Value, and Product Ease of Use, but they also scored lower in Importance. Users place less importance on this new feature relative to quality, value, and ease of use. Data show that users' satisfaction with the L&T feature is not as strongly correlated as quality and ease of use to their likelihood to recommend the product and is therefore not as important to driving growth of product sales.

The labels on the quadrants in Figure 10.1 tell us exactly which aspects of the user experience to improve next. Features that plot in the upper left quadrant, labeled FIX, are the highest priority because they have the highest importance and lowest satisfaction.

Data in Figure 10.1 indicate that if we were to redesign the L&T feature, we should invest in L&T Relevance as it plotted higher on the Importance axis than L&T Discoverability and Ease of Use. Discoverability and ease of use of the L&T feature are in the HOLD quadrant, indicating that these should be prioritized last.

10.1.3 Prioritizing Investments in Interface Design

So how much does the user interface of a software product contribute to users' willingness to recommend the product? We had been told by our peers in the business intelligence department that the strongest predictors to a user's willingness to recommend a product are:

1. Helpful and responsive customer support (Support)
2. Useful functionality at a good price (Value).

We ran a multiple-regression on our survey data set (Figure 10.2) and found that variables for the software user experience contribute 36% to the likelihood to recommend ($n = 2170$). Product Value accounted for 13% and Support accounted for another 9%. To verify the contribution of software user experience to willingness to recommend, we ran another multiple regression on data from a second, similar survey ($n = 1061$) and found the contribution of user experience variables to be 40%.

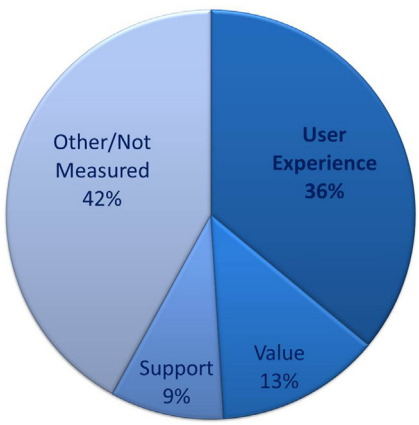


Figure 10.2 Simulated analysis of aspects of the customer experience contributing to customers' likelihood to recommend a product. Note that some graph data are simulated.

We then ran a third survey 1 year later. Regression formulas from the first survey and the third survey are shown, where LTR represents Likelihood to Recommend. In Year 1 we calculated the improvement targets shown in Figure 10.3 (left). We set a target of 5% increase in users' *likelihood to recommend* our product and we knew how to achieve that increase from

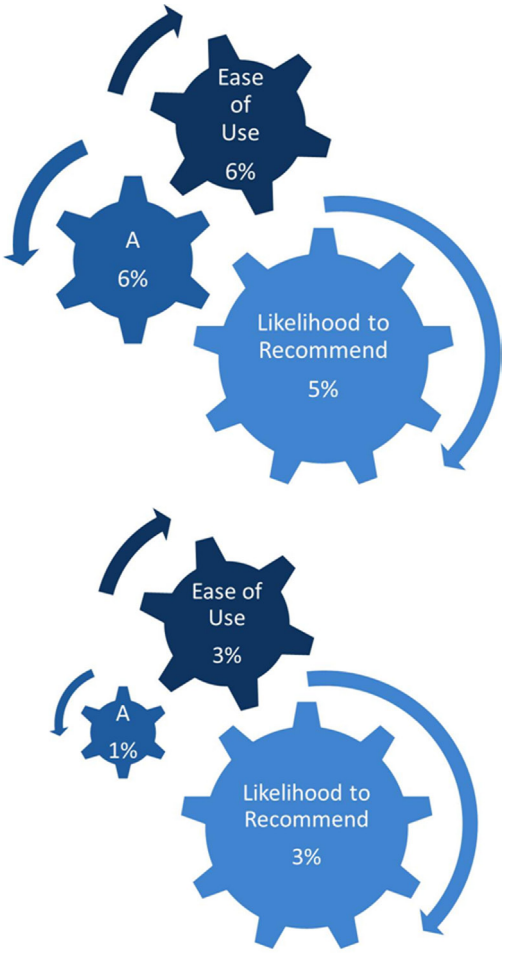


Figure 10.3 Target Increase in Likelihood to Recommend (top) vs Actual Increase (bottom).

the regression formula: assuming that the other contributing factors remain constant, if we could increase the satisfaction scores for the overall product ease of use, for the usability of Feature 1 and for the usability of Feature 2, then we would see an increase in users' Likelihood to Recommend of 5%.

In Year 2, we reran the analysis. We found that the actual increase in Likelihood to Recommend was 3%. This 3% increase was driven by a 3% increase in ease of use, a 1% increase in Feature 1's usability, and a 0% increase in Feature 2's usability, as summarized by Figure 10.3 (right). Regression formulas for the product studied are shown here:

$$\text{Year 1} - \text{Product} \times \text{LTR} = 2.8 + 0.39(\text{Ease of Use}) + 0.13(\text{Feature 1}) + 0.19(\text{Feature 2})(R^2 = 37\%)$$

$$\text{Year 2} - \text{Product} \times \text{LTR} = 2.8 + 0.39(\text{Ease of Use}) + 0.11(\text{Feature 1}) + 0.24(\text{Feature 2})(R^2 = 36\%)$$

10.1.4 Discussion

Thus, we found that running the multivariate analysis showed that the user experience contributed 36% to increasing product recommendations. At Year 2, we hadn't met our target of increasing Likelihood to Recommend our product by 5%, but by investing in ease of use and in a few key features we were able to improve the Likelihood to Recommend by 3%. The Net Promoter model had provided us with a way to define and prioritize investment in user experience design and had given us a way to track the return of that investment year after year.

We wanted to test the Net Promoter model further. Could the model be used as a predictor of sales growth, as it was originally intended (Reichheld, 2003)? We know the average sales price of our products. We know, from the multivariate analysis, that interface design contributes 36% to motivating users to recommend our product. If we knew how many promoters refer the product actively, we could estimate the revenue gains associated with improved user experience of our software.

What we did next is determine if there is a link between "promoters" and an increase in customer referrals. In our survey, we asked if the respondent—all were existing customers—had referred the product to a friend in the last year (Owen & Brooks, 2008). From these data we derived the proportion of

customers obtained through referrals and who likely refer others. This allowed us to approximate the number of referrals necessary to acquire one new customer (see Figure 10.4). Data used to derive this number are proprietary. For the purpose of this chapter, we use the number eight: we need eight referrals to acquire one new customer. In the NPS model, it is *promoters* who refer a product actively. But we didn't want to assume that every respondent who answered 9 or 10 to the *likelihood to recommend* question, that is, every promoter, had referred our product actively. The actual percentage of promoters who referred our product actively within the last year was 63%. From this, we derived that the total number of promoters needed to acquire one new customer was 13.

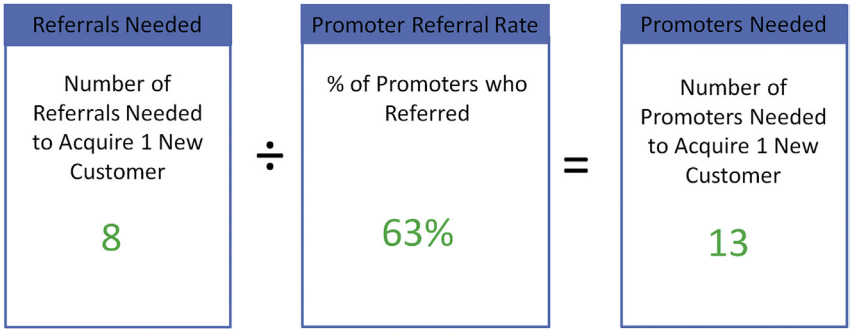


Figure 10.4 How many promoters are necessary to acquire one new customer?

10.1.5 Conclusion

By calculating the number of promoters required to acquire a new customer, we were able to connect the proverbial dots in the software business: a good user experience design drives our users to recommend our products and product recommendations increase customer acquisition, which increases revenue growth. Through multivariate analysis, we have shown that experience design contributes 36% to motivating users to recommend our product. Since we know the average sales price of our product, we were able to estimate the revenue gains associated with improving the user experience of our software. We quantified the value of a good user experience. By tying user experience to customer acquisition, we are able to prioritize design investment in ease of use and in research to improve the user experience of our products.

In summary, this case study shows:

- Multivariate analysis of user experience attributes can be used to prioritize investment in user experience design and research.
- User experience attributes, such as ease of use, contribute significantly to customer loyalty.
- Knowing the average sales price of our products and the number of promoters needed to acquire one new customer, we can quantify the return on investment of a good user experience.

At Autodesk, we've found that calculating a net promoter score isn't as useful as graphing and using key driver charts. Key driver charts target the aspects of the user experience that are needed most urgently in design improvements. By calculating drivers from year to year, we see how our investments in key areas pay out by increasing our users' likelihood to recommend our products. We watch a features move from the FIX quadrant safely into the LEVERAGE quadrant. Inspiring more customers to promote our product through designing excellent user experiences is what motivates us. It's not about a *score* or solely about acquiring new customers, it's about designing software experiences that are so good, our users will promote them actively.

REFERENCES

Bradner, E. (2010,). *Recommending net promoter*. Retrieved on 23.10.2011 from DUX: Designing the User Experience at Autodesk <http://dux.typepad.com/dux/2010/11/recommending-net-promoter.html>.

Reichheld, F. (2003). The one number you need to grow. *Harvard Business Review*.

Owen, R., & Brooks, L. (2008). *Answering the ultimate question*. San Francisco: Jossey-Bass.

Sauro, J. (2011). *Usability and net promoter benchmarks for consumer software*. Retrieved on 23.10.2011 from Measuring Usability <http://www.measuringusability.com/software-benchmarks.php>.

Sauro, J., & Kindlund, E. (2005). Using a single usability metric (SUM) to compare the usability of competing products. In *Proceeding of the human computer interaction international conference (HCII 2005)*, Las Vegas, NV.

Biography

Erin Bradner works for Autodesk, Inc.—makers of AutoCAD and a world leader in 3D design software for manufacturing, building, engineering, and entertainment. Erin manages user experience research across several of Autodesk's engineering and design products. She actively researches topics ranging from the future of computer-aided design, to how best to integrate marking menus into AutoCAD, to the contribution of user experience to likelihood to recommend a product. Erin has a Ph.D. in Human–Computer Interaction and 15 years of experience using both quantitative and qualitative research methods. Prior to Autodesk, Erin consulted for IBM, Boeing, and AT&T.

10.2 MEASURING THE EFFECT OF FEEDBACK ON FINGERPRINT CAPTURE

By Mary Theofanos, Yee-Yin Choong, and Brian Stanton, National Institute of Standards and Technology

The National Institute of Standards and Technology's Biometrics Usability Group is studying how to provide real-time feedback to fingerprint users in order to improve biometric capture at U.S. ports of entry. Currently, the U.S. Visitor and Immigrant Status Indicator Technology (US-VISIT) program collects fingerprints from all foreign visitors entering the United States using an

operator-assisted process. US-VISIT is considering unassisted biometric capture for specific applications. But ensuring acceptable quality images requires that users receive real-time feedback on performance. Many factors influence image quality, including fingerprint positioning, alignment, and pressure. What form this informative feedback should take is the challenge for an international audience.

To address this need, Guan and colleagues (2011) designed an innovative, cost-efficient, real-time algorithm for fingertip detection, slap/thumb rotation detection, and finger region intensity estimation that feeds rich information back to the user instantaneously during the acquisition process by measuring objective parameters of the image. This study investigates whether such rich, real-time feedback is enough to enable people to capture their own fingerprints without the assistance of an operator. A second objective is to investigate if providing an overlay guide will help people in better positioning their hands for fingerprint self-capture.

10.2.1 Methodology

EXPERIMENTAL DESIGN

We used a within-subject, single factor design with 80 participants who performed two fingerprint self-capture tasks: one with a fingerprint overlay displayed on a monitor to guide them on positioning their fingers during the capture process and the other task without the overlay. Order of receiving conditions was reversed for half the participants. The dependent variables are:

- *Task completion rate*—ratio of the number of participants who completed the self-capture task versus the number who did not complete or completed the task with assistance
- *Errors*—number of hand-positioning corrections until an acceptable fingerprint image is recorded
- *Quality of the fingerprint image*—the NIST fingerprint imaging quality (NFIQ) scores of the fingerprint images
- *Attempt time*—time from the moment a participant presents her hand until the end of the capture
- *Task completion time*—total time it takes to complete a capture task
- *User satisfaction*—user's ratings from the post-task questionnaire

PARTICIPANTS

Eighty adults [36 females and 44 males; ages ranging from 22 to 77(mean = 46.5)] were recruited from the general population (Washington, DC, area). Participants were distributed diversely across education, occupation, and ethnicity. Fifty-four participants indicated that they had been fingerprinted before: 18 had prior experience with inked and rolled fingerprinting and 36 did not indicate the type of their fingerprint experience. All fingerprint experiences were assisted.

Figure 10.5 Experimental setup.

MATERIALS[1]

We used a CrossMatch Guardian Fingerprint scanner as used by US-VISIT. Specifications include 500 ppi resolution, effective scanning area 3.2″ × 3.0″ (81 × 76 mm), single prism, single imager, uniform capture area. The system runs on an Intel core 2 CPU 4300 @1.8-GHz processor PC, with 3.23 GB RAM and a 20-inch LCD monitor.

Figure 10.5 shows the experiment configuration: scanner on a height-adjustable table with height set to 39 inches (common counter height at US-VISIT facilities). The scanner was placed at the recommended 20° angle (Theofanos et al., 2008). A webcam was mounted on the ceiling above the scanner to record participants' hand movements.

PROCEDURE

Each participant was instructed to perform two self-capture tasks using on-screen instructions. Participants were informed verbally that both tasks required them to capture four fingerprint images following the same sequence: right slap (RS), right thumb (RT), left slap (LS), and left thumb (LT).

The test scenario is described in Figure 10.6: task 1 includes the overlay and task 2 includes the nonoverlay. Half of the participants were assigned randomly to start without an overlay, followed by a task with the overlay. The other half of the participants received the reverse.

When the participant was ready, a generic fingerprint capture symbol as in Figure 10.7 was displayed, marking the start of the process.

Participants filled out a post-task questionnaire and discussed their overall impressions with the test administrator.

RESULTS

Applying the ISO (1998) definition of usability—"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"—we measured effectiveness, efficiency, and user satisfaction. The α of all tests for statistical significance was set to 0.05. Data were not distributed normally; thus, a nonparametric test of difference, the Wilcoxon matched-pairs signed-ranks test, was used on all statistical within-subject comparisons.

[1]Specific products and/or technologies are identified solely to describe the experimental procedures accurately. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

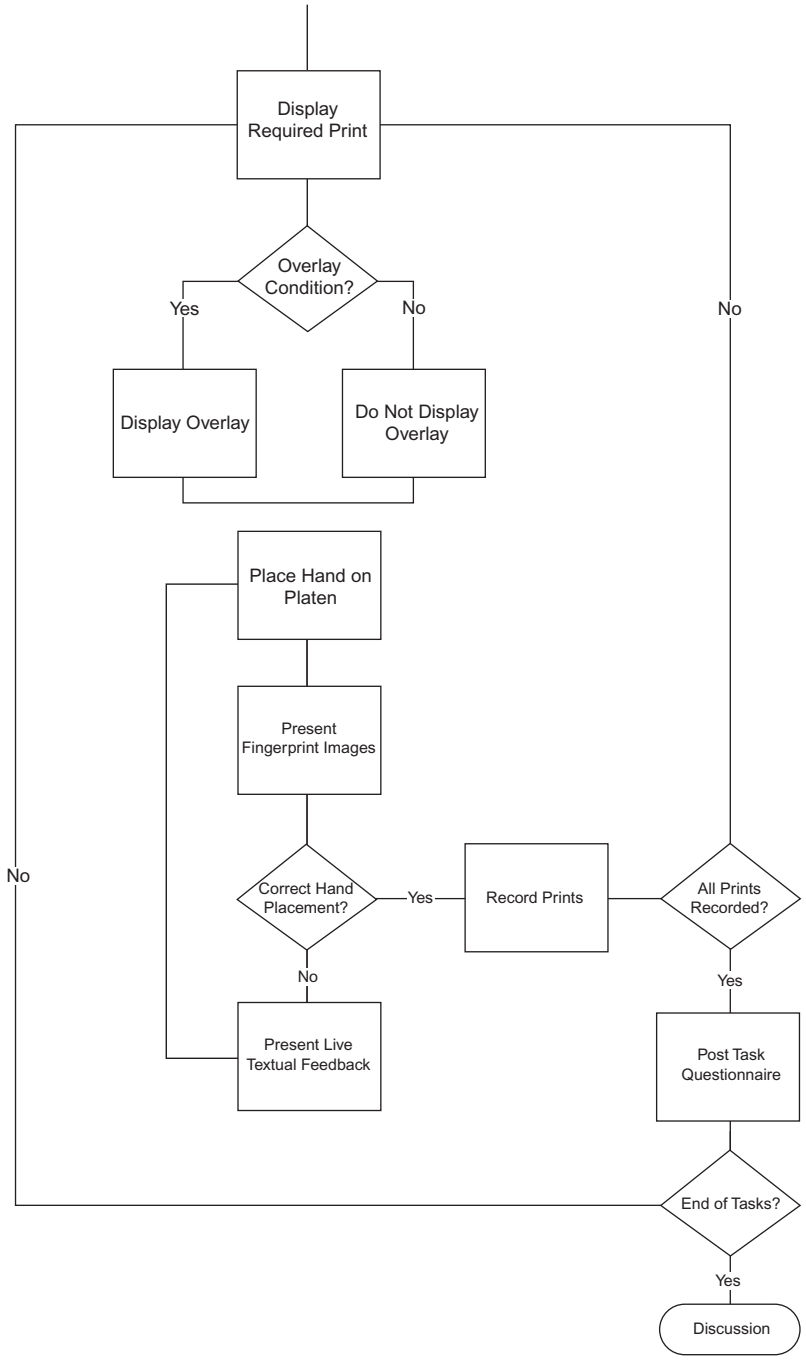
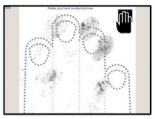


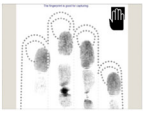
Figure 10.6 Test scenario.

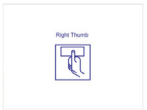

(a) Start of the capture process


(b) RS sym bol


(c) Scanning area with RS overlay

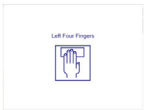

(d) Correction example – "Rotate counterclockwise"


(e) Example of an acceptable RS

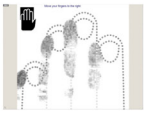

(f) RT symbol


(g) Scanning area with RT overlay


(h) Correction example – "Move up"


(i) LS symbol


(j) Scanning area with LS overlay


(k) Correction example – "Move right"


(l) LT symbol


(m) Scanning area with LT overlay


(n) Correction example – "Press more"

(o) End of the capture process

Figure 10.7 Overlay condition—fingerprint self-capture process with examples.[2]

EFFECTIVENESS

Three dependent variables are related to the effectiveness of the self-capture system: number of participants that completed the tasks (task completion rate), errors, and quality of the fingerprint image.

TASK COMPLETION RATE

Overall, 37 out of 40 (92.5%) participants for Task 1 and 39 out of 39 (100%) participants for Task 2 completed the self-capture tasks successfully by following the on-screen instructions without assistance or prompts from the test administrator.

ERRORS

The average number of hand-positioning corrections (errors) for each fingerprint is shown in Table 10.1. As shown in Table 10.2, errors can be classified into four categories.

Figure 10.8 shows the seven most common errors. For slaps (both right and left hands), not enough pressure applied was the most common error: RS: 214, LS: 101. There were many occurrences where not all four fingers were detected at the same time, indicating that participants had some difficulties placing their fingers evenly on the scanner: RS: 132, LS: 51.

QUALITY OF THE FINGERPRINT IMAGE

We used NIST fingerprint imaging software to compute the NFIQ (Tabassi et al., 2004) score for each finger. NFIQ scores range from 1 (highest quality) to 5 (lowest quality). The medians of individual NFIQ scores are shown in Figure 10.9.

As there is not yet consensus in the biometrics community on how to determine the quality of a slap image, we used a proposed quality scoring method under consideration by US-VISIT to assess the overall quality of the images. A slap is accepted if the index finger and middle finger have an NFIQ value of 1 or 2 and the ring finger and little finger have an NFIQ score of 1, 2, or 3. A thumb is accepted if it has an NFIQ value of 1 or 2. The results of applying the criteria

[2]The fingerprint images were blurred to prevent possible identification of the participants.

	Mean	STD	Min	Max
Right slap (RS)	6.4	6.689	0	31
Right thumb (RT)	0.663	1.158	0	6
Left slap (LS)	2.911	3.689	0	18
Left thumb (LT)	0.55	1.993	0	14
Total errors	10.50	9.69	0	49

Table 10.1 Errors by condition.

Error Category	Error Condition	Error Text Example
Pressure or angle that affects the contrast of an image	More finger area needs to be in contact with the scanner	"Press more" "Lower the angle of your fingers"
Detection of fingers	Not all fingers are detected	"Not all four fingers are detected" "Fingers are spread too wide. Please place all four fingers inside the box"
Movement of fingers	Vertical and horizontal misalignment	"Move your hand to the left" "Move your hand up"
Rotation of fingers	Fingers are not positioned upright	"Rotate your hand clockwise" "Rotate your thumb counterclockwise" "The direction of your hand is not correct. Please place four fingers upright in the center"

Table 10.2 Error category, condition, and text.

are also shown in Figure 10.9. The acceptance rates are RT, 78.8%; RS, 67.5%; LT, 76.3%; and LS, 68.4%.

EFFICIENCY

Two dependent variables are related to the efficiency of the self-capture system: attempt time and task completion time (Table 10.3).

ATTEMPT TIME

Attempt time is the time it takes from the moment a participant presents her hand until the end of capture of a fingerprint image. As expected, it took longer to capture slaps than thumbs. Means of the attempt time in seconds are RS, 28.66; RT, 5.87; LS, 17.14; and LT, 5.54.

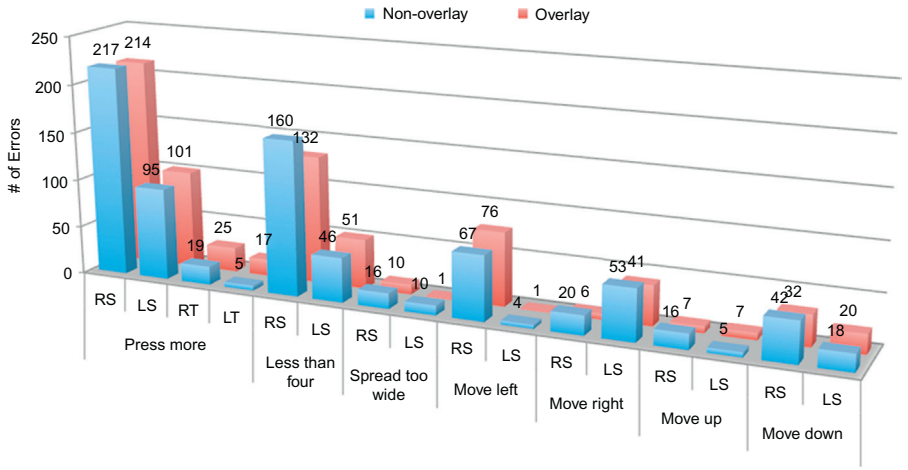


Figure 10.8 Most common correction errors.

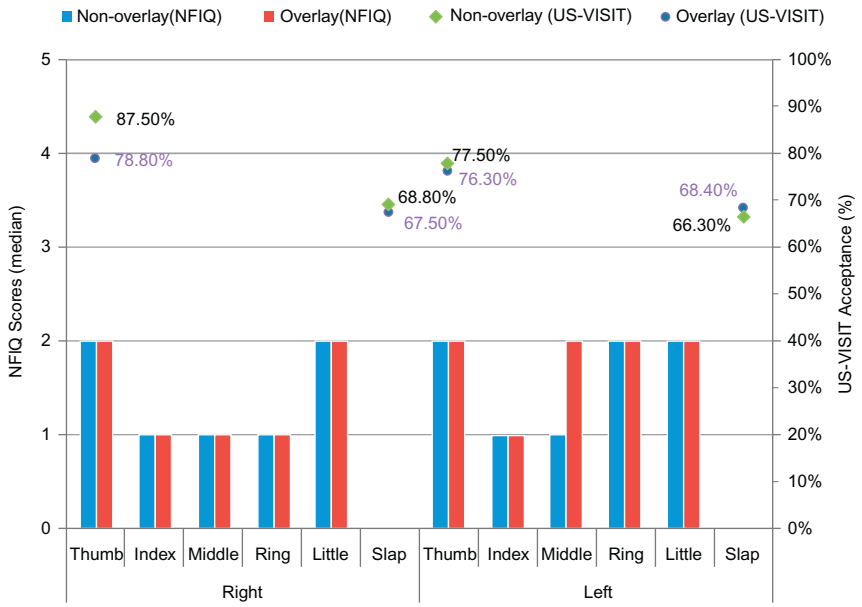


Figure 10.9 Fingerprint image quality.

TASK COMPLETION TIME

Task completion time is the total time spent to complete a capture task of four fingerprint images. As in Table 10.3, on average, it took participants approximately 1½ minutes to complete a self-capture task, including four fingerprint images.

	Time (sec)			
	Mean	Mean	Min	Max
RS (attempt time)	28.66	28.66	0.05	153.80
RT (attempt time)	5.87	5.87	2.23	28.14
LS (attempt time)	17.14	17.14	2.56	113.13
LT (attempt time)	5.54	5.54	2.172	55.02
Task completion time	94.71	94.71	32.61	270.47

Table 10.3 Time by conditions.

SATISFACTION

Participants filled out a questionnaire regarding their experience with the self-capture task after each task. The questionnaire consisted of six questions with a five-point semantic distance scale. Mean ratings across all participants are summarized in Table 10.4. Overall, the participants responded very positively to the self-capture tasks.

	Questions	Mean Ratings	Scales
Q1	How comfortable were you with the interaction with the fingerprint device?	4.24	1–uncomfortable 5–very comfortable
Q2	How did the time it took to have your fingerprint recorded compare with what you expected?	3.32	1–much more than expected 5–a lot less than expected
Q3	How would you rate the difficulty in positioning yourself so that your fingerprint could be recorded?	1.59	1–not difficult 5–very difficult
Q4	It was clear when the fingerprint process began.	3.51	1–unclear 5–intuitive
Q5	It was clear when the fingerprint process ended.	3.94	1–unclear 5–intuitive
Q6	How confident are you that you completed the fingerprint task as intended?	4.24	1–not confident 5–certain

Table 10.4 Post-task satisfaction questions and ratings.

During the discussion after the test, we asked each participant if the overlay assisted the self-capture process. Forty-six participants (57.5%) found the overlay helpful in guiding them to better position their hand on the scanner; six of

those participants indicated that it would be more helpful if the overlay were directly on the scanner (rather than projected on the monitor). Twenty-eight participants (35%) did not find the overlay helpful; nine of those participants indicated that the overlay would be helpful if it were directly on the scanner. Six participants (7.5%) indicated that the overlay was not a factor, as the process is quite simple and straightforward; one participant from this group indicated that it would be more helpful if the overlay were on the scanner.

10.2.2 Discussion

In this study, we examined whether people can perform fingerprint self-captures successfully using the proposed real-time feedback system. We found that the participants were very effective and efficient in performing the self-capture tasks with great satisfaction.

The participant completion rate was high (92.5% in Task 1, then improved to 100%) with fewer than 11 errors on average. When examining only positioning errors (errors related to pressure or angles were excluded), the average errors dropped (mean = 5.94). Fingerprint image quality was comparable, if not better, to images taken in the attended situation. In a study with an attended setup, Stanton et al. (2012) reported that the acceptance rates of slaps ranged from 55 to 63%, based on the US-VISIT acceptance criteria. In this self-capture study, acceptance rates ranged from 67.5 to 68.4% for slaps, higher than those in Stanton et al. (2012).

Using the on-screen instructions, participants were able to position their hands accordingly, make adjustments when needed, and capture fingerprint images in approximately 1½ minutes. Ratings on the post-task questionnaire indicate that the participants felt comfortable and confident and interacted without much difficulty with the self-capture process. It was clear to the participants when the capture process began and ended. In debriefing, participants indicated that the self-capture process was easy, straightforward, and quick. They praised the experience of "do-it-yourself" as it gave them a sense of being in control and being trusted; as one participant put it: "The self capture process was very neat. It is easy enough that anybody can do it. It is elementary, easy to use, even children can do it."

Our second research question was to investigate whether an overlay facilitates the self-capture fingerprint process. The overlay condition did not show consistent advantages or disadvantages of performance, that is, time, errors, and image quality, over the nonoverlay condition. However, more participants (57.5%) perceived that having the overlay helped them with the positioning of their hands and providing visual feedback. One reason for the discrepancy between performance and preference is the experimental configuration. The setup required participants to place their hands on the scanner (often looking down) and look up at the LCD monitor for the fingerprint image and feedback for corrections if needed. The overlay guide was superimposed onto the screen, which added another level of hand–eye coordination. Participants realized

corrections were needed according to the visual feedback of their fingerprint image in relation to the overlay on the screen, but had to move their hands to make the actual corrections on the scanner. Participants were observed being more careful in placing their hands on the scanner in the overlay condition as if they wanted to make sure their hands were aligned properly with the overlay, whereas participants positioned their hands more freely in the nonoverlay condition. Sixteen participants indicated that the overlay guide would be very helpful if it were placed on the scanner in order to align their hand on the overlay as they placed their hand on the scanner instead of looking up and down trying to make a perfect alignment.

We observed that participants learned to use the system very quickly. Within-subject comparisons were performed to examine the ease of learning of the system. When comparing the performance between Task 1 and Task 2, it was found that Task 2 performed significantly better on RS (attempt time and errors), RT (errors), task completion time, and total errors. Learning was even evident within a task.

10.2.3 Conclusion

The real-time feedback fingerprint system is a highly usable system and shows evidence of great potential for fingerprint self-captures. By following the on-screen, real-time instructions, participants quickly learned and felt comfortable and confident in capturing their own fingerprints without any assistance. The next step is to determine if users will benefit more in a language-free environment in which all instructions are presented in graphical format (symbols or icons, without any textual elements). With the findings from this study, we are planning future research to answer this question. Although the overlay guide did not show consistent advantages or disadvantages with respect to performance, it was perceived as helpful with hand positioning and provided visual feedback of where users' hands were in relation to the scanner area. Use of the overlay guide is recommended for use in the fingerprint self-capture process; however, we would recommend that it be placed directly on the scanner.[3]

ACKNOWLEDGMENT

This work was funded by the Department of Homeland Security Science and Technology Directorate.

REFERENCES

Guan, H., Theofanos, M., Choong, Y. Y., & Stanton, B. (2011). *Real-time feedback for usable fingerprint systems*. International Joint Conference on Biometrics (IJCB), pp. 1–8.

[3]The material in this chapter was taken from Y.Y. Choong, M.F. Theofanos, and H. Guan, Fingerprint Self Capture: Usability of a Fingerprint System with Real-Time Feedback. IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), September 23–26, 2012. Please refer to that paper for the complete report.

International Organization for Standards (ISO)(1998). 9241–11 *Ergonomic requirements for office work with visual display terminals (VDTs)*. Part 11. Guidance on usability. Geneva, Switzerland.

Stanton, B., Theofanos, M., Steves, M., Chisnell, D., & Wald, H. (2012). *Fingerprint scanner affordances*, NISTIR (to be published), National Institute of Standards and Technology, Gaithersburg, MD.

Tabassi, E., Wilson, C., & Watson, C. (2004). *Fingerprint image quality*. NISTIR 7151, National Institute of Standards and Technology, Gaithersburg, MD <http://www.nist.gov/customcf/get_pdf.cfm?pub_id=905710>.

Theofanos, M., Stanton, B., Sheppard, C., Micheals, R., Zhang, N. F., Wydler, J., et al. (2008). *Usability testing of height and angles of ten-print fingerprint capture*, NISTIR 7504, National Institute of Standards and Technology, Gaithersburg, MD.

Biographies

Mary Frances Theofanos is a computer scientist at the National Institute of Standards and Technology where she is the program manager of the Common Industry Format Standards for usability and the principal architect of the Usability and Security Program evaluating human factors and usability of cyber security and biometric systems. She spent 15 years as a program manager for software technology at the Oak Ridge National Laboratory complex of the U.S. DOE. She received a Master's in Computer Science from the University of Virginia.

Brian Stanton obtained his Master's degree in Cognitive Psychology from Rensselaer Polytechnic Institute and is a cognitive scientist in the Visualization and Usability Group at the National Institute of Standards and Technology where he works on the Common Industry Format project developing usability standards and investigates usability and security issues ranging from password rules and analysis to privacy concerns. He has also worked on biometric projects for the Department of Homeland Security, Federal Bureau of Investigation's hostage rescue team, and with latent fingerprint examiners. Previously he worked in private industry designing user interfaces for air traffic control systems and B2B web applications.

Yee-Yin Choong is a research scientist at the National Institute of Standards and Technology. Her research focuses on applying human factors and usability disciplines to technologies, including graphical user interface design, symbols and icons design, biometrics technology usability, and cyber security and usability. Yee-Yin holds graduate degrees in Industrial Engineering from Pennsylvania State University and Purdue University, respectively.

10.3 REDESIGN OF A WEB EXPERIENCE MANAGEMENT SYSTEM

By Tanya Payne, Grant Baldwin, and Tony Haverda, OpenText

Web Experience Management is an enterprise software product designed to create, edit, and manage websites. Generally, the websites it manages are quite

large and complex. For example, they use large databases to store the website assets and presentation rules to decide what dynamic content is displayed. The product has both a console and a preview (close to WYSIWYG) interface for interacting with the content. The console view is designed for managing lists of content and bulk actions where the preview view is for editing content and the design.

The previous release of the Web Experience Management system in-context tools palette was widely criticized as being too large and "in the way" all the time. The User Experience team was tasked with redesigning the in-context tools so that they were easy to use as well as small and out of the way. The original "tools palette" can be seen in Figure 10.10.

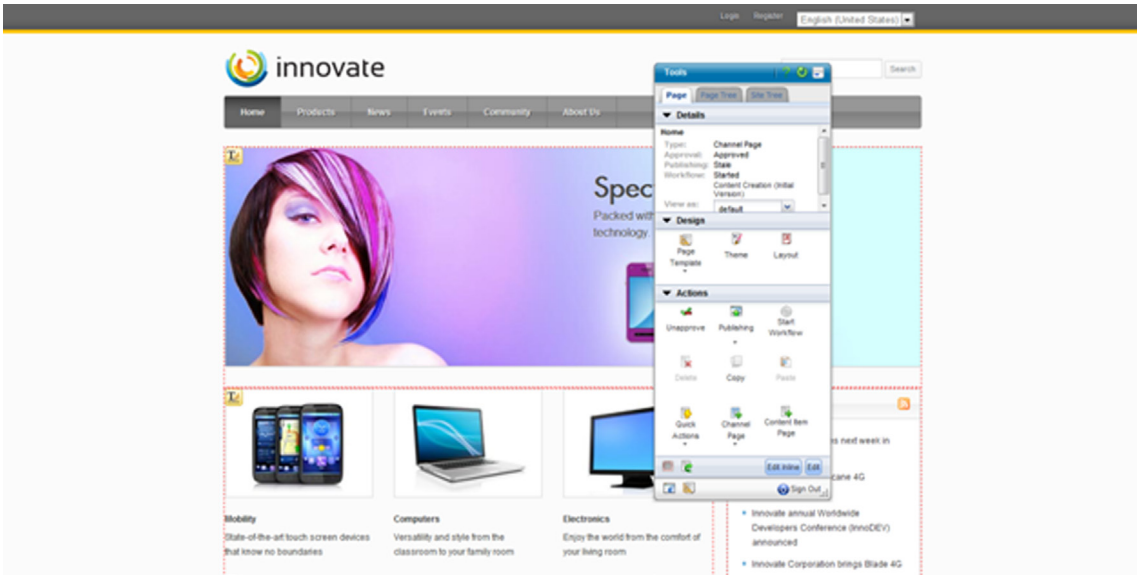


Figure 10.10 The large tools palette can be seen on top of a web page. It is easy to see that the tools palette covers a great deal of the page.

The design team focused on creating a minimal sized "toolbar," effectively reducing the existing large tools palette to the size of a menu bar. The full functionality and options in the existing tools palette could be accessed through slide-out expansion "drawers" as a user made selections from the primary icons on the new toolbar. This approach reduced the size and overall presence of the in-context tools dramatically.

10.3.1 Test Iterations

Using Axure-generated HTML prototypes, we performed a total of six rounds of usability testing on the toolbar approach during the design phase. We iterated

on the design between rounds of testing, adding functionality as more requirements were incorporated. If a particular area of the product performed below our expectations, we often retested it in subsequent rounds. Because our focus was improving the design, if data integrity conflicted with the right thing to do for the design, we naturally always chose the design. Since we work in an Agile environment, we had to keep the usability testing cycles quite short, usually less than 2 weeks per round, often aiming for 1 week. Generally, we tried to keep one or two iterations ahead of the actual coding work that was going on.

For the purpose of this case study, we focused on four rounds of testing because they repeated the same workflow task with four different designs. We will refer to these as Rounds 1–4.

Employing a resource we often use to expand the coverage of our team, Round 1 of testing was performed in person by a graduate student from the School of Information at the University of Texas at Austin under the mentorship of Associate Professor Randolph Bias. We conducted Rounds 2, 3, and 4 of testing remotely via a conference call and WebEx. We shared our desktop via WebEx and gave the participant control of the mouse and keyboard.

There were a total of 25 participants across the four rounds of testing. We had 3 participants in Round 1, 9 participants in Round 2, 4 participants in Round 3, and 9 participants in Round 4. Participant groups varied for the rounds of testing, partially due to budget constraints, but also due to the fact that users of the Web Experience Management product can vary significantly. Users can range from being long-time, full-time users of the system to brand new, occasional users of the system. Rounds 2 and 3 involved current users of the system, as well as users of competitors' systems recruited by a market research firm. Round 1 involved representative users from the University of Texas, and Round 4 involved current customers exclusively.

10.3.2 Data Collection

Even though all the usability tests were "formative" in nature, we collected usability metrics for each of the rounds of testing similar to the methodology reported by Bergstrom, Olmsted-Hawala, Chen, and Murphy (2011). From our perspective, usability metrics are just another way of communicating what happened during a usability test. Of course, we also collect qualitative data and that data still represent the bulk of our formative usability test results and recommendations. However, we have found metrics, being numbers, are concise and easy for management and developers to digest. Also, we find them quick and easy to collect and analyze.

We have standardized a set of metrics at OpenText and these were reported and tracked across the rounds of tests to communicate improvements in the design to product owners. Following the ISO definition of usability of "effectiveness," "efficiency," and "satisfaction," we collected task completion rate, time on task, the Single Ease Question (SEQ; Sauro & Dumas, 2009; Sauro & Lewis,

2012) after every task and the System Usability Scale at the end of the test. The SEQ is a single question asking participants to rate the difficulty of the task on a seven-point scale, where 1 is "very difficult" and 7 is "very easy."

We collected the data using Excel spreadsheets based on a template we use for every usability test that includes formulas for calculating our metrics. As a result, we can report the quantitative results almost immediately after finishing a study. Since we had product release goals, the team was very interested in hearing quickly if we were getting closer to our goals.

The only downside with our spreadsheet methodology is that one of our experimenters had a difficult time with the spreadsheet's keyboard shortcuts in the beginning of these studies. As a result, we lost many time-on-task measurements and do not have enough data to present those findings here. However, we often see interesting differences in time on task, even when testing mockups.

10.3.3 Workflow

For the purposes of this case study, we focused on one aspect of the project: workflow design and results. The workflow task in all four rounds of usability testing was essentially the same: accept a workflow task assigned to the participant, approve the page being edited, add a note to the workflow, and finish the workflow task assigned. Task instructions given to the participant were:

> *You've received an automatic workflow item! When you made changes to the homepage, a workflow automatically triggered and sent a workflow to you to approve the page. Please complete the workflow and add the note "fixed typo, changed image" so that your editor knows what you've changed.*

Workflow was a difficult task for us to get right from a design perspective. In the new slim design, we had very little room to work with to communicate complex task requirements, and we had to work within the existing system limitations. There wasn't time or resources to allow for a complete rewrite of the code. Figures 10.11–10.15 show screenshots of the design and how it changed over the four testing iterations. We started out with a very modular approach, requiring the user to find each piece of the functionality on the toolbar and ended with a more "wizard"-like approach where the user is guided through the process. In each version of the design, the toolbar is shown at the bottom of the screen in gray. The "task panel" (or later "Task Inbox") is the blue panel just above the toolbar on the left-center portion of the screen. Early versions of the design (1 and 2) included some task functions ("Accept Task") within the task panel, while later versions (3 and 4) moved those actions into a second "Task Editor" window (the panel on the right side of Design 3, and Screen 2 for Design 4). The designs, along with a brief description, appear in Figures 10.11–10.15.

WORKFLOW DESIGN 1

The first design was focused on trying to stay with the very modular design of the new toolbar. The toolbar can be seen at the bottom of Figure 10.11 with the

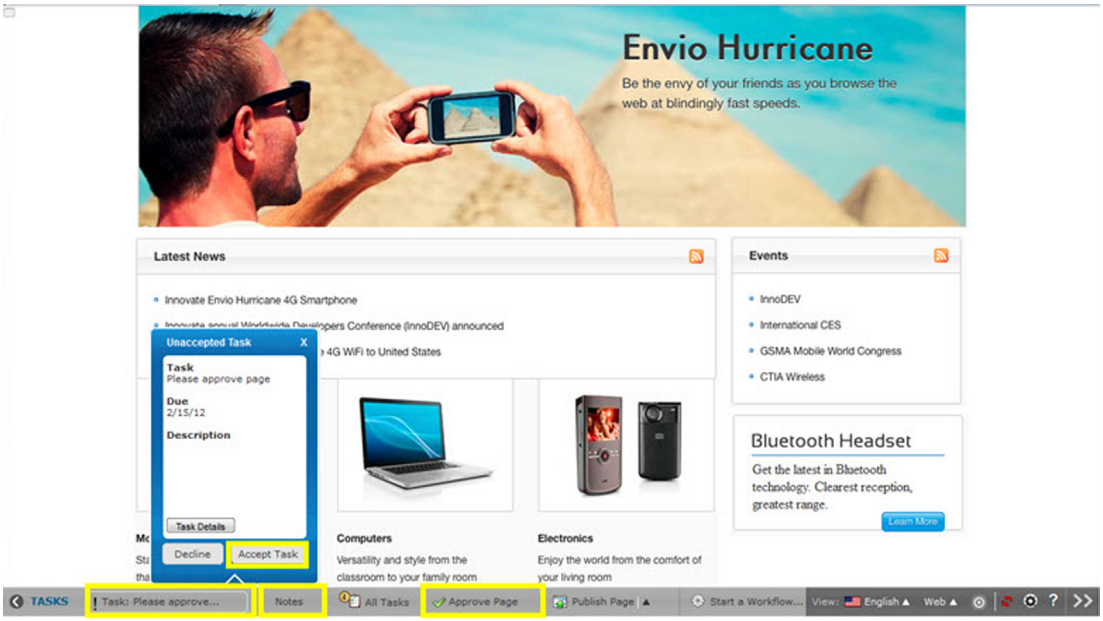


Figure 10.11 An early version of the new "toolbar" design can be seen on top of a web page. The "unaccepted task" window slides out on selection of the "Task: Please approve…." Yellow boxes indicate required selections for accepting a workflow task.

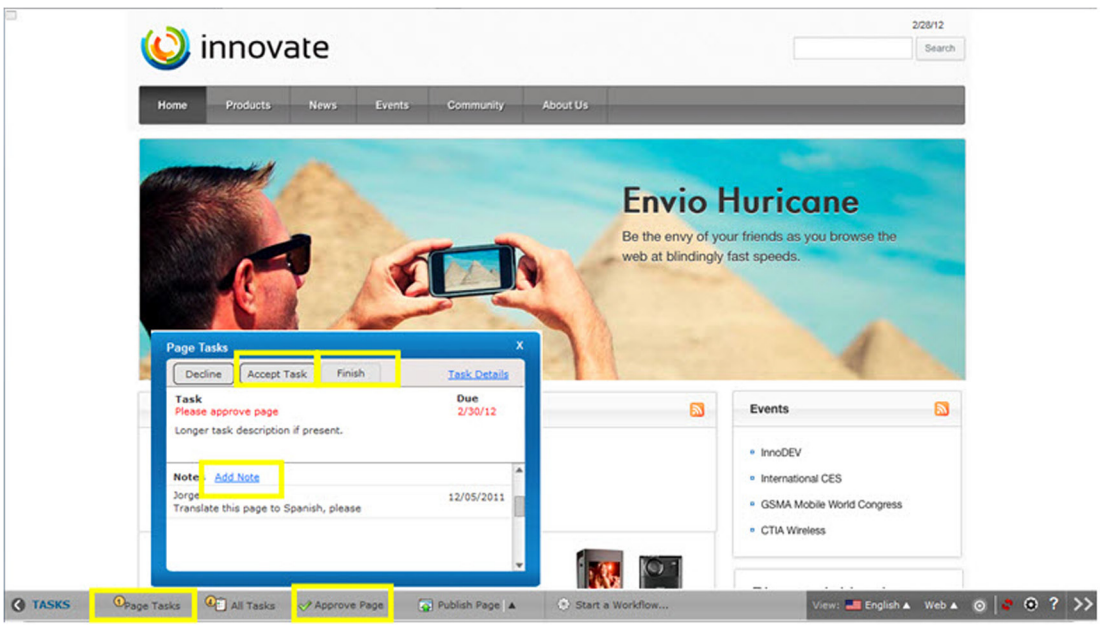


Figure 10.12 Second iteration of the workflow functionality in a toolbar design. Yellow boxes indicate required selections for accepting a workflow task.

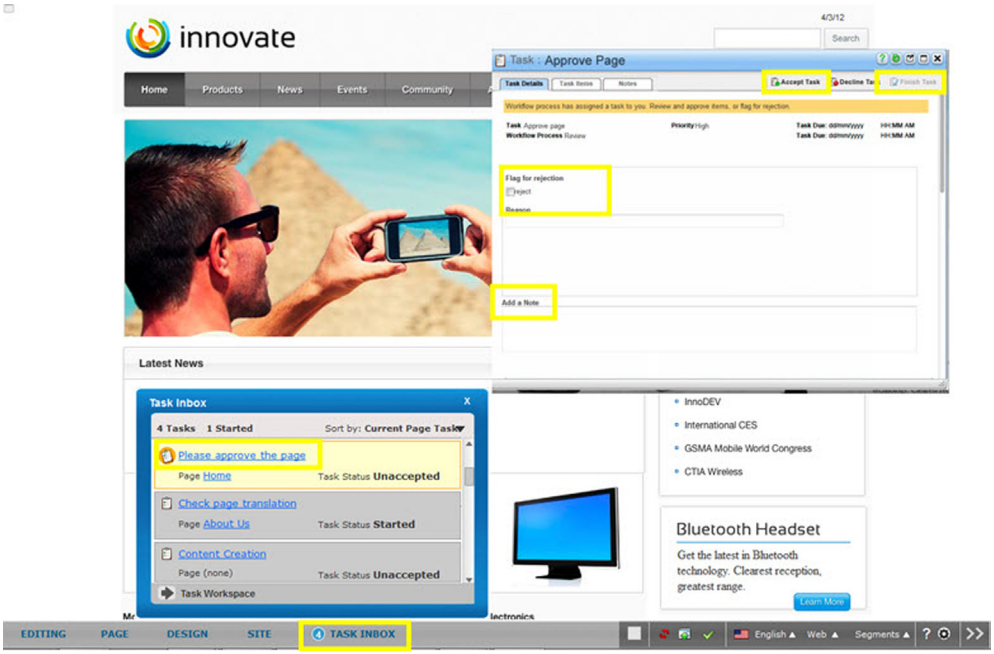


Figure 10.13 Third iteration of the workflow functionality in a toolbar design. Yellow boxes indicate required selections for accepting a workflow task.

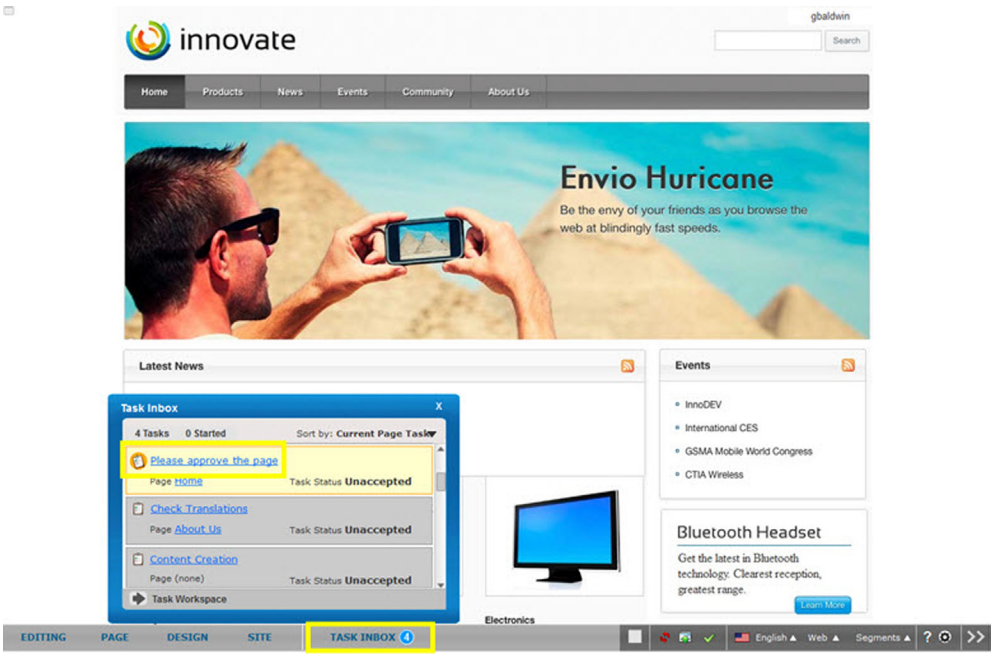


Figure 10.14 Fourth iteration of the workflow functionality in a toolbar design. A more "modular" or "wizard" approach was taken, so two images have been used to describe the interaction. Yellow boxes indicate required selections for accepting a workflow task.

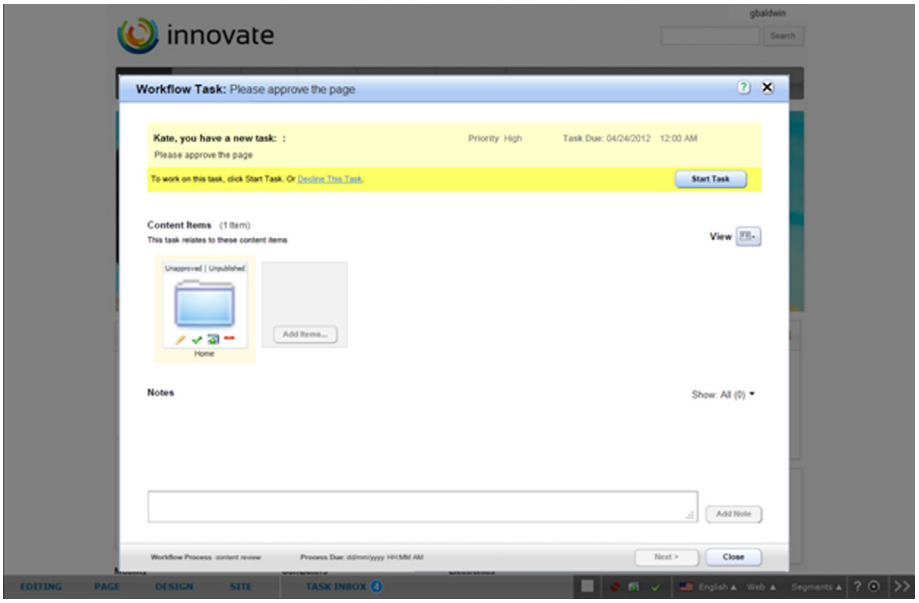


Figure 10.15 Fourth iteration of the workflow functionality in a toolbar design. A more "modular" or "wizard" approach was taken, so we have two images to describe the interaction. This page shows the actual task screen. Yellow boxes indicate required selections for accepting a workflow task.

word "tasks" to the left. Users had to perform individual steps: open the task, accept the task, approve the page and add a note, and then finish the task (the accept task became a finish task button upon selection) using different buttons. The buttons required are highlighted in yellow.

WORKFLOW DESIGN 2

The second design, as shown in Figure 10.12, attempted to streamline the experience of accepting a task and adding notes, while leaving the "approve page" button outside of the workflow. Again, participants needed to select the tasks, accept the task, add a note, approve the page, and finish the task. Buttons required to perform the tasks are highlighted in yellow.

WORKFLOW DESIGN 3

The third set of workflow designs removed the concept of a "tasks" area of the toolbar and moved all of the actions into the "editor" of the content. The concept of an automatic accept with a "flag" for rejection was also explored here, as we had seen this example at a customer site. In this design, participants were required to select tasks, select the correct task and then accept the task, reject the design, add a note, and finish the task (see Figure 10.13). Buttons required to do the tasks are highlighted in yellow.

WORKFLOW 4 SCREEN 1

The fourth and final design took the idea of putting workflow into a single popup further, making the popup a bit more like a "wizard" experience. Participants needed to select "tasks" and select the correct task, after which a popup came up, represented by Figure 10.14.

WORKFLOW 4 SCREEN 2

In screen 2 (see Figure 10.15), participants started at the top of the screen with a "start task" button. Then participants moved down the screen to approve the item by clicking directly on a green check box associated with the item and added a note. At the time the item was "approved," the "next" button at the bottom of the screen was replaced with a "finish task" button.

10.3.4 Results

We were able to demonstrate an improvement in the design of workflow, as indicated by the SEQ and task completion rate (see Figures 10.16 and 10.17).

The SEQ yielded our most interesting results. The mean SEQ score for the workflow task increased in each round of testing, indicating that participants found the task easier with each design iteration. In the initial round of testing, the mean SEQ for workflow was a 3.0. In Round 2, the mean SEQ increased to 3.9, and in Round 3, to 4.5. By Round 4 of testing, the mean SEQ had increased to 5.9.

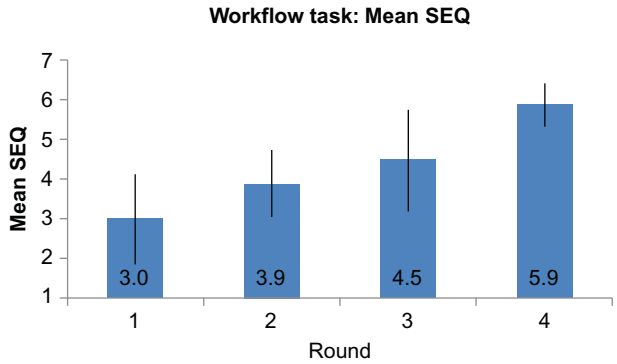


Figure 10.16 Mean SEQ ratings increased from Round 1 to Round 4.

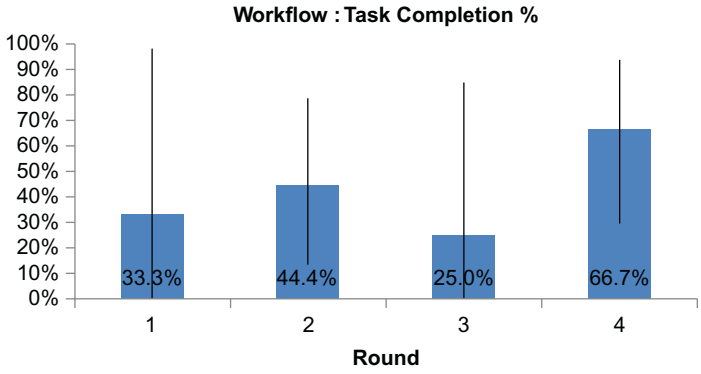


Figure 10.17 Task completion rates were slightly higher in Round 4.

The task completion rate for workflow also increased from Round 1 to Round 4, but not in the same way as the SEQ scores. The task completion rate went from 33% in Round 1 to 44% in Round 2, but decreased to 25% in Round 3. Finally, the task completion rate was highest in Round 4, at 67%. The drop in the task completion rate in Round 3 was interesting; although fewer participants completed the task successfully, the SEQ score was higher than Rounds 1 and 2. Participants rated the design in Round 3 easier to use than Round 1 or 2, even though it was actually more difficult for them to use.

Because we were conducting formative testing, we were not overly concerned with statistically significant differences between rounds of testing. However, we still calculated 95% confidence intervals as a way of assessing the variability in our data. Even with the small and unequal sample sizes between rounds, we were able to resolve some differences in SEQ scores. The confidence intervals for SEQ (shown as error bars in Figure 10.16) suggest that participants thought the workflow task in Round 4 was substantially easier than in Round 1 or 2. In contrast, the confidence intervals for task completion rate (error bars in Figure 10.17) were much larger and suggest that the differences we saw were small.

Because the new toolbar design was such a large departure from previous designs, we also looked for any differences in data between current users of the system and users of competitive systems during Rounds 2 and 3. We did not see any differences between the different user groups.

10.3.5 Conclusions

Like Bergstrom and colleagues (2011), we found quantitative metrics to be useful in formative testing with rapid iteration cycles. For us, that included rounds of testing conducted both by us and by partners at the University of Texas at Austin. By using task-level measures such as task completion rate and SEQ, we could retest certain aspects of the design, such as workflow, across multiple design iterations independently from the rest of the product. That allowed us to track the progress we made in our design, while also allowing us to add or remove tasks to test other parts of the product.

Because our test changed from round to round, we found that using the SEQ was very useful for us. The SEQ provided us with a metric we could use at the task level to get participants' subjective impressions of ease of use. We found that the SEQ could be sensitive enough to resolve differences between different designs, even with only three or four participants per round of testing.

REFERENCES

Bergstrom, J., Olmsted-Hawala, E., Chen, J., & Murphy, E. (2011). Conducting iterative usability testing on a web site: Challenges and benefits. *Journal of Usability Studies, 7*(1), 9–30.

ISO 9241-124. *Ergonomics of human-system interaction*.

Sauro, J., & Dumas, J. S. (2009). *Comparison of three one-question, post-task usability questionnaires*. Computer human interaction conference. <http://www.measuringusability.com/papers/Sauro_Dumas_CHI2009.pdf)>.

Sauro, J., & Lewis, J. R. (2012). *Quantifying the user experience: practical statistics for user research*. Morgan Kaufmann.

Biographies

Tanya Payne has been working in the User Experience field for about 17 years. During that time she's worked as a contractor, a consultant, and an in-house employee. She's worked on a variety of consumer and enterprise products,

including cell phones, printers, RISC600 servers, children's touch screen paint product, and content management applications. Tanya received her Ph.D. in Cognitive Psychology from the University of New Mexico. She currently works at OpenText as a Senior User Experience Designer.

Grant Baldwin has been a User Experience professional at OpenText for 2 years, working on a range of enterprise software applications. Grant has an M.A. in Cognitive Psychology from The University of Texas at Austin and a B.S. from Ohio State University. He currently works at OpenText as a User Experience Designer.

Tony Haverda has worked in the User Experience field for over 23 years. Most recently he has been senior manager of User Experience Design at Open Text for the past 4 years. Tony holds a M.S. degree in Industrial Engineering–Human Factors and a B.S. degree in Computer Science, both from Texas A&M University. He current is senior manager of the User Experience Design Group at OpenText.

10.4 USING METRICS TO HELP IMPROVE A UNIVERSITY PROSPECTUS

By Viki Stirling and Caroline Jarrett, Open University

The Open University is the U.K.'s largest university, with over 200,000 students, and the only one dedicated solely to distance learning. Its online prospectus receives approximately six million visitors each year. Ninety percent of the students register online, accounting for approximately £200 million (about US $300 million) of registrations each year.

The team, with overall responsibility for development of the Open University's web presence, is led by Ian Roddis, head of Digital Engagement in the communications team. He co-ordinates the efforts of stakeholder groups, including developers, user experience consultants, academics, and many others. The team has been committed to user-centered design for many years now by involving users directly in usability tests, participatory design sessions, and other research and indirectly through a variety of different data sets, including search logs and web tracking. But the real value comes from triangulation, using several different sets of data together—as illustrated in Figure 10.18, from Jarrett and Roddis (2002).

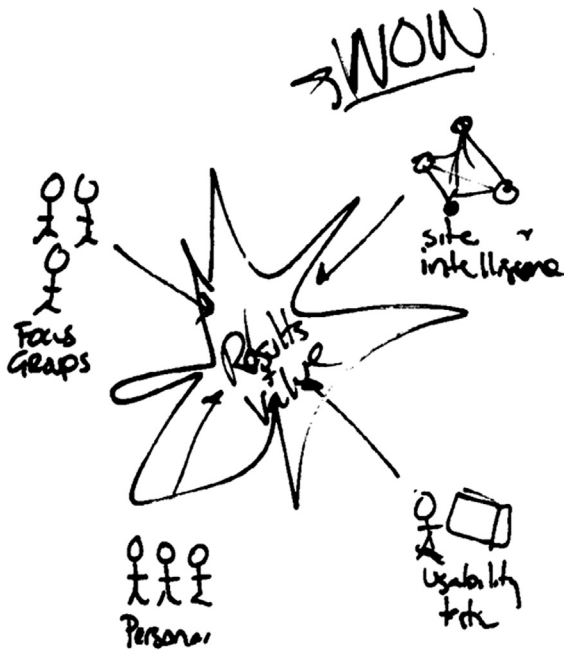


Figure 10.18 WOW: Results and value—sketch from 2002 presentation on the value of UX measurement and triangulation.

10.4.1 Example 1: Deciding on Actions after Usability Testing

One of our earliest examples of triangulation started with a usability test. The prospectus homepage consisted of a long list of subjects (Figure 10.19).

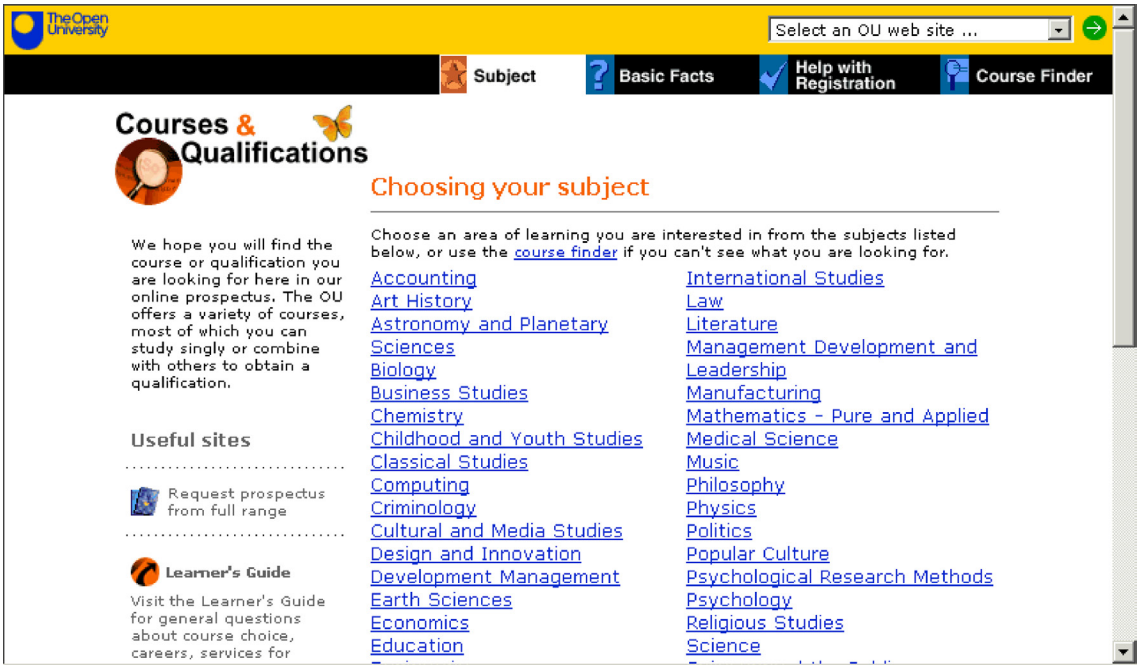


Figure 10.19 Original list of subjects on the prospectus homepage, as seen on a typical screen.

Most people who consider university study start by looking for the subject they are interested in. When we asked participants in a usability test to look for the subject they wanted, we observed that some of them struggled:

- When viewed on a typical screen at that time, some of the list was "below the fold" and not visible to the user (Figure 10.20).
- The list was presented in alphabetical order, which meant that some related subjects (e.g., Computing and Information Technology) were separated from each other.

We could have done more testing with more participants to measure exactly how much of a problem this was, but instead we decided to use web analytics to investigate the actual behavior of site visitors.

Figure 10.20 Scrolling down revealed "missing" subjects, such as Information Technology, more sciences, Social Work, and Teacher Training.

WEB ANALYTICS TOOLS AT THE OPEN UNIVERSITY

The Open University uses commercial web analytics tools, reviewing the choice of tools from time to time. Our current tracking tool is Digital Analytix from comScore. We tag each web page that we want to track, and we ask web visitors to give us permission to use cookies to track their visits. The tool then logs each page visit and the path taken by each visitor through the website.

We can also distinguish between visits by logged-in visitors (students and staff) and by other visitors. We distinguish between a single visit—the path someone takes through our site in one continuous experience—and the experience of a visitor—the aggregation of multiple visits from a computer where someone has given us permission to use cookies.

It can be tricky to distinguish different types of visit and visitor, so we find that it's best to try to focus on the big overall picture and not stress too much about finer details.

For example, we discovered that 37% of visits that involved Information Technology also involved Computing, but that only 27% of visits that involved Computing also involved Information Technology. In addition, we found that Computing was receiving 33% more visitors than Information Technology. This confirmed what we'd seen in usability testing: our participants were more likely to click on Computing (above the fold) than on Information Technology (below the fold).

We looked at the content of these two subjects and discovered that prospective students should really think about both of them before choosing either. From this type of analysis, across the entire list of subjects, we recommended a new design with a much shorter list of subject areas based on actual user behavior, and the clusters of subjects they tended to view together (see Figure 10.21).

The previous organization of subject areas reflected the internal structure of the university at that time; for example, the Mathematics and Computing faculty

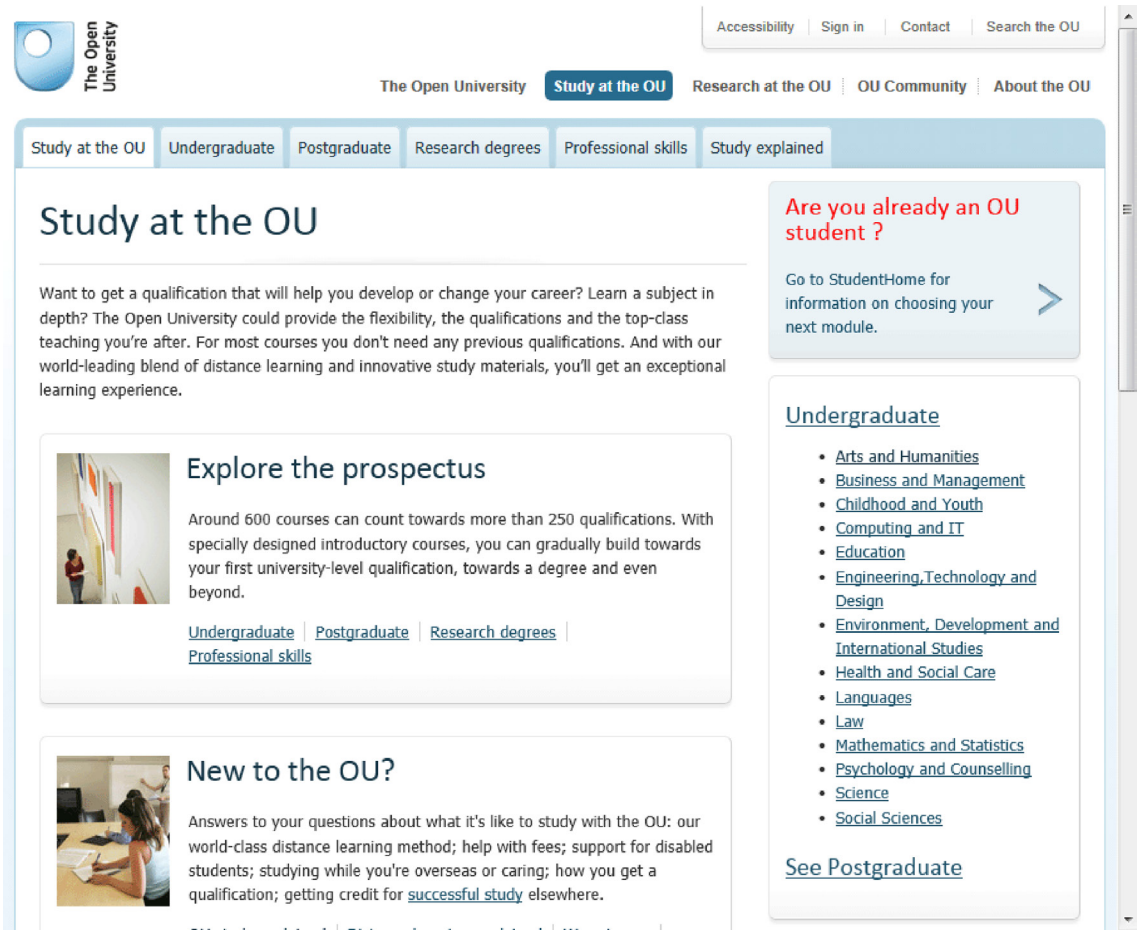


Figure 10.21 The prospectus homepage in 2012; a short but effective list of subjects.

taught Computing, but the Technology faculty taught Information Technology. The revised organization aligns with visitor expectations and needs and has performed well (with a few tweaks) ever since.

10.4.2 Example 2: Site-Tracking Data

The usability test described in Section 10.4.1 was a major initiative that required a lot of data to persuade many different stakeholders—the type of thing you only want to do occasionally.

This second example is more typical of our everyday work. Some stakeholders came to Viki Stirling, who looks after analytics and optimization, with a problem: they weren't getting the expected level of conversion from part of their website.

Viki took the site-tracking data and fed the appropriate tracking data into NodeXL, a visualization tool.

Looking at the flows by visits, the problem jumped out at her immediately: A lot of visits arrive at a particular page, but few continue after that (highlighted in red in Figure 10.22). The big arrows from node to node should continue, getting

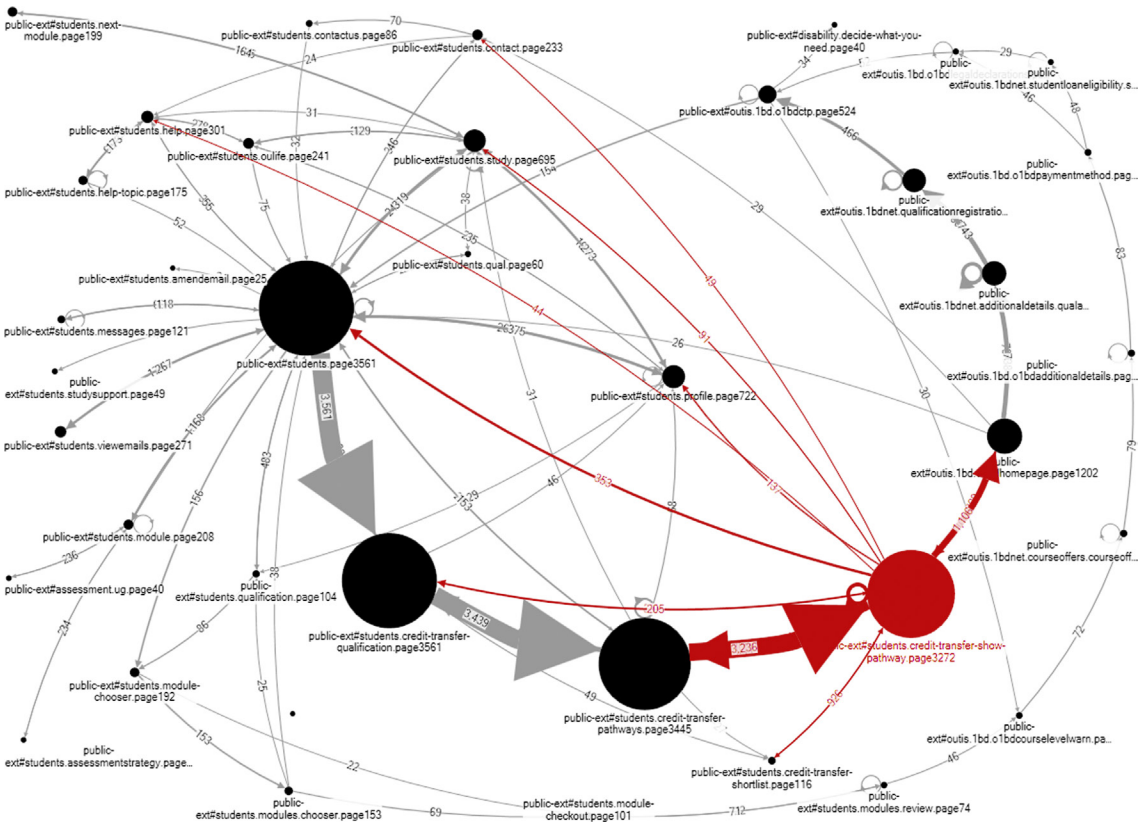


Figure 10.22 This view of flows by visit shows one page where plenty of visits arrive, but few move on to the next step.

only slightly smaller at each step. At the problematic page, we suddenly see that the larger arrow flows backward up the chain, with only a small arrow moving on to the next step.

When she investigated the problematic page, it was obvious how to revise it. But Viki was suspicious: although this task isn't common, it's important for the relatively few visitors who attempt it. She investigated further, looking at the flows by visitor (see Figure 10.23). This revealed that the previous step in the process was also causing problems: visitors are moving backward and forward from that step, clearly trying to make progress but failing. Once again, a look at the relevant web page quickly revealed the necessary changes.

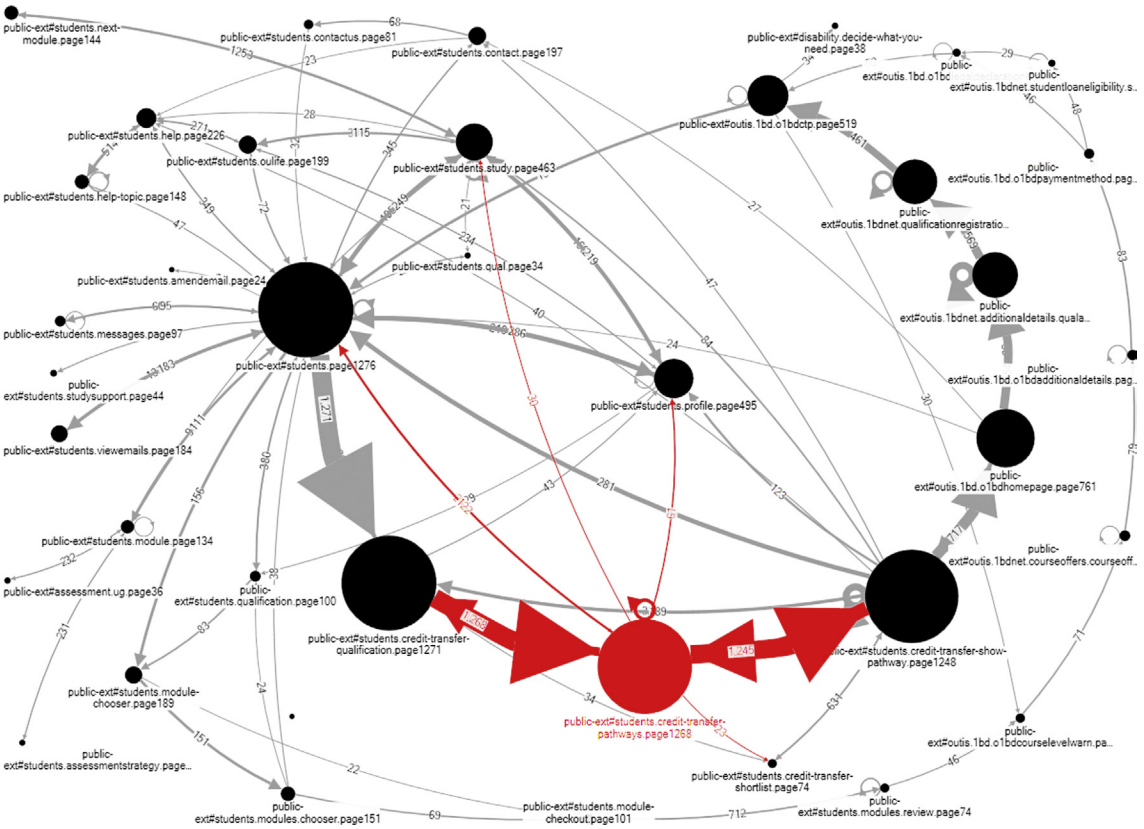


Figure 10.23 Flows by visitor show that an earlier step in the process is also causing difficulty.

From the UX point of view, we might immediately ask: why didn't the stakeholders do usability testing, which would probably have revealed these problems ahead of time? The answer is that, of course, the Open University does lots of usability testing, but they face a challenge familiar to any organization with a huge and complex website, which is one of prioritization. In this example, the

problematic task is rather unusual and relevant only to a small number of users at a very specific point in their progression from enquirer to student.

10.4.3 Example 3: Triangulation for Iteration of Personas

The two previous examples demonstrate the use of measurement techniques for specific changes. Our third example illustrates the use of metrics for one of the UX tools we use all the time: personas.

We first started using personas after Caroline Jarrett learned about them from Whitney Quesenbery at the Society for Technical Communication Conference in 2002. They were based on our experience of usability test participants over a few years—by that point we had been usability testing since 1998—and Sarah Allen validated them against various internal data sources at the time. With Whitney's help, we've been using, updating, and revalidating the personas ever since. Pruitt and Adlin (2006) include a short overview of our experience with personas.

For example, the Open University introduced Foundation Degrees, shorter degree programs focused on training for particular jobs that are somewhat similar to the U.S. "Associates Degree." To help with our design activities around Foundation Degrees, we added in a persona, "Winston," who was interested in the Foundation Degree in Materials Fabrication and Engineering. But we discovered that we weren't meeting Winstons in usability tests. Viki Stirling had the idea of doing some visit tracking to see whether the routes through the site that we envisaged for the personas were actually sufficiently based in data. She discovered that most of them were, but Winston really wasn't justified, the numbers just weren't there. Winston became Win, interested in the Foundation Degree in Early Years (see Figure 10.24).

Segment: Not Employed Adults (C2)
- 24-49
- Not employed
- Considering HE
- No OU experience
- No degree
- Progress career

Awareness/Exploration

Win, turn my job into a career

The Open University

Tell us a bit about yourself	I'm 32, I'm married with two children: Lewis, 10 and Florence, 7. We live in Cardiff. Since Lewis was a baby, I've been working as a registered childminder but my current kids are going to school next year so I'm really looking for a job now.
Have you got any qualifications?	National Childminding Association - Diploma for the Children and Young People's Workforce
What is your ambition?	I love working with children, but being a childminder at home doesn't pay well. I want to turn that into a career.
Why didn't you go to university?	Where I lived, you were pleased to survive school and we didn't come out with anything. It was hard enough to get any type of job, never mind uni.
What do you want to know?	How many hours per week? How long will it take? How will it help my career?
How did you find out about the OU?	I searched for 'early years distance learning'

Figure 10.24 Persona "Win" at the start of her journey to becoming a student.

Lindsay's reasons for studying are slightly different to Win's, and she's focusing slightly more on costs and fees—but overall, she's close enough

that we can be confident that a design intended for persona Win will also work for real aspiring students like Lindsay.

10.4.4 Summary

Most user experience techniques are valuable on their own, and we're happy to use them individually, as illustrated by our everyday example, number 2 above.

We find that the real value comes from comparing what we learn from larger scale quantitative techniques with what we learn from small-scale, qualitative techniques—and continuing to do that over many years.

ACKNOWLEDGMENTS

We thank our colleagues at the Open University: Sarah Allen and Ian Roddis, and at Whitney Interactive Design: Whitney Quesenbery.

REFERENCES

Jarrett, C., & Roddis, I. (2002). *How to obtain maximum insight by cross-referring site statistics, focus groups and usability techniques. Web based surveys and usability testing*. San Francisco, CA: Institute for International Research.

Pruitt, J., & Adlin, T. (2006). *The persona lifecycle: Keeping people in mind throughout product design*. San Francisco: Morgan Kaufmann.

Biographies

Viki Stirling, as eBusiness Manager of Analytics and Optimization in the Digital Engagement team at Open University, is responsible for leading the understanding of actual customer on/off-site behaviors. She manages the integration and implementation of online business analytics, both quantitative (web analytics) and qualitative (sentiment), to provide insight and recommendations that inform institutional strategy, addresses the university's business objectives, and improves e-business performance. As she is particularly interested in the relationship between analytics and user experience, she regularly provides analytics insight to support usability testing, persona development, and optimization of the customer journey.

Caroline Jarrett, after 13 years as a project manager, started her business, Effortmark Limited. She became fascinated with the problem of getting accurate answers from users when she was consulting with HM Revenue and Customs (the U.K. tax authority) on how to deal with large volumes of tax forms. She became an expert in forms design and is coauthor of "Forms That Work: Designing Web Forms for Usability." Along the way, she completed an MBA with the Open University, which led to coauthoring the textbook "User Interface Design and Evaluation" and to consulting on the user experience of their vast and complex website. Caroline is a Chartered Engineer, Fellow of the Society for Technical Communication, and the cofounder of the Design to Read project, which aims to bring together practitioners and researchers working on designing for people who do not read easily.

10.5 MEASURING USABILITY THROUGH BIOMETRICS

By Amanda Davis, Elizabeth Rosenzweig, and Fiona Tranquada, Design and Usability Center, Bentley University

A group of Bentley University researchers from the Design and Usability Center (DUC) wanted to understand how the emotional experience of using a digital textbook compared to a printed textbook. In 2011, our team of graduate students (Amanda Davis, Vignesh Krubai, and Diego Mendes), supervised by DUC principal consultant Elizabeth Rosenzweig, explored this question using a unique combination of affective biometric measurement and qualitative user feedback. This case study describes how these techniques were combined to measure emotional stimulation and cognitive load.

10.5.1 Background

As user experience research achieves greater prominence in business organizations, we are often asked to help gauge the emotional experience of a product, as well as its usability. Usability professionals have a variety of tools and techniques available to understanding human behavior. However, the tools used commonly to measure the emotion of participants while attempting to complete tasks rely on either an observer's interpretation of how the participant is feeling or the participants' description of their reactions (e.g., through think-aloud, protocolor post-task ratings). These interpretations are subject to phenomena such as the observer effect, the participants' inclination to please, and the time passed since they had the reaction. Other tools, such as the Microsoft Product Reaction Cards, show the direction of a participant's response (positive or negative) to a product, but not the magnitude of that emotion (Benedek & Miner, 2002).

Adding biometric measures to user research provides a way to measure users' arousal as they use a product. Arousal describes the overall activation (emotional stimulation and cognitive load) experienced by a user, as measured by biometric measures such as electrodermal activity (EDA). These measures capture physiological changes that co-occur with emotional states (Picard, 2010). Because biometric measures are collected in real time during a user's interaction with a product, they provide a direct measurement of arousal that is not affected by observer or participant interpretation.

This case study describes initial research to gauge the effectiveness of a new technique that assigns meaning to biometric measures. We hypothesized that by combining biometric measures with feedback from the Microsoft Product Reaction Cards, we could gain a detailed description of a user's arousal, the interaction that activated a change in arousal, and assignment of emotion (positive or negative) to that interaction. For example, if we saw a user's arousal level increase sharply while attempting a search task, the Microsoft Product Reaction Cards selected would indicate whether the arousal increased due to frustration or pleasure. This combination would let practitioners quickly identify areas of a

product that participants found more or less engaging or frustrating, even if the participants do not articulate their reaction.

Our Bentley DUC team partnered with Pearson Education, a textbook publisher that had recently moved into the digital textbook space on the iPad. We focused this new technique on a usability study that would highlight any differences in arousal between digital and paper textbooks.

10.5.2 Methods

PARTICIPANTS

We recruited 10 undergraduates who owned and used iPads. Each 60-minute session was one on one with a moderator from the Bentley Design and Usability Center in the room with the participant.

TECHNOLOGY

To gather affective measurement and user feedback, we used two innovative tools. Affectiva's Q Sensor was used to identify moments of increased arousal. The words selected by participants using the Microsoft's Product Cards enabled us to understand the direction (positive or negative) of their emotions.

The Affectiva Q Sensor is a wearable, wireless biosensor that measures emotional arousal via skin conductance. The unit of measure is electrodermal activity, which increases when the user is in a state of excitement, attention, or anxiety and reduces when the user experiences boredom or relaxation. Since EDA captures both cognitive load and stress (Setz, Arnrich, Schumm, La Marca, Troster, & Ehlert, 2009), we used this technology to accurately pinpoint moments of user engagement with digital and printed textbooks. Affectiva's analysis software provides markers used to indicate areas of interest in the data. Depending on the study, areas of interest may include task start and end times. These markers can be set during a study or post-test.

To better understand the emotions of the user, we utilized a toolkit developed by Microsoft called Microsoft Product Cards. These cards are given to users to form the basis for discussion about a product (Benedek & Miner, 2002). The main advantage of this technique is that it does not rely on a questionnaire or rating scales, and users do not have to generate words themselves. The 118 product reaction cards targeted a 60% positive and 40% neutral balance. A study out of Southern Polytechnic in Georgia found that cards encourage users to tell a richer and more revealing description of their experiences (Barnum & Palmer, 2010). This user feedback helped the DUC team assign specific emotions to the Q Sensor's readings. Without these cards, we would have needed to make inferences about the peaks and lulls found in the Q Sensor data.

PROCEDURE

Each session was structured as follows:

1. When participants arrived, we attached the Q Sensor biosensors to each of their hands. Participants were asked to walk down the hallway and

back so a small amount of electrolyte solution (sweat) would be generated. This sweat was necessary to establish a connection between the skin surface and the Q Sensor's electrodes.

2. Participants attempted seven tasks ("homework questions") on the digital textbook and the paper textbook. Half of the participants conducted tasks using the digital textbook first, while the other set of participants used the printed textbook first. The first set of tasks included four tasks and the second set of tasks included three tasks. Participants were asked to think aloud as they completed their tasks.
3. After they completed their tasks on either the digital textbook or the printed textbook, participants used the Product Reaction Cards to indicate their reaction to the experience.
4. After participants had used both textbooks, we asked a few open-ended questions about the comparative experience, what they liked best and least, and which version of the textbook they would prefer if they had to select one.

10.5.3 Biometric Findings

Q SENSOR DATA RESULTS

For Q Sensor data analysis, we divided Q Sensor data by task. As each participant was wearing two gloves, we were able to collect two different sets of data for each task. Of the 140 tasks data points (10 participants with two gloves, across seven tasks), 102 data points from 9 participants remained after removing poor quality biometric data and missed tasks. Figure 10.25 provides an example of the Q Sensor analysis software.

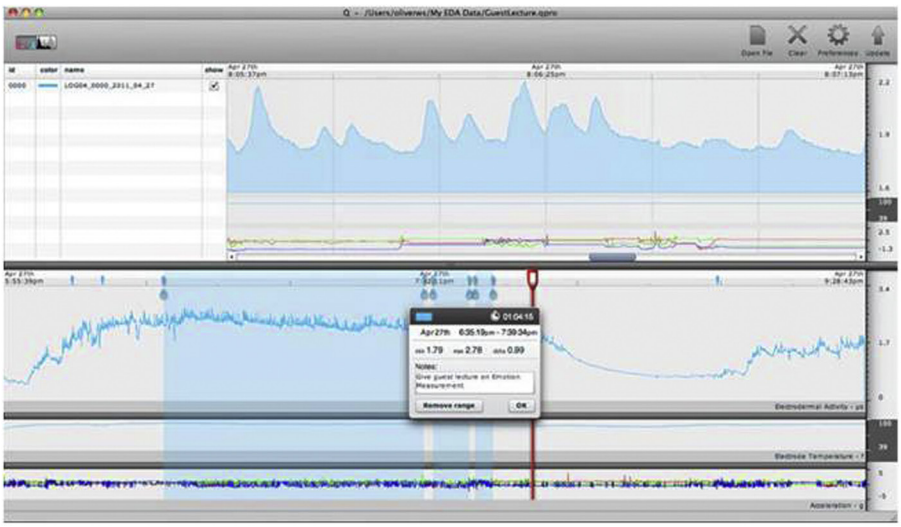


Figure 10.25 Affectiva's analysis software showing a single participant's results from the Q Sensor. The bottom half screen shows the participant's electrodermal activity during the session, while the top right zooms in to a particular shorter period of time.

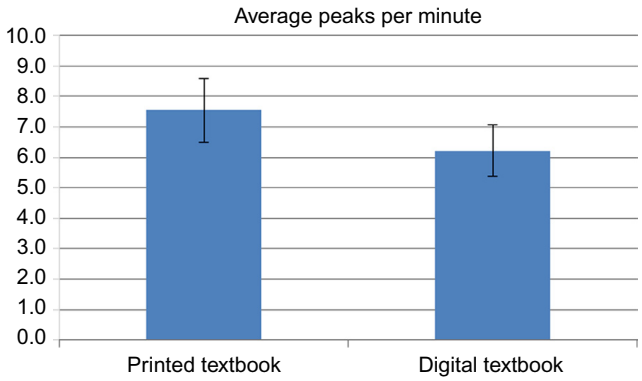


Figure 10.26 Average peaks per minute with 95% confidence limits for printed textbook and digital textbook tasks.

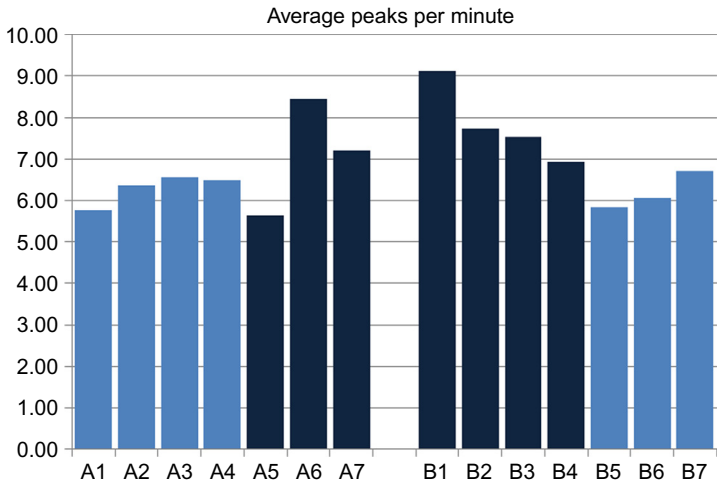


Figure 10.27 Average peaks per minute per task broken out by group. Group A used the digital textbook first; group B used the digital textbook first.

We then compared the number of peaks per minute for each participant's tasks using the digital textbook to tasks using the printed textbook. Results showed that the digital and the paper textbooks had average peaks per minute of 6.2 and 7.6, respectively. However, using a paired samples *t* test, this difference was not statistically significant ($p = 0.23$) at a 95% confidence interval (see Figure 10.26). Figure 10.27 shows the average number of peaks by participant for the two groups. Comparing the peaks per minute across the different tasks, the paper textbook had higher peaks per minute than the iPad textbook on six out of the seven tasks.

10.5.4 Qualitative Findings

Once we observed that average peaks per minute were trending higher for the paper textbook than the digital textbook, we compiled the qualitative feedback that we had collected from the Microsoft Reaction Cards, as well as the poststudy questions. While participants described the digital version as "organized," "easy to use," and "efficient," participants described the paper textbook as "slow," "time-consuming" and "old." Figures 10.28 and 10.29 show word clouds for the paper text book and digital textbook, respectively. The larger the font, the most frequently the card was selected. The shade of the text does not have any meaning.

Affectiva Q Sensor data showed us that participants experienced higher arousal while using the printed textbook, but combining those results with qualitative data from the Microsoft Product Reaction Cards, as well as moderated discussion, revealed that that the higher levels were due to negative emotions from a difficulty in performing search and comprehension tasks.

At the end of the session, the moderators asked participants to choose between digital and printed versions of the textbook; surprisingly, participants were split with five preferring the digital and five preferring the printed textbook. Although this was not the focus of our research, this split suggests that the

Figure 10.28 Word cloud from Microsoft Reaction Cards for paper textbook.

Figure 10.29 Word cloud from Microsoft Reaction Cards for digital textbook.

decision between paper and digital textbooks relies on more than the difference in emotional experiences.

10.5.5 Conclusions and Practitioner Take-Aways

This study successfully tested the feasibility of integrating biometric measures with qualitative user feedback. Results showed that the benefits of this technique over standard usability testing include:

- Direct measurements of a participant's arousal
- Triangulation of sources to explain and validate findings

Specifically, additional information gained from the Q Sensor for this study redirected and clarified the impressions that we had based on observation and the participants' thinking aloud. Based on those measures, we expected that the digital textbook would have been more stimulating based on how the participants described and interacted with it. However, Q Sensor data revealed that participants had a higher arousal level while using the paper textbook. Other qualitative data indicated a negative direction for those emotions, as participants struggled with their tasks. This negative emotion was stronger than the pleasurable emotions felt while using the digital textbook.

This approach would be ideal for projects whose goal is to understand participant emotional responses and the severity of those reactions throughout their interaction with a product. By associating metrics across data sets, researchers can pinpoint a participant's exact emotional reaction, and what was causing that reaction, at any point during their session. This unique combination of metrics provides a new window into a participant's emotional reactions above and beyond what is articulated during a standard think-aloud usability study.

However, these techniques require additional time to set up the study appropriately and to analyze the results. For example, researchers will want to plan on using time markers with the Q Sensor. We learned during our data analysis that we could have saved significant efforts postanalysis by adding more markers to Q Sensor data during the sessions. For this study, the team spent approximately 2 work weeks scrubbing, combining, and analyzing the data. However, a more recent project that's used these same techniques only took us 3 work days as we used more Q Sensor markers during the sessions. This method probably won't make sense for a basic formative usability study, but we believe would offer benefits for projects with a larger scope.

We are continuing to refine and build out these techniques through additional projects and are applying them to new domains.

ACKNOWLEDGMENTS

Thanks to the Design and Usability Center at Bentley University for their support, to Affectiva for use of the Q Sensor and analysis support, and Pearson Education for providing the digital and printed textbooks. Also, thanks to Vignesh Krubai, Diego Mendes, and Lydia Sankey for their contributions to this research.

REFERENCES

Barnum, C., & Palmer, L. (2010). More than a feeling: Understanding the desirability factor in user experience. *CHI*, 4703–4715.

Benedek, J., & Miner, T. (2002). Measuring desirability: New methods for evaluating desirability in a usability lab setting. *Proceedings of Usability Professionals Association*, 8–12.

Picard, R. (2010). Emotion research by the people, for the people. *Emotion Review*, *2*, 250–254.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., & Ehlert, U. (2010). Discriminating stress from cognitive load using a wearable EDA device. *IEEE Transactions on Information Technology in Biomedicine, 14*(2), 410–417.

Biographies

Amanda Davis is a research associate at the Design and Usability Center at Bentley University. Amanda specializes in applying eye-tracking and biometric measurement tools to usability. She holds a B.A. in Economics from Wellesley College, where she focused on behavioral economics. She is currently pursuing an M.S. in Human Factors in Information Design at Bentley University.

Elizabeth Rosenzweig is a principal usability consultant at the Design and Usability Center at Bentley University and founding director of World Usability Day. She holds four patents in intelligent user interface design. Her work includes design, research, and development in areas such as digital imaging, voting technology, mobile devices, and financial and health care systems.

Fiona Tranquada is a senior usability consultant at the Design and Usability Center at Bentley University. She leads user research projects for clients across many industries, including financial services, health care, and e-commerce. Fiona received an M.S. in Human Factors and Information Design degree from Bentley University

This page intentionally left blank

CHAPTER 11

Ten Keys to Success

CONTENTS

Some of the concepts and approaches we've described may be new to some readers, and maybe even a bit overwhelming at first, so we wanted to highlight 10 key elements that will help you succeed. These are lessons we've learned—sometimes the hard way—over the years.

11.1 MAKE DATA COME ALIVE

One of the most important factors that will determine how much impact you have with your research is the extent to which you can make the data come alive for your stakeholders. It's easy for eyes to start glazing over when looking at a bunch of numbers. However, it is very different when you bring the data to life by showing the actual experiences users are having with a product. Even though this is anecdotal, it can have a tremendous impact on getting your point across. Essentially you are putting a real face to your data. It is much harder to ignore your metrics when someone has a deeper level of understanding, or even emotional attachment to the data. Tomer Sharon, in his book "It's Our Research: Getting Buy-in for User Experience Research Projects" (2012) does an excellent job of explaining how critical it is for UX professionals to make your data come alive.

Several techniques can be very helpful in making this happen. First, we recommend that when conducting a usability test you bring key decision makers

Measuring the User Experience. DOI: http://dx.doi.org/10.1016/B978-0-12-415781-1.00011-X

into the lab to observe as many sessions as possible. If you don't have a lab, arrange to use a conference room for the day. A screen-sharing application and a conference call can make for a very effective makeshift observation gallery. Send a reminder message to those you have invited the day before the first session. Nothing speaks louder than observing the user experience firsthand.

Once key decision makers start to see a consistent pattern of results, you won't need to spend much effort convincing them of the need for a design change. But be careful when someone only observes a single usability session. Watching one participant struggle can be dismissed easily as an edge case (e.g., "Our users will be much smarter than that person!"). Conversely, seeing someone fly easily through the tasks can lead to a false sense of security that there are no usability issues with the design. The power of observation is in consistent patterns of results. When key decision makers attend a session, invite them to "come to at least one more session" to get a fuller picture of the results.

Another excellent way to sell UX research is with short video clips. Embedding short video clips into a presentation can make a big difference. The most effective way to illustrate a usability issue is by showing short clips of two or three different participants encountering the same problem. Showing reliable patterns is essential. In our experience, participants who are more animated usually make for better clips. But avoid the temptation to show a dramatic or humorous clip that is not backed by solid data. Make sure each clip is short—ideally less than a minute, and perhaps just 30 seconds. The last thing you want is to lose the power of a clip by dragging it out too long. Before showing a clip, provide appropriate context about the participant (without revealing any private information) and what he or she is trying to do.

If bringing observers into the lab or putting video clips in front of them doesn't work, try presenting a few key UX metrics. Basic metrics around task success, efficiency, and satisfaction generally work well. Ideally, you'll be able to tie these metrics to return on investment (ROI). For example, if you can show how a redesign will increase ROI or how abandonment rates are higher on your product compared to your competition, you'll get the attention of senior management.

TIPS FOR GETTING PEOPLE TO OBSERVE USER SESSIONS

- *Provide a place for observing.* Even if it's a remote session, provide a room with projection or a large screen for observers to watch the session as a group. An important part of observing a usability session is interaction among the observers.
- *Provide food.* For some odd reason, more observers show up when test sessions are scheduled during the lunch hour and food is provided for everyone!
- *Get the sessions on their calendars.* Many people live by their online calendars. If it's not on the calendar, it doesn't happen (for them). Send meeting invitations using your company's scheduling system. Send out a reminder the day before the first session.

- *Provide information.* Observers need to understand what's going on. Make sure that a session schedule, moderator's guide, and any other relevant information are readily available to the observers, both before and during the sessions.
- *Engage the observers.* Give the observers something to do besides just watching. Provide whiteboards or sticky notes for them to record issues. If there are breaks between sessions, have them do a quick review of the key takeaways from the last session.

11.2 DON'T WAIT TO BE ASKED TO MEASURE

Many years ago, one of the best things we ever did was to collect UX data without being asked for it directly. At that time, we started to sense a certain level of hesitancy or even skepticism about purely qualitative findings. Also, the project teams started to ask more questions, specifically around design preferences and the competitive landscape that we know could only really be answered with quantitative data. As a result, we took it upon ourselves to start collecting UX metrics central to the success of the design we were working on.

What is the best way to do this? We recommend starting off with something small and manageable. It's critical that you be successful in your first uses of metrics. If you're trying to incorporate metrics in routine formative testing, start with categorizing types of issues and issue severity. By logging all the issues, you'll have plenty of data to work with. Also, it's easy to collect System Usability Scale (SUS) data at the conclusion of each usability session. It only takes a few minutes to administer the survey, and it can provide valuable data in the long run. That way you will have a quantitative measure across all of your tests and you can show trends over time. As you get comfortable with some of the more basic metrics, you can work your way up the metrics ladder.

A second phase might include some efficiency metrics such as completion times and lostness. Consider some other types of self-reported metrics, such as usefulness–awareness gaps or expectations. Also, explore different ways to represent task success, such as through levels of completion. Finally, start to combine multiple metrics into an overall UX metric or even build your own UX scorecard.

Over time you'll build up a repertoire of different metrics. By starting off small, you'll learn which metrics work for your situation and which don't. You'll learn the advantages and disadvantages of each metric and start to reduce the noise in the data collection process. In our work, it has taken us many years to expand our metrics toolkit to where it is today so don't worry if you're not collecting all the metrics you want at first; you'll get there eventually. Also, be aware that your audience will have an adjustment period. If your audience is only used to seeing qualitative findings, it may take them a while to get adjusted to seeing metrics. If you throw too much at them too quickly, they may become resistant or think you just got back from math camp.

11.3 MEASUREMENT IS LESS EXPENSIVE THAN YOU THINK

No one can use the excuse that metrics take too long to collect or are too expensive. That might have been true 10 years ago, but no longer. There are many new tools available to UX researchers that make data collection and analysis quick and easy, and won't break your budget. In fact, in many cases, running a quantitative-based UX study costs less than a traditional usability evaluation.

Online tools such as UserZoom (www.userzoom.com) and Loop11 (www.loop11.com) are excellent ways to collect quantitative data about how users are interacting with a website or prototype. Studies can be set up in a matter of minutes or hours, and the cost is fairly low, particularly when you compare it to the time setting up a traditional usability evaluation. These tools also provide ways to analyze click paths, abandonment rates, self-reported measures, and many other metrics. In our book "Beyond the Usability Lab" (2010) we highlight many of these tools and provide a step-by-step guide to using online usability testing tools.

Sometimes you are less concerned about actual interaction and more about reaction to different designs. In this situation we recommend taking advantage of many of the online survey tools that now allow you to embed images into the survey and asking questions about those images. Online tools such as Qualtrics (www.qualtrics.com), Survey Gizmo (www.surveygizmo.com), and Survey Monkey (www.surveymonkey.com) all provide the ability to embed images. In addition, some interactive capabilities allow the participant to click on various elements within the image based on questions you provide. The cost of these survey tools is very reasonable, particularly if you sign up for a yearly license.

Many other tools are also very reasonably priced and do an excellent job of collecting data about the user experience. For example, Optimal Workshop (www.optimalworkshop.com) provides a robust suite of tools to build and test any information architecture. If you can't afford your own eye-tracking hardware, EyeTrackShop (www.eyetrackshop.com) allows you to conduct webcam-based eye tracking. This technology has the potential to bring eye-tracking research to a much larger group of researchers without access to hardware. In lieu of traditional usability testing we suggest looking at Usertesting.com (www.usertesting.com) as a way to get very quick feedback about your product in a matter of hours. This tool also has a way of embedding questions into the script, as well as analyzing videos by demographics. While there is certainly some work on the researcher's end, the price can't be beat.

11.4 PLAN EARLY

One of the key messages of this book has been the importance of planning ahead when collecting any metrics. The reason we stress this is because it is so tempting to skip, and skipping it usually has a negative outcome. If you go into a UX study not sure which metrics you want to collect and why, you're almost certainly going to be less effective.

Try to think through as many details as you can before the study. The more specific you can be, the better the outcome. For example, if you're collecting task success metrics and completion times, make sure that you define your success criteria and when exactly you'll turn off the clock. Also, think about how you're going to record and analyze the data. Unfortunately, we can't provide a single, comprehensive checklist to plan out every detail well in advance. Every metric and evaluation method requires its own unique set of plans. The best way to build your checklist is through experience.

One technique that has worked well for us has been "reverse engineering" the data. This means sketching out what the data will look like before conducting the study. We usually think of it as key slides in a presentation. Then we work back from there to figure out what format the data must be in to create the charts. Next, we start designing the study to yield data in the desired format. This isn't faking the results but rather visualizing what the data might look like. Another simple strategy is to take a fake data set and analyze it to make sure that you can perform the desired analysis. This might take a little extra time, but it could help save more time when you actually have the real data set in front of you.

Of course, running pilot studies is also very useful. By running one or two pilot participants through the study, you'll be able to identify some of the outstanding issues that you have yet to address in the larger study. It's important to keep the pilot as realistic as possible and to allow enough time to address any issues that arise. Keep in mind that a pilot study is not a substitute for planning ahead. A pilot study is best used to identify smaller issues that can be addressed fairly quickly before data collection begins.

11.5 BENCHMARK YOUR PRODUCTS

User experience metrics are relative. There's no absolute standard for what is considered "good user experience" and "bad user experience." Because of this, it's essential to benchmark the user experience of your product. This is done constantly in market research. Marketers are always talking about "moving the needle." Unfortunately, the same is not always true in user experience. But we would argue that user experience benchmarking is just as important as market research benchmarking.

Establishing a set of benchmarks isn't as difficult as it may sound. First, you need to determine which metrics you'll be collecting over time. It's a good practice to collect data around three aspects of user experience: effectiveness (i.e., task success), efficiency (i.e., time), and satisfaction (i.e., ease-of-use ratings). Next, you need to determine your strategy for collecting these metrics. This would include how often data are going to be collected and how the metrics are going to be analyzed and presented. Finally, you need to identify the type of participants to include in your benchmarks (broken up into distinct groups, how many you need, and how they're going to be recruited). Perhaps the most important thing to remember is to be consistent from one benchmark to another. This

makes it all the more important to get things right the first time you lay out your benchmarking plans.

Benchmarking doesn't always have to be a special event. You can collect benchmark data (anything that will allow you to compare across more than one study) on a much smaller scale. For example, you could routinely collect SUS data after each usability session, allowing you to easily compare SUS scores across projects and designs. It isn't directly actionable, but at least it gives an indication of whether improvements are being made from one design iteration to the next and how different projects stack up against each other.

Running a competitive user experience study will put your data into perspective. What might seem like a high satisfaction score for your product might not be quite as impressive when compared to the competition. Competitive metrics around key business goals always speak volumes. For example, if your abandonment rates are much higher than your competition, this can be leveraged to acquire budget for future design and user experience work.

11.6 EXPLORE YOUR DATA

One of the most valuable things you can do is to explore your data. Roll up your shirt sleeves and dive into the raw data. Run exploratory statistics on your data set. Look for patterns or trends that are not so obvious. Try slicing and dicing your data in different ways. The keys to exploring your data are to give yourself enough time and not to be afraid to try something new.

When we explore data, especially large data sets, the first thing we do is to make sure we're working with a clean data set. We check for inconsistent responses and remove outliers. We make sure all the variables are well labeled and organized. After cleaning up the data, the fun begins. We start to create some new variables based on the original data. For example, we might calculate top-2-box and bottom-2-box scores for each self-reported question. We often calculate averages across multiple tasks, such as total number of task successes. We might calculate a ratio to expert performance or categorize time data according to different levels of acceptable completion times. Many new variables could be created. In fact, many of our most valuable metrics have come through data exploration.

You don't always have to be creative. One thing we often do is run basic descriptive and exploratory statistics (explained in Chapter 2). This is easy to do in statistical packages such as SPSS and even in Excel. By running some of the basic statistics, you'll see the big patterns pretty quickly.

Also, try to visualize your data in different ways. For example, create different types of scatterplots and plot regression lines, and even play with different types of bar charts. Even though you might never be presenting these figures, it helps give you a sense of what's going on.

Go beyond your data. Try to pull in data from other sources that confirm or even conflict with your assertions. More data from several other sources lend

credibility to the data you share with your stakeholders. It's much easier to commit a multimillion-dollar redesign effort when more than one data set tells the same story. Think of UX data as just one piece of the puzzle—the more pieces of the puzzle, the easier it is to fit it all together and get the big picture.

We can't stress enough the value in going through your data firsthand. If you're working with a vendor or business sponsor who "owns the data," ask for the raw data. Canned charts and statistics rarely tell the whole story. They're often fraught with issues. We don't take any summary data at face value; we need to see for ourselves what's going on.

11.7 SPEAK THE LANGUAGE OF BUSINESS

User experience professionals must speak the language of business to truly make an impact. This means not only using the terms and jargon that management understands and identifies with but, more important, adopting their perspective. In the business world, this usually centers on how to decrease costs and/or increase revenue. So if you're asked to present your findings to senior management, you should tailor your presentation to focus on how the design effort will result in lower costs or increased revenue. You need to approach UX research as an effective means to an end. Convey the perspective that UX is a highly effective way to reach business goals. If you keep your dialogue too academic or overly detailed, what you say probably won't have the impact you're hoping for.

Do whatever you can to tie your metrics to decreased costs or increased sales. This might not apply to every organization but certainly to the vast majority. Take the metrics you collect and calculate how costs and/or revenue is going to change as a result of your design efforts. Sometimes it takes a few assumptions to calculate an ROI, but it's still an important exercise to go through. If you're worried about your assumptions, calculate both a conservative and an aggressive set of assumptions to cover a wider range of possibilities. Case study 10.1 is an excellent example of connecting the dots between UX metrics and business goals.

Also, make sure the metrics relate to the larger business goals within your organization. If the goal of your project is to reduce phone calls to a call center, then measure task completion rates and task abandonment likelihood. If your product is all about e-commerce sales, then measure abandonment rates during checkout or likelihood to return. By choosing your metrics carefully, you'll have greater impact.

11.8 SHOW YOUR CONFIDENCE

Showing the amount of confidence you have in your results will lead to smarter decisions and help enhance your credibility. Ideally, your confidence in the data should be very high, allowing you to make the right decisions. Unfortunately, this is not always the case. Sometimes you may not have a lot of confidence in your results because of a low sample size or a relatively large amount of variance in the data. By calculating and presenting the confidence intervals, you'll have a

much better idea of how much faith or confidence to place in the data. Without confidence intervals, deciding whether some differences are real is pretty much a wild guess, even what may appear to be big differences.

No matter what your data show, show confidence intervals whenever possible. This is especially important for relatively small samples (e.g., less than 20). The mechanics of calculating and presenting confidence intervals is pretty simple. The only thing you need to pay attention to is the type of data you are presenting. Calculating a confidence interval is different if data are continuous (such as completion time) or binary (such as binary task success). By showing the confidence intervals, you can (hopefully) explain how the results generalize to a larger population.

Showing your confidence goes beyond calculating confidence intervals. We recommend that you calculate *p* values to help you decide whether to accept or reject your hypotheses. For example, when comparing average task completion times between two different designs, it's important to determine whether there's a significant difference using a *t* test or ANOVA. Without running the appropriate statistics, you just can't really know.

Of course, you shouldn't misrepresent your data or present it in a misleading way. For example, if you're showing task success rates based on a small sample size, it might be better to show the numbers as a frequency (e.g., six out of eight) as compared to a percentage. Also, use the appropriate level of precision for your data. For example, if you're presenting task completion times, and the tasks are taking several minutes, there's no need to present the data to the third decimal position. Even though you can, you shouldn't.

11.9 DON'T MISUSE METRICS

User experience metrics have a time and a place. Misusing metrics has the potential of undermining your entire UX program. Misuse might take the form of using metrics where none are needed, presenting too much data at once, measuring too much at once, or over-relying on a single metric.

In some situations it's probably better not to include metrics. If you're just looking for some qualitative feedback at the start of a project, metrics might not be appropriate. Or perhaps the project is going through a series of rapid design iterations. Metrics in these situations might only be a distraction and not add enough value. It's important to be clear about when and where metrics serve a purpose. If metrics aren't adding value, don't include them.

It's also possible to present too much UX data at once. Just like packing for a vacation, it's probably wise to include all the data you want to present and then chop it in half. Not all data are equal. Some metrics are much more compelling than others. Resist the urge to show everything. That's why appendices were invented. We try to focus on a few key metrics in any presentation or report. By showing too much data, the most important message is lost.

Don't try to measure everything at once. There are only so many aspects of the user experience that you can quantify at any one time. If a product or business sponsor wants you to capture 100 different metrics, make them justify why each and every metric is essential. It's important to choose a few key metrics for any one study. The additional time to run the study and perform the analyses may make you think twice about including too many metrics at once.

Don't over-rely on a single metric. If you try to get a single metric to represent the entire experience, you're likely to miss something big. For example, if you only collect data on satisfaction, you'll miss everything about the actual interaction. Sometimes satisfaction data might take aspects of the interaction into account, but it often misses a lot as well. We recommend that you try to capture a few different metrics, each tapping into a different aspect of the user experience.

11.10 SIMPLIFY YOUR PRESENTATION

All your hard work comes down to the point where you have to present results. How you choose to communicate your results can make or break a study. There are a few key things you should pay special attention to. First and foremost, your goals need to match those of your audience.

Often you need to present findings to several different types of audiences. For example, you may need to present findings to the project team, consisting of an information architect, design lead, project manager, editor, developer, business sponsor, and product manager. The project team is most concerned with detailed usability issues and specific design recommendations. Bottom line, they want to know the weaknesses with the design and how to fix them.

TIPS FOR AN EFFECTIVE PRESENTATION OF USABILITY RESULTS

- *Set the stage appropriately.* Depending on your audience, you might need to explain or demo the product, describe the research methods, or provide other background information. It all comes down to knowing your audience.
- *Don't belabor procedural details, but make them available.* At a minimum, your audience will usually want to know something about the participants in the study and the tasks they were asked to perform.
- *Lead with positive findings.* Some positive results come out of almost every study. Most people like to hear about features of the design that worked well.
- *Use screenshots.* Pictures really do work better than words in most cases. A screenshot that you've annotated with notes about usability issues can be very compelling.
- *Use short video clips.* The days of an elaborate production process to create a highlights videotape are, thankfully, mostly gone. With computer-based video, it's

much easier and more compelling to embed short clips directly in the appropriate context of your presentation.

- *Present summary metrics.* Try to come up with one slide that clearly shows the key usability data at a glance. This might be a high-level view of task completion data, comparisons to objectives, a derived metric representing overall usability, or a usability scorecard.

You also may need to present to the business sponsors or product team. They're concerned about meeting their business goals, participants' reactions to the new design, and how the recommended design changes are going to impact the project timeline and budget. You may present to senior management too. They want to ensure that the design changes will have the desired impact in terms of overall business goals and user experience. When presenting to senior managers, generally limit the metrics and focus instead on the big picture of the user experience by using stories and video clips. Too much detail usually doesn't work.

Most usability tests produce a long list of issues. Many of those issues do not have a substantial impact on the user experience, for example, minor violations of a company standard or one term on a screen that you might consider jargon. Your goal for a test presentation should be to get the major issues, as you see them, addressed, not to "win" by getting all of the issues fixed. If you present a long list of issues in a presentation, you may be seen as picky and unrealistic. Consider presenting a top 5 or at most a top 10 list and leave minor issues for an off-line discussion.

When presenting results, it's important to keep the message as simple as possible. Avoid jargon, focus on the key message, and keep the data simple and straightforward. Whatever you do, don't just describe the data. It's a surefire way to put your audience to sleep. Develop a story for each main point. Every chart or figure you show in a presentation has a story to it. Sometimes the story is that the task was difficult. Explain why it was difficult and use metrics, verbatims, and video clips to show why it was difficult, possibly even highlighting design solutions. Paint a high-level picture for your audience. They will want perhaps two or three findings to latch onto. By putting all the pieces of the puzzle together, you can help them move forward in the decision making.

References

Albert, W., & Dixon, E. (2003). Is this what you expected? The use of expectation measures in usability testing. *Proceedings of Usability Professionals Association 2003 Conference*, Scottsdale, AZ.

Albert, W., Gribbons, W., & Almadas, J. (2009). Pre-conscious assessment of trust: a case study of financial and health care web sites. *Human factors and ergonomics society annual meeting proceedings*, 53, 449–453. Also <http://www.measuringux.com/Albert_Gribbons_Preconsciousness.pdf>.

Albert, W., & Tedesco, D. (2010). Reliability of self-reported awareness measures based on eye tracking. *Journal of Usability Studies*, *5*(2), 50–64.

Aldenderfer, M., & Blashfield, R. (1984). *Cluster analysis (quantitative applications in the social sciences)*. Beverly Hills, CA: Sage Publications, Inc.

American Institutes for Research. (2001). *Windows XP Home Edition vs. Windows Millennium Edition (ME) public report*. New England Research Center, Concord, MA. Available at <http://download.microsoft.com/download/d/8/1/d810ce49-d481-4a55-ae63-3fe2800cbabd/ME_Public.doc>.

Andre, A. (2003). When every minute counts, all automatic external defibrillators are not created equal. Published in June, 2003 by Interface Analysis Associates <http://www.usernomics.com/iaa_aed_2003.pdf>.

Bangor, A., Kortum, P., & Miller, J. A. (2009). Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of Usability Studies*, *4*, 3.

Bargas-Avila, J. A. & Hornbæk, K. (2011). Old wine in new bottles or novel challenges? a critical analysis of empirical studies of user experience, *CHI '11 Proceedings of the 2011 annual conference on human factors in computing systems*, 2689–2698.

Barnum, C., Bevan, N., Cockton, G., Nielsen, J., Spool, J., & Wixon, D. (2003). The "magic number 5": is it enough for web testing? 2003, April 5–10, Ft. Lauderdale, FL: *CHI*.

Benedek, J., & Miner, T. (2002). Measuring desirability: new methods for evaluating desirability in a usability lab setting. *Usability professionals association 2002 conference,* Orlando, FL, July 8–12. Also available at <*http://www.microsoft.com/usability/UEPostings/DesirabilityToolkit.doc*>. Also see the appendix listing the Product Reaction Cards at <http://www.microsoft.com/usability/UEPostings/ProductReactionCards.doc>.

Bias, R., & Mayhew, D. (2005). *Cost-justifying usability, Second edition: an update for the Internet age*. San Francisco: Morgan Kaufmann.

Birns, J., Joffre, K., Leclerc, J., & Paulsen, C. A. (2002). Getting the whole picture: Collecting usability data using two methods – concurrent think aloud and retrospective probing. *Proceedings of the 2002 Usability Professionals' Association Conference*, Orlando, FL. Available from <http://concordevaluation.com/papers/paulsen_thinkaloud_2002.pdf>.

Breyfogle, F. (1999). *Implementing six sigma: smarter solutions using statistical methods*. New York: John Wiley and Sons.

Brooke, J. (1996). SUS: a quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester & I. L. McClelland (Eds.), *Usability evaluation in industry*. London: Taylor & Francis.

Burby, J., & Atchison, S. (2007). *Actionable web analytics: using data to make smart business decisions*. Indianapolis, IN: Sybex.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction*. London: Lawrence Erlbaum Associates.

Catani, M., & Biers, D. (1998).Usability evaluation and prototype fidelity. In *Proceedings of the human factors and ergonomic society*.

Chadwick-Dias, A., McNulty, M., & Tullis, T. (2003). Web usability and age: how design changes can improve performance. *Proceedings of the 2003 ACM conference on universal usability*, Vancouver, BC, Canada.

Chin, J. P., Diehl, V. A., & Norman, K. L. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. *ACM CHI'88 proceedings*, 213–218.

Clifton, B. (2012). *Advanced web metrics with Google analytics*. Indianapolis, IN: Sybex.

Cockton, G., & Woolrych, A. (2001). Understanding inspection methods: lessons from an assessment of heuristic evaluation. *Joint Proceedings of HCI and IHM: people and computers, XV.*

Cox, E. P. (1980). The optimal number of response alternatives for a scale: a review. *Journal of Marketing Research*, *17*(4), 407–422.

Cunningham, K. (2012). *The accessibility handbook*. Sebastopol, CA: O'Reilly Media.

Dennerlein, J., Becker, T., Johnson, P., Reynolds, C. J., & Picard, R. W. (2003). Frustrating computer users increases exposure to physical factors. In *Proceedings of the international ergonomics association*, August 24–29, Seoul.

Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., et al. (2008). Response rate and measurement differences in mixed mode surveys using mail, telephone, interactive voice response, and the internet. Available at <http://www.sesrc.wsu.edu/dillman/papers/2008/ResponseRateandMeasurement.pdf>.

Ekman, P., & Friesen, W. (1975). *Unmasking the face*. Englewood Cliffs, NJ: Prentice-Hall.

Everett, S. P., Byrne, M. D., & Greene, K. K. (2006). Measuring the usability of paper ballots: efficiency, effectiveness, and satisfaction. P*roceedings of the human factors and ergonomics society 50th annual meeting*. Santa Monica, CA: Human Factors and Ergonomics Society.

Few, S. (2006). *Information dashboard design: the effective visual communication of data*. Sebastopol, CA: O'Reilly Media, Inc.

Few, S. (2009). *Now you see it: simple visualization techniques for quantitative analysis*. Oakland, CA: Analytics Press.

Few, S. (2012). *Show me the numbers: designing tables and graphs to enlighten* (2nd ed.). Oakland, CA: Analytics Press.

Finstad, K. (2010). Response interpolation and scale sensitivity: evidence against 5-point scales. *Journal of Usability Studies*, *5*(3), 104–110.

Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., et al. (2001). What makes web sites credible? a report on a large quantitative study. *Proceedings of CHI'01, human factors in computing systems*, 61–68.

Foraker. (2010). Usability ROI case study: breastcancer.org discussion forums. Retrieved 4/18/2013 from <http://www.usabilityfirst.com/documents/U1st_BCO_CaseStudy.pdf>.

Foresee. (2012). ACSI e-government satisfaction index (Q4 2012). <http://www.foreseeresults.com/research-white-papers/_downloads/acsi-egov-q4-2012-foresee.pdf>.

Friedman, H. H., & Friedman, L. W. (1986). On the danger of using too few points in a rating scale: a test of validity. *Journal of Data Collection*, *26*(2), 60–63.

Garland, R. (1991). The mid-point on a rating scale: is it desirable? *Marketing Bulletin*(2), 66–70. Research Note 3.

Guan, Z., Lee, S., Cuddihy, E., & Ramey, J. (2006). The validity of the stimulated retrospective think-aloud method as measured by eye tracking. In *Proceedings of the ACM SIGCHI conference on human factors in computing systems, 2006* (pp.1253–1262). New York, New York, USA. ACM Press. Available from <http://dub.washington.edu:2007/pubs/chi2006/paper285-guan.pdf>.

Gwizdka, J., & Spence, I. (2007). Implicit measures of lostness and success in web navigation. *Interacting with Computers, 19*(3), 357–369.

Hart, T. (2004). Designing "senior friendly" websites: do guidelines help? *Usability News, 6.1.* <http://psychology.wichita.edu/surl/usabilitynews/61/older_adults-withexp.htm>.

Henry, S. L. (2007). *Just ask: integrating accessibility throughout design*. Raleigh, NC: Lulu.com.

Hertzum, M., Jacobsen, N., & Molich, R. (2002). Usability inspections by groups of specialists: perceived agreement in spite of disparate observations. *CHI*, Minneapolis.

Hewett, T. T. (1986). The role of iterative evaluation in designing systems for usability. In M. D. Harrison & A. F. Monk (Eds.), *People and computers: designing for usability (pp. 196–214)*. Cambridge: Cambridge University Press.

Holland, A. (2012a). *Ecommerce button copy test: did 'Personalize Now' or 'Customize It' get 48% more revenue per visitor?* Retrieved on 4/18/2013 from <http://whichtestwon.com/archives/14511>.

Holland, A. (2012b). *Online newspaper layout test: should photos alternate sides or always appear to the right of stories?* Retrieved on 4/18/2013 from <https://whichtestwon.com/archives/18744>.

Hornbæk, K., & Frøkjær, E. (2008). A study of the evaluator effect in usability testing. *Human-Computer Interaction, 23*(3), 251–277.

Human Factors International. (2002). *HFI helps staples.com boost repeat customers by 67%*. Retrieved 4/18/2013 from <http://www.humanfactors.com/downloads/documents/staples.pdf>.

Hyman, I. E., Boss, S. M., Wise, B. M., McKenzie, K. E., & Caggiano, J. M. (2010). Did you see the unicycling clown? Inattentional blindness while walking and talking on a cell phone. *Applied Cognitive Psychology, 24*, 597–607.

ISO/IEC 25062 (2006). Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – Common Industry Format (CIF) for usability test reports.

Jacobsen, N., Hertzum, M., & John, B. (1998). The evaluator effect in usability studies: problem detection and severity judgments. In *Proceedings of the human factors and ergonomics society*.

Kapoor, A., Mota, S., & Picard, R. (2001). Towards a learning companion that recognizes affect. *AAAI Fall Symposium*, November, North Falmouth, MA.

Kaushik, A. (2009). *Web analytics 2.0: the art of online accountability and science of customer centricity*. Indianapolis, IN: Sybex.

Kirkpatrick, A., Rutter, R., Heilmann, C., Thatcher, J., & Waddell, C. (2006). *Web accessibility: web standards and regulatory compliance*. New York, NY: Apress Media.

Kohavi, R., Crook, T., & Longbotham, R. (2009). *Online experimentation at Microsoft*, Third workshop on Data Mining Case Studies and Practice. Retrieved on 4/18/2013 from <http://robotics.stanford.edu/~ronnyk/ExP_DMCaseStudies.pdf>.

Kohavi, R., Deng, A., Frasca, B., Longbotham, R., Walker, T., & Xu, Y. (2012). Trustworthy online controlled experiments: five puzzling outcomes explained. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining (KDD '12)*. ACM, New York, NY, USA, 786–794.

Kohavi, R., & Round, M. (2004). *Front line internet analytics at Amazon.com*. Presentation at Emetrics Summit 2004. Retrieved on 4/18/2013 from <http://ai.stanford.edu/~ronnyk/emetricsAmazon.pdf>.

Kohn, L. T., Corrigan, J. M., & Donaldson, M. S. (Eds.), (2000). *Committee on quality of health care in America, institute of medicine. "To err is human: building a safer health system."*. Washington, DC: National Academies Press.

Kruskal, J., & Wish, M. (2006). *Multidimensional scaling (quantitative applications in the social sciences)*. Beverly Hills, CA: Sage Publications, Inc..

Kuniavsky, M. (2003). *Observing the user experience: a practitioner's guide to user research*. San Francisco: Morgan Kaufmann.

LeDoux, L., Mangan, E., & Tullis, T. (2005). Extreme makeover: UI edition. Presentation at Usability Professionals Association (UPA) 2005 Annual Conference, Montreal, QUE, Canada. Available from <http://www.upassoc.org/usability_resources/conference/2005/ledoux-UPA2005-Extreme.pdf>.

Lewis, J. (1994). Sample sizes for usability studies: additional considerations. *Human Factors*, *36*, 368–378.

Lewis, J. R. (1991). Psychometric evaluation of an after-scenario questionnaire for computer usability studies: the ASQ. *SIGCHI Bulletin*, *23*(1), 78–81. Also see <http://www.acm.org/~perlman/question.cgi?form=ASQ>.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*(*1*), 57–78. Also see http://www.acm.org/~perlman/question.cgi?form=CSUQ.

Lewis, J. R. & Sauro, J. (2009). The factor structure of the system usability scale. *Proceedings of the human computer interaction international conference (HCII 2009)*, San Diego CA, USA.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, *140*, 55.

Lin, T., Hu, W., Omata, M., & Imamiya, A. (2005). Do physiological data relate to traditional usability indexes? In *Proceedings of OZCHI2005*, November 23–25, Canberra, Australia.

Lindgaard, G., & Chattratichart, J. (2007). Usability testing: what have we overlooked? In *Proceedings of ACM CHI conference on human factors in computing systems*.

Lindgaard, G., Fernandes, G., Dudek, C., & Brown, J. (2006). Attention web designers: you have 50 milliseconds to make a good first impression!. *Behaviour & Information Technology*, *25*, 115–126.

Lund, A. (2001). Measuring usability with the USE questionnaire. *Usability and user experience newsletter* of the STC Usability SIG. See <http://www.stcsig.org/usability/newsletter/0110_measuring_with_use.html>.

Martin, P., & Bateson, P. (1993). *Measuring behaviour* (2nd ed.). Cambridge, UK, and New York: Cambridge University Press.

Maurer, D., & Warfel, T. (2004). Card sorting: a definitive guide. *Boxes and Arrows*, April 2004. Retrieved on 4/18/2013 from <http://boxesandarrows.com/card-sorting-a-definitive-guide/>.

Mayhew, D., & Bias, R. (1994). *Cost-justifying usability*. San Francisco: Morgan Kaufmann.

McGee, M. (2003). Usability magnitude estimation. *Proceedings of human factors and ergonomics society annual meeting*, Denver, CO.

McLellan, S., Muddimer, A., & Peres, S. C. (2012). The effect of experience on system usability scale ratings. *Journal of Usability Studies*, *7*(*2*), 56–67. <http://www.upassoc.org/upa_publications/jus/2012february/JUS_McLellan_February_2012.pdf>.

Miner, G., Elder, J., Hill, T., Nisbet, R., Delen, D., & Fast, A. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Elsevier Academic Press. ISBN 978-0-12-386979-1.

Molich, R. (2011). *CUE-9: The evaluator effect*. <http://www.dialogdesign.dk/CUE-9.html>.
Molich, R., Bevan, N., Butler, S., Curson, I., Kindlund, E., Kirakowski, J., et al., (1998). *Comparative evaluation of usability tests. Usability professionals association 1998 Conference*, 22–26 June 1998 Washington, DC: Usability Professionals Association, pp. 189–200.
Molich, R., & Dumas, J. (2008). Comparative usability evaluation (CUE-4). *Behaviour & Information Technology*, *27*, 263–281.
Molich, R., Ede, M. R., Kaasgaard, K., & Karyukin, B. (2004). Comparative usability evaluation. *Behaviour & Information Technology*, *23*(1), 65–74.
Molich, R., Jeffries, R., & Dumas, J. (2007). Making usability recommendations useful and usable. *Journal of Usability Studies*, *2*(4), 162–179. Available at <http://www.upassoc.org/upa_publications/jus/2007august/useful-usable.pdf>..
Mueller, J. (2003). *Accessibility for everybody: understanding the Section 508 accessibility requirements*. New York, NY: Apress Media.
Nancarrow, C., & Brace, I. (2000). Saying the "right thing": coping with social desirability bias in marketing research. *Bristol Business School Teaching and Research Review*(Summer), 3.
Nielsen, J. (1993). *Usability engineering*. San Francisco: Morgan Kaufmann.
Nielsen, J. (2000). Why you only need to test with 5 users. *AlertBox*, March 19. Available at <http://www.useit.com/alertbox/20000319.html>.
Nielsen, J. (2001). Beyond accessibility: treating users with disabilities as people. *AlertBox*, November 11, 2001. Retrieved on 4/18/2013, from <http://www.nngroup.com/articles/beyond-accessibility-treating-users-with-disabilities-as-people/>.
Nielsen, J. (2005). Medical usability: how to kill patients through bad design, Alertbox, April 11, 2005 <http://www.nngroup.com/articles/medical-usability/>.
Nielsen, J., Berger, J., Gilutz, S., & Whitenton, K. (2008). *Return on Investment (ROI) for usability* (4th ed). Freemont, CA: Nielsen Norman Group.
Nielsen, J., & Landauer, T. (1993). A mathematical model of the finding of usability problems. *ACM proceedings, Interchi* 93, Amsterdam.
Norgaard, M., & Hornbaek, K. (2006). What do usability evaluators do in practice? An explorative study of think-aloud testing. In *Proceedings of designing interactive systems*, pp. 209–218. University Park, PA.
Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL: University of Illinois Press.
Otter, M., & Johnson, H. (2000). Lost in hyperspace: metrics and mental models. *Interacting with Computers*, *13*, 1–40.
Petrie, H., & Precious, J. (2010). Measuring user experience of websites: think aloud protocols and an emotion word prompt list. In *Proceedings of ACM CHI 2010 Conference on human factors in computing systems, 2010*. pp. 3673–3678.
Reichheld, F. F. (2003). One number you need to grow. *Harvard Business Review*, December 2003.
Reynolds, C. (2005). Adversarial Uses of Affective Computing and Ethical Implications. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge. Available at <http://affect.media.mit.edu/pdfs/05.reynolds-phd.pdf>.
Sangster, R. L., Willits, F. K., Saltiel, J., Lorenz, F. O., & Rockwood, T. H. (2001). *The effects of Numerical Labels on Response Scales*. Retrieved on 3/30/2013 from <http://www.bls.gov/osmr/pdf/st010120.pdf>.
Sauro, J. (2009). Composite operators for keystroke level modeling, *Proceedings of the human computer interaction international conference (HCII 2009)*, San Diego CA, USA.
Sauro, J. (2010). *Does better usability increase customer loyalty? The net promoter score and the system usability scale (SUS)*. Retrieved on 4/1/2013 from <http://www.measuringusability.com/usability-loyalty.php>.

Sauro, J. & Dumas J. (2009). Comparison of three one-question, post-task usability questionnaires, *Proceedings of the conference on human factors in computing systems (CHI 2009)*, Boston, MA.

Sauro, J., & Kindlund, E. (2005). A method to standardize usability metrics into a single score. *Proceedings of the conference on human factors in computing systems (CHI 2005)*, Portland, OR.

Sauro, J., & Lewis, J. (2005). Estimating completion rates from small samples using binomial confidence intervals: comparisons and recommendations. *Proceedings of the human factors and ergonomics society annual meeting*, Orlando, FL.

Sauro, J., & Lewis, J. R. (2011). When designing usability questionnaires, does it hurt to be positive?, *Proceedings of the conference on human factors in computing systems (CHI 2011)*, Vancouver, BC, Canada.

Schwarz, N., Knäuper, B., Hippler, H. J., Noelle-Neumann, E., & Clark, F. (1991). Rating scales: numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*, 570–582.

Section 508. (1998). Workforce Investment Act of 1998, Pub. L. No. 105–220, 112 Stat. 936 (August 7). Codified at 29 U.S.C. § 794d.

Shaikh, A., Baker, J., & Russell, M. (2004). What's the skinny on weight loss websites? *Usability News*, 6.1, 2004. Available at <http://psychology.wichita.edu/surl/usabilitynews/61/diet_domain.htm>.

Smith, P. A. (1996). Towards a practical measure of hypertext usability. *Interacting with Computers, 8*(4), 365–381.

Snyder, C. (2006). Bias in usability testing. *Boston Mini-UPA Conference*, March 3, Natick, MA.

Sostre, P., & LeClaire, J. (2007). *Web analytics for dummies*. Hoboken, NJ: Wiley.

Spencer, D. (2009). *Card sorting: designing usable categories*. Brooklyn, NY: Rosenfeld Media.

Spool, J., & Schroeder, W. (2001). Testing web sites: five users is nowhere near enough. *CHI 2001*, Seattle.

Stover, A., Coyne, K., & Nielsen, J. (2002). Designing usable site maps for Websites. Available from <http://www.nngroup.com/reports/sitemaps/>.

Tang, D., Agarwal, A., O'Brien, D., & Meyer, M. (2010). Overlapping experiment infrastructure: more, better, faster experimentation. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '10)*. ACM, New York, NY, USA, 17–26.

Teague, R., De Jesus, K., & Nunes-Ueno, M. (2001). Concurrent vs post-task usability test ratings. *CHI 2001 extended abstracts on human factors in computing systems*, p. 289–290.

Teague, R., DeJesus, K., & Nunes-Ueno, M. (2001). Concurrent vs. post-task usability test ratings. *Proceedings of CHI 2001*, 289–290.

Tedesco, D., & Tullis, T. (2006). A comparison of methods for eliciting post-task subjective ratings in usability testing. *Usability Professionals Association (UPA) 2006 annual conference*, Broomfield, CO, June 12–16.

Trimmel, M., Meixner-Pendleton, M., & Haring, S. (2003). Stress response caused by system response time when searching for information on the internet: psychophysiology in ergonomics. *Human Factors, 45*(4), 615–621.

Tufte, E. R. (1990). *Envisioning information*. Chesire, CT: Graphics Press.

Tufte, E. R. (1997). *Visual explanations: images and quantities, evidence and narrative*. Chesire, CT: Graphics Press.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed). Chesire, CT: Graphics Press.

Tufte, E. R. (2006). *Beautiful evidence*. Chesire, CT: Graphics Press.

Tullis, T. S. (1985). Designing a menu-based interface to an operating system. *Proceedings of the CHI '85 conference on human factors in computing systems,* San Francisco.

Tullis, T. S. (1998). A method for evaluating Web page design concepts. *Proceedings of CHI '98 conference on computer-human interaction*, Los Angeles, CA.

Tullis, T. S. (2007). Using closed card-sorting to evaluate information architectures. *Usability Professionals Association (UPA) 2007 Conference,* Austin, TX. Retrieved on 4/18/2013 from <http://www.eastonmass.net/tullis/presentations/ClosedCardSorting.pdf>.

Tullis, T. S. (2008a). *SUS scores from 129 conditions in 50 studies*. Retrieved on 3/30/2013 from <http://www.measuringux.com/SUS-scores.xls>.

Tullis, T. S. (2008b). *Results of online usability study of Apollo program websites*. <http://www.measuringux.com/apollo/>.

Tullis, T. S. (2011). *Worst usability issue*. Posted July 4, 2011. <http://www.measuringux.com/WorstUsabilityIssue/>.

Tullis, T. S., Mangan, E. C., & Rosenbaum, R. (2007). An empirical comparison of on-screen keyboards. *Human factors and ergonomics society 51st annual meeting,* October 1–5, Baltimore. Available from <http://www.measuringux.com/OnScreenKeyboards/index.htm>.

Tullis, T. S., & Stetson, J.. (2004). A comparison of questionnaires for assessing Website usability. *Usability Professionals Association (UPA) 2004 conference*, June 7–11, Minneapolis, MN. Paper available from <http://home.comcast.net/~tomtullis/publications/UPA2004TullisStetson.pdf>. Slides: <http://www.upassoc.org/usability_resources/conference/2004/UPA-2004-TullisStetson.pdf>.

Tullis, T. S., & Tullis, C. (2007). Statistical analyses of e-commerce websites: can a site be usable and beautiful? *Proceedings of HCI international 2007 conference*, Beijing, China.

Tullis, T. S., & Wood, L. (2004). How many users are enough for a card-sorting study? *Proceedings of Usability Professionals Association Conference*, June 7–11, Minneapolis, MN. Available from http://home.comcast.net/~tomtullis/publications/UPA2004CardSorting.pdf.

Van den Haak, M. J., de Jong, M. D. T., & Schellens, P. J. (2004). Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: a methodological comparison. *Interacting with Computers, 16,* 1153–1170.

Vermeern, A., van Kesteren, I., & Bekker, M. (2003). Measuring the evaluator effect in user testing. In M. Rauterber et al. (Eds.), *Human-computer interaction–INTERACT'03*. pp. 647–654. Published by IOS Press, (c)IFIP.

Virzi, R. (1992). Refining the test phase of the usability evaluation: how many subjects is enough? *Human Factors, 34*(4), 457–468.

Vividence Corp. (2001). Moving on up: move.com improves customer experience. Retrieved October 15, 2001, from <http://www.vividence.com/public/solutions/our+clients/success+stories/movecom.htm>.

Ward, R., & Marsden, P. (2003). Physiological responses to different WEB page designs. *International Journal of Human-Computer Studies, 59,* 199–212.

Wilson, C., & Coyne, K. P. (2001). Tracking usability issues: to bug or not to bug? *Interactions*, May–June.

Withrow, J., Brinck, T., & Speredelozzi, A. (2000). *Comparative usability evaluation for an e-government portal.* Diamond Bullet Design Report, #U1-00-2, Ann Arbor, MI., December. Available at <http://www.simplytom.com/research/U1-00-2-egovportal.pdf>.

Wixon, D., & Jones, S. (1992). Usability for fun and profit: a case study of the design of DEC RALLY, Version 2. Digital Equipment Corporation.

Wong, D. (2010). *The Wall Street Journal guide to information graphics: the do's and don'ts of presenting data, facts, and figures*. New York, NY: W. W. Norton & Company.

Woolrych, A., & Cockton, G. (2001). Why and when five test users aren't enough. In *Proceedings of IHM-HCI2001*, 2, pp. 105–108. Toulouse, France: Ce´padue`s-E´ditions.

This page intentionally left blank

Index

Note: Page numbers followed by "*b*" refer to boxes.

D

E

F

T

U

V

W

X

Z